APPLIED SURVEY DATA ANALYSIS USING STATA:

The Kauffman Firm Survey Data

Joseph Farhat
Alicia Robb

AUGUST 2014

The Kauffman Firm Survey

2004 2005 2006 2007 2008 2009 2010 2011

Ewing Marion KAUFFMAN Foundation

Preface

While entrepreneurial activity is an important part of our economy, data about U.S. businesses in their early years of operation have been extremely limited. Only recently has it become apparent what important contributions new and young businesses make to job creation and innovation activities. As part of an effort to understand the dynamics of new businesses in the United States, the Ewing Marion Kauffman Foundation sponsored the Kauffman Firm Survey (KFS), a panel study of new businesses founded in 2004 that were tracked annually over their first eight years of operation. Tracking businesses over time allows us to follow business evolutions that would not be apparent in cross-sectional snapshots, the more typical collection method. The KFS dataset provides researchers with a unique opportunity to study a panel of new businesses from startup to sustainability (or exit), with longitudinal data centering on topics such as how businesses are financed; the products, services, and innovations these businesses possess and develop in their early years of existence; and the characteristics of those who own and operate them. The Kauffman Firm Survey (KFS) is currently the largest, longest longitudinal survey of new businesses in the world. Data are available through calendar year 2011, the eighth year of operations for continuing businesses. Additionally, since our panel came into existence before the most recent recession, following these businesses allows us to get a picture of how young businesses in the U.S. were affected by the crisis.

We hope that you find the following chapters useful in analyzing the KFS data. Feel free to contact us with comments, suggestions, and/or questions through the KFS website: http://www1.kauffman.org/kfs

Joseph Farhat, Ph.D.

Alicia Robb, Ph.D.

Contents

1.1. Introduction

The Kauffman Firm Survey (KFS), the largest longitudinal study of newly formed businesses, has received considerable attention from researchers in the field of entrepreneurship. Capitalizing on the richest longitudinal study of new businesses, hundreds of researchers are using the data on topics spanning several disciplines. The KFS was constructed using complex survey sample designs where the population of interest was stratified, both explicit and implicit, based on industrial technology level and gender and oversampled within high- and medium- tech industries.

In this chapter, we present a simplified description of the KFS sampling process as well as a multi-step approach that establishes the final weights in the KFS. Next, we examine the impact of ignoring the probability-based weights on the parameter estimates and their standard errors. We conclude with an examination of the design effects' (the finite population correction and stratification) impact on the standard errors. We compare the results when ignoring the sample design effects with the ones that incorporate the sample design effects and show how ignoring the design effects can lead to misleading conclusions.

1.2. The Kauffman Firm Survey

The Kauffman Firm Survey (KFS) was commissioned by the Ewing Marion Kauffman Foundation and was conducted every year from 2005 to 20123 by Mathematica Policy Research, Inc. (MPR). The main objective of the survey was to further understand entrepreneurial activity, to longitudinally track new firms, to understand the dynamics of business development at the owner and the business level in the United States, and to close the informational gap related to new business development (Haviland and Savych, 2007). By capturing the same type of information from the same business over time through data collection at multiple intervals (waves), the longitudinal nature of the KFS data provides opportunities for studying individual-level change over time as well as identifying the underlying dynamics of change.

The KFS longitudinal data is organized in major sections that provide information about business characteristics, strategy and innovation, business organization and human resource benefits, business finances, work behavior, and ownership and demographics of up to ten active-owner-operators.[1] In the KFS, an active-owner-operator is defined as an owner who provides regular assistance or advice regarding the day-to-day operations of the business, rather than providing only money or occasional operating assistance.

[1] The primary sampling units in the KFS are businesses and not owners.

The KFS is a true longitudinal study with a very special feature—it is a single-cohort panel (a type of single indefinite life panels) that tracks the same group of businesses from a common starting point (birth) and records a wide range of information about them over time.[2] Like most longitudinal panel data, the KFS provides the researcher with an opportunity to analyze individual-level change, and it allows for the aggregation of data for businesses over time by examining the occurrence of special events, frequency, timing, and duration, controlling for omitted variables and heterogeneity, and utilizing dynamic panel models. Unlike most longitudinal panel data, the longitudinal nature of the KFS has greater analytical potential to analyze change over time because it remains a single-cohort panel and, thus, can avoid any problems of population composition changes.

1.3. The KFS Target Population and Sample Design

To obtain a sample, we must begin by defining a target population. In any business survey, the target population is the group of businesses the researcher is interested in describing and making statistical inferences about. For KFS, the target population is all new businesses started as independent business, through the purchase of an existing business, or by the purchase of a franchise in the 2004 calendar year in the United States. The KFS target population does not include new businesses that were started as a branch or subsidiary owned by an existing business or a business inherited or a business created as a not-for-profit organization. Notably, a target population could be a subset—by the use of inclusion or exclusion criteria—of a larger population. For example, the target population of the KFS is a subset of a larger population—namely, all new businesses started in 2004 in the United States.

A valid sample must be a representative subset of the target population. Because no single comprehensive national business register of newly formed businesses is available as a frame, the Dun and Bradstreet (D&B) database was chosen as the sampling frame source.[3]

To ensure that a business qualified as part of the target population, inclusion and exclusion criteria must be used to screen eligible businesses. For the KFS, the inclusion and exclusion criteria were:

- Include businesses that were started as independent business, or by the purchase of an existing business, or by the purchase of a franchise in the 2004 calendar year.

[2] However, the "unit of analysis for the KFS design is the sampled business so that if the same business changed ownership from one reporting period to another, it would remain in the sample" (Kauffman Firm Survey Fifth Follow-up Methodology Report); data for businesses that sold or merged were not collected.

[3] A sample frame is a list of elements of the population with appropriate contact information.

- Exclude businesses that were started as a branch or a subsidiary owned by an existing business, that were inherited, or that were created as a not-for-profit organization in the 2004 calendar year.

Then

- Include businesses that have a valid business legal status (sole proprietorship, limited liability company, subchapter S corporation, C-corporation, general partnership, or limited partnership) in 2004.

Then

- Include businesses that have at least one of the following activities:
- Acquired employer identification number during the 2004 calendar year;
- Organized as sole proprietorships reporting that 2004 was the first year they used Schedule C or Schedule C-EZ to report business income on a personal income tax return;
- Reported that 2004 was the first year they made state unemployment insurance payments; or
- Reported that 2004 was the first year they made federal insurance contribution act payments.

In response to the Kauffman Foundation's interest in understanding the dynamics of high-technology, medium-technology, and woman-owned businesses, the KFS is a stratified sample based on industrial technology level (High-Tech, Medium-Tech, and Non-Tech) and gender, which oversamples businesses in high- and medium-tech industries (given a higher selection probability).[4] Table 1 shows the SIC codes used to construct the tech strata of businesses in the D&B sample frame.

Stratification involves dividing the population into non-overlapping groups (strata) defined by selected characteristics. Dividing the population into strata and selecting within strata ensures that the same proportion of respondents in strata and reduces the possibility that the sample will be disproportionately concentrated on one part of the population.

Oversampling a key population subgroup in survey data in response to the small size of a subgroup or for a special interest in that subgroup is a common practice in policy-making surveys. Statistically speaking, the KFS oversampled high-technology and medium-technology businesses to improve the precision of stand-alone analysis and comparative analysis and to improve the precision of cross-sectional and

[4] The technology categories are based on the designation identified by the business's Standard Industry Classification (SIC) code, developed in the early 1990s by researchers from Bureau of Labor Statistics. For details, see Hadlock et al. "High Technology Employment: Another View." *Monthly Labor Review,* July 1991, pp. 26-30.

longitudinal analyses of these sub-groups. It is important to emphasize that woman-owned businesses were not oversampled in the KFS.

Table 1

High Tech	
Two digits SIC	Industry
28	Chemicals and allied products
35	Industrial machinery and equipment
36	Electrical and electronic equipment
38	Instruments and related products
Medium Tech	
Three digits SIC	Industry
131	Crude Petroleum and natural gas operations
211	Cigarettes
229	Miscellaneous textile goods
261	Pulp mills
267	Miscellaneous converted paper products
291	Petroleum refining
299	Miscellaneous petroleum and coal products
335	Nonferrous rolling and drawing
348	Ordnance and accessories, not elsewhere classified
371	Motor vehicles and equipment
372	Aircraft and parts
376	Guided missiles, space vehicles, parts
379	Miscellaneous transportation equipment
737	Computer and data processing services
871	Engineering and architectural services
873	Research and testing services
874	Management and public relations
899	Services, not elsewhere classified
Not High Tech	
	Includes all other industries not listed above

In the KFS, combining the stratification and oversampling yields a disproportionate stratified sample. In disproportionate stratified sampling, the size of each stratum is not proportionate (does not have the same sampling fractions) to its representation in the target population. Thus, weights are used to make the KFS sample a representative sample of the target population.

The precision of generalizing the KFS sample results to the target population depends on the weights selected by the researcher. Ignoring the weights in analyzing the KFS data results in a stratum that is overrepresented or underrepresented, or it could produce skewed results and understate the variances.

The KFS aimed to interview 5,000 businesses that started in 2004. Table 2 summarizes the number of observations used at each step of the process to achieve the final sample. Out of the 251,282 businesses in the sample frame (D&B database), a

stratified sample of 32,469 businesses was selected. The sample was released in waves until the target sample size was achieved. As Table 2 shows, the high and medium tech industries were oversampled.[5] [6]

Table 2

The technology and gender	***D&B Database***		***Sample Count***		***Located***	
ownership strata	N	%	n	%	n	%
High tech, woman owned	527	0.21	527	1.6	491	1.7
High tech, not woman owned	3,342	1.33	3,342	10.3	3,149	10.7
High tech	**3,869**	**1.54**	**3,869**	**11.9**	**3,640**	**12.3**
Medium tech, woman owned	5,547	2.21	1,266	3.9	1,132	3.8
Medium tech, not woman owned	24,114	9.60	6,308	19.4	5,707	19.3
Medium tech	**29,661**	**11.80**	**7,574**	**23.3**	**6,839**	**23.2**
Non tech, woman owned	41,967	16.70	2,760	8.5	2,527	8.6
Non tech, not woman owned	175,785	69.96	18,266	56.3	16,520	56
Non tech	**217,752**	**86.66**	**21,026**	**64.8**	**19,047**	**64.5**
Total	**251,282**	**100.00**	**32,469**	**100.0**	**29,526**	**100**
The technology and gender	***Completes***		***Ineligible***		***Eligible***	
ownership strata	n	%	n	%	n	%
High tech, woman owned	287	1.80	184	1.6	103	2.1
High tech, not woman owned	1,764	10.90	1,162	10.3	602	12.2
High tech	**2,051**	**12.70**	**1,346**	**12.0**	**705**	**14.3**
Medium tech, woman owned	722	4.50	451	4.0	271	5.5
Medium tech, not woman owned	3,288	20.40	2,230	19.9	1,058	21.5
Medium tech	**4,010**	**24.80**	**2,681**	**23.9**	**1,329**	**27**
Non tech, woman owned	1,496	9.30	983	8.8	513	10.4
Non tech, not woman owned	8,599	53.20	6,218	55.4	2,381	48.3
Non tech	**10,095**	**62.50**	**7,201**	**64.1**	**2,894**	**58.7**
Total	**16,156**	**100.00**	**11,228**	**100.0**	**4,928**	**100**

MPR was able to locate 29,526 businesses out of the 32,469 that were released for data collection. Of those located, 16,156 completed the baseline survey.[7] The screening criteria section in the baseline survey indicated that 11,228 businesses were ineligible, resulting in 4,928 businesses as the final sample of eligible businesses.

As the last column in Table 2 shows, the distribution of the observations across the technology and gender ownership strata do not represent the target population; thus, a weighting procedure must be used to correct for sample design (over-sampling) and for non-response (attrition) bias. The use of weights in the KFS compensates for this

[5] Based on the results of a Pilot Test, MPR assumed a 40% response rate and a 40% eligibility rate and retained a 100% reserve sample.

[6] For the Baseline Survey, MPR received two sampling frames of businesses started in 2004 from D&B (in June 2005 and November 2005), totaling roughly 250,000 businesses. MPR balanced the sample size between the two files to reduce unequal sampling weights. The November 2005 D&B file included 62,990 additional businesses with start dates in 2004, resulting in a total pool of 251,282 businesses from the combined June and November files

[7] Completed cases include businesses with complete data for applicable questions. These include eligible and ineligible completes.

differential representation, thereby producing estimates that relate to the target population.[8] Establishing the weights in the KFS will be discussed in the next section.

1.4. Weighting

In complex sample survey data, weighting adjustments are used in studies where the sample is not selected via a random sampling method with an equal probability of selection.[9] A weight is a value assigned to each case in each wave of the survey to remove selection bias and response bias (attrition) from a survey sample and to map the sample back to represent the target population.

A multi-step approach establishes the final weights in the KFS. For the baseline survey, the first step was to create the initial sampling weights (base weight, $w_{t,B}$) to account for unequal sampling probabilities (oversampling). These initial sampling weights are defined as the inverse of the probability of selection, which was calculated in each stratum. According to the theory of design-based inference for probability samples, using the inverse probability weights will yield unbiased estimates of target population statistics. For example, Table 2 shows that the probability of selecting high tech, woman-owned businesses in this sample is equal to one (527/527); thus, the initial sampling weight for this strata is one. Meanwhile, the probability of selecting non-tech, woman-owned businesses is equal to 0.06 (2,760/41,967), and the inverse of the probability of selection is around 15; thus, each business we sampled in this strata represents 15 businesses in the target population.

In the second step, the initial sampling weights ($w_{t,B}$) need to be adjusted to compensate for the businesses that cannot be located and the businesses that did not respond. To determine the probability of locating a business, a logistic propensity model was used for each technology stratum. The fitted binary model ("located" versus "not located," over business characteristics) gives the propensity to locate a business, thereby allowing us to calculate the location adjustment factor as the inverse of the propensity scores ($w_{t,L}$). Next, among located businesses, the fitted binary model ("respondent" versus "non-respondent," over business characteristics) gives the propensity to respond and its inverse is used as the response adjustment factor ($w_{t,R}$). Step one and two, together, represent the joint conditional probability that a business was selected for sampling, was located, and responded to the survey.

The last step in weighting adjustments is post-stratification ($w_{t,P}$); we re-weight the data in each technology group to make the data even more representative of the

[8] We use the terms parameter, statistic, estimate and estimator interchangeably

[9] Adjustments refer to the adjustment for unequal inclusion probabilities, located adjustment, non-response adjustment, and post-stratification adjustment to the weights.

population and to match the population totals.[10] The final weights ($w_{t,F}$) in KFS are the product of the base weight, location adjustment weight, a non-response adjustment weight, and the post-stratification weight:

$$w_{t,F} = w_{t,B} * w_{t,L} * w_{t,R} * w_{t,P} \quad (1)$$

For the first follow-up survey, and up to the seventh follow-up, a similar strategy was applied, wherein the final weights ($w_{t,F}$) for wave *t* are the product of the baseline final weight, location adjustment weight, a non-response adjustment weight, and the post-stratification weight:

$$w_{t,F} = w_{0,F} * w_{t,L} * w_{t,R} * w_{t,P} \quad (2)$$

Table 3

The technology and gender	Unweighted		Weighted (Baseline)	
ownership sampling strata	n	%	N	%
High tech, woman owned	103	2.1	190	0.3
High tech, not woman owned	602	12.2	1,123	1.5
High tech	**705**	**14.3**	**1,313**	**1.8**
Medium tech, woman owned	271	5.5	2,026	2.8
Medium tech, not woman owned	1,058	21.5	7,649	10.4
Medium tech	**1,329**	**27.0**	**9,675**	**13.2**
Non tech, woman owned	513	10.4	14,366	19.6
Non tech, not woman owned	2,381	48.3	47,924	65.4
Non tech	**2,894**	**58.7**	**62,290**	**85.0**
Total	**4,928**	**100.0**	**73,278**	**100.0**

Table 3 depicts the number of unweighted observations in the KFS sample and the equivalent number of businesses in the target population. The estimated target population size in the KFS is 73,278 businesses, which is the estimated number of new businesses in 2004 that meets the KFS new-business screening criteria. Further, the final sample that represents the population is 4,928 businesses, out of which 705 are high-tech, 1,329 are medium-tech, and 2,894 are non-tech businesses. Using the raw survey data sample without correction for the oversampled high-tech and medium-tech businesses provides a biased representation of the target population, and this bias is typically corrected by weighting. After considering the weights, the non-tech businesses represent 85% (rather than 58.7%) of the sample, which is the same as the target population.

[10] "Starting from the third follow-up survey a raking adjustment within the six sampling strata was used to achieve better precision" (KFS Fifth Follow-up Methodology Report, March 29, 2011).

1.4.1. Types of Weights Provided by the KFS

In general, longitudinal data can be analyzed, either as a cross-section or longitudinally. The KFS includes two types of weights; longitudinal weights provide the weight for businesses (longitudinal respondent) that completed the survey in every follow-up from the baseline survey up to the current follow-up. Meanwhile, cross-sectional weights provide the weight for each business that completed the survey in a particular follow-up.[11]

Similar to cross-sectional surveys, longitudinal panel surveys could be used for measuring cross-sectional variation. The major feature of longitudinal panel surveys that distinguishes them from cross-sectional surveys is their capacity to measure longitudinal variation—that is, variation over time at the level of the individual sample member. For example, the baseline survey in the KFS provides the same information as the one-time cross-sectional survey of new businesses founded in 2004; both assess current target population conditions and measure cross-sectional variation among new businesses in 2004. The KFS design allows for measurement of variation among sample members (cross-sectional variation) and variation within sample members across time (longitudinal variation).

Table 4 and Table 5 provide a list of the weights provided on the KFS datasets together with a description of those weights. The difference between cross-sectional weights and longitudinal weights reflects the difference in the sample represented by each type of weight. Thus, each type of weight is related to different research questions and estimation objectives. Each of the seven longitudinal weights represents the same longitudinal sampled observations. For the purposes of panel analyses, longitudinal respondents are generally of interest.

The eight cross-sectional weights in the KFS represent different cross-sectional sampled observations and different cases will contribute in the parameters estimates. Nonetheless, these weights should not be used for longitudinal analyses because they are designed to analyze each wave of the KFS as a cross-section.[12]

---

[11] Completed cases are the businesses that responded to those follow-ups, including businesses that ceased operations.

[12] Cross-sectional analysis can use either cross-sectional weight or longitudinal weight; the former includes many more cases.

Table 4

Cross-sectional weights	Description
wgt_final_0	The cross-section population weight for all businesses who responded in baseline survey.
wgt_final_1	The cross-section population weight for all businesses who responded, permanently stopped operation, temporarily stopped operation, and sold or merged in first follow-up survey.
wgt_final_f2_2	The cross-section population weight for all businesses who responded, permanently stopped operation, temporarily stopped operation, and sold or merged in second follow-up survey and all businesses that permanently stopped operation or sold or merged in any of previous follow-ups.
wgt_final_f3_3	The cross-section population weight for all businesses who responded, permanently stopped operation, temporarily stopped operation, and sold or merged in third follow-up survey and all businesses that permanently stopped operation or sold or merged in any of previous follow-ups.
wgt_final_f4_4	The cross-section population weight for all businesses who responded, permanently stopped operation, temporarily stopped operation, and sold or merged in fourth follow-up survey and all businesses that permanently stopped operation or sold or merged in any of previous follow-ups.
wgt_final_f5_5	The cross-section population weight for all businesses who responded, permanently stopped operation, temporarily stopped operation, and sold or merged in fifth follow-up survey and all businesses that permanently stopped operation or sold or merged in any of previous follow-ups.
wgt_final_f6_6	The cross-section population weight for all businesses who responded, permanently stopped operation, temporarily stopped operation, and sold or merged in sixth follow-up survey and all businesses that permanently stopped operation or sold or merged in any of previous follow-ups.
wgt_final_f7_7	The cross-section population weight for all businesses who responded, permanently stopped operation, temporarily stopped operation, and sold or merged in seventh follow-up survey and all businesses that permanently stopped operation or sold or merged in any of previous follow-ups.

Table 5

Longitudinal weights	Description
wgt_final_1	The longitudinal population weight for all businesses responded in the baseline, and first follow-up surveys and permanently stopped operation, temporarily stopped operation, and sold or merged in first follow-up survey.
wgt_final_f12_long_2	The longitudinal population weight for all businesses responded in the baseline, first and second follow-up surveys and businesses who responded in every follow-up from the baseline up to the follow-up when they permanently stopped operations or sold or merged, and businesses who responded to in every follow-up from the baseline up to the second follow-up and report that they are temporarily stopped operations in second follow-up
wgt_final_f123_long_3	The longitudinal population weight for all businesses responded in the baseline , first, second and third follow-up surveys and businesses who responded in every follow-up from the baseline up to the follow-up when they permanently stopped operations or sold or merged, and businesses who responded to in every follow-up from the baseline up to the third follow-up and report that they are temporarily stopped operations in third follow-up
wgt_final_f1234_long_4	The longitudinal population weight for all businesses responded in the baseline , first, second, third and fourth follow-up surveys and businesses who responded in every follow-up from the baseline up to the follow-up when they permanently stopped operations or sold or merged, and businesses who responded to in every follow-up from the baseline up to the fourth follow-up and report that they are temporarily stopped operations in fourth follow-up
wgt_final_f5_long_5	The longitudinal population weight for all businesses responded in the baseline , first, second, third, fourth, and fifth follow-up surveys and businesses who responded in every follow-up from the baseline up to the follow-up when they permanently stopped operations or sold or merged, and businesses who responded to in every follow-up from the baseline up to the fifth follow-up and report that they are temporarily stopped operations in fifth follow-up
wgt_final_f6_long_6	The longitudinal population weight for all businesses responded in the baseline , first, second, third, fourth, fifth and sixth follow-up surveys and businesses who responded in every follow-up from the baseline up to the follow-up when they permanently stopped operations or sold or merged, and businesses who responded to in every follow-up from the baseline up to the sixth follow-up and report that they are temporarily stopped operations in sixth follow-up
wgt_final_f7_long_7	The longitudinal population weight for all businesses responded in the baseline , first, second, third, fourth, fifth, sixth and seventh follow-up surveys and businesses who responded in every follow-up from the baseline up to the follow-up when they permanently stopped operations or sold or merged, and businesses who responded to in every follow-up from the baseline up to the seventh follow-up and report that they are temporarily stopped operations in seventh follow-up

To examine which cases will contribute in the parameters estimates, using cross-sectional weights and longitudinal weights, one must understand which cases receive a weight and, of those cases that receive a weight, which ones would contribute in the parameters estimates. [13]

The KFS assigns a cross-sectional weight during follow-up for any business that:

- responded to the current follow-up and is still in operation,
- responded to the current follow-up and has permanently stopped operation,
- responded to the current follow-up and has temporarily stopped operation,
- responded to the current follow-up and has sold or merged, and
- Any business that permanently stopped operation or has sold or merged in any of the previous follow-ups.

Table 6 shows the number of businesses that were assigned cross-sectional weights in each follow-up and the sum of weights for those businesses using the cross-sectional weights for that follow-up. Across all waves, businesses that did not respond to the follow-up survey receive a weight of zero.

Longitudinal weights in the KFS (here, only the longitudinal weights for the most recent follow-up survey will be discussed) are assigned for businesses that:

- responded to the survey in every follow-up from the first follow-up to the seventh follow-up,
- responded to the survey in every follow-up from the first follow-up to the follow-up when they permanently stopped operations or sold or merged, and
- responded to the survey in every follow-up from the first follow-up to the seventh follow-up and have reported that they have temporarily stopped operations in the seventh follow-up.

Table 7 presents the number of businesses assigned longitudinal weights in the seventh follow-up. Table 7 indicates that the KFS panel data consist of 3,140 businesses.

[13] "Weights" refer to a weight greater than zero.

Table 6

Business Status	**Baseline** [a]		**First Follow-Up**[b]	
	n	Weight (N)	n	Weight (N)
Responded	4,928	73,278	3,998	66,952
Did not respond			561	0
Sold or Merged			43	691
Permanently stopped operations			260	4,616
Temporarily stopped operations			66	1,020
Total	4,928	73,278	4,928	73,278
Response rate[i]			0.89	
Business Status	**Second Follow-Up** [c]		**Third Follow-Up** [d]	
Responded : No Data			75	1383
Responded	3,390	57,954	2,915	50,452
Did not respond	743	0	825	0
Sold or Merged	47	982	45	687
Permanently stopped operations	321	6,270	299	5,763
Temporarily stopped operations	124	2,246	98	1,687
Stopped operation or sold or merged in any of previous follow-ups.	303	5,827	671	13,307
Total	4,928	73,278	4,928	73,278
Response rate	0.85		0.83	
Business Status	**Fourth Follow-Up** [e]		**Fifth Follow-Up** [f]	
Responded : No Data	49	866	51	939
Responded	2,606	44,634	2,408	40,738
Did not respond	816	0	743	0
Sold or Merged	40	648	36	614
Permanently stopped operations	344	6,354	250	4,498
Temporarily stopped operations	58	1,155	41	813
Stopped operation or sold or merged in any of previous follow-ups.	1,015	19,621	1,399	25,675
Total	4,928	73,278	4,928	73,278
Response rate	0.83		0.85	
Business Status	**Sixth Follow-Up**[g]		**Seventh Follow-Up**[h]	
Responded : No Data	40	837	25	458
Responded	2,126	35,682	2,007	32,681
Did not respond	776	0	676	0
Sold or Merged	38	612	40	670
Permanently stopped operations	218	3,935	209	3,900
Temporarily stopped operations	45	899	30	531
Stopped operation or sold or merged in any of previous follow-ups	1,685	31,314	1,941	35,038
Total	4,928	73,278	4,928	73,278
Response rate	0.84		0.86	

[a] Calculated using wgt_final_0.
[b] Calculated using wgt_final_1.
[c] Calculated using wgt_final_f2_2.
[d] Calculated using wgt_final_f3_3.
[e] Calculated using wgt_final_f4_4.
[f] Calculated using wgt_final_f5_5.
[g] Calculated using wgt_final_f6_6.
[h] Calculated using wgt_final_f7_7.

[i] Response rate is defined as the count of respondents who were interviewed in any given survey year (included in the calculations are stopped operation , sold or merged respondents) as a proportion of the count of eligible businesses at the time of the Baseline Survey

Table 7

Business Status	n	Weight (N)
Permanently stopped operations in the first follow-up	260	6,856
Sold or Merged in first follow-up	43	1,096
Permanently stopped operations in second follow-up	247	7,036
Sold or Merged in second follow-up	36	1,122
Permanently stopped operations in third follow-up	188	4,809
Sold or Merged in third follow-up	36	694
Permanently stopped operations in fourth follow-up	213	5,124
Sold or Merged in fourth follow-up	25	520
Permanently stopped operations in fifth follow-up	141	3,607
Sold or Merged in fifth follow-up	23	542
Permanently stopped operations in sixth follow-up	133	3,139
Sold or Merged in sixth follow-up	20	462
Permanently stopped operations in seventh follow-up	114	2,703
Sold or Merged in seventh follow-up	17	359
Temporarily stopped operations in seventh follow-up	14	317
Responded to first follow-up to seventh follow-up	1,630	34,892
Total	3,140	73,278
Response rate	0.64	

1.4.2. Sample Representativeness and Attrition

All weights, regardless of type (cross-sectional or longitudinal) should be able to map the sample in each follow-up, back to represent the target population at the baseline.

Table 8 presents a comparison of a range of owners, businesses, and industry characteristics of the survey target population at the baseline. Columns three through ten show the owners, businesses, and industry characteristics of the target population using the cross-sectional weights; the eleventh column shows these characteristics using the seventh follow-up longitudinal weights, and the last column shows the sample characteristics (unweighted). As Table 8 shows, all weights in the KFS map the sample in each follow-up, back to represent the characteristics of the target population at the baseline. However, the sample number varies among the cross-sectional weights because the cross-sectional weights consider businesses that responded to a particular follow-up, whereas the number in sample decreases over time for the longitudinal weights; they only consider businesses that responded to all the previous follow-ups.

A comparison of weighted versus unweighted data shows that to generate estimates that are unbiased estimates of the target population, one has to weight the KFS data. An important point identified in Table 8 is that the effect of weighting the data is specific to each characteristic. Some characteristics more common among the over-sampled businesses will appear less common when the data are weighted, while characteristics more common among the under-sampled businesses will appear more common when the data are weighted. For example, having a patent is more common among high-tech businesses (about 14%), yet it is only about 2% in the target population (weighted data). For the characteristics that vary at random among over-sampled and under-sampled businesses, the weighted and unweighted point's estimates will be very close to each other.

Overall, the above analysis of the weights indicates that the weighting scheme used for compensating for sample selection and attrition in the KFS has allowed the samples to remain representative longitudinally and cross-sectionally.

It is also important to note that weighting not only affects point estimates, it also affects the precision of these estimates (the standard error).

Table 8

		Weighted									Unweighted
Characteristics		Mean[a]	Mean[b]	Mean[c]	Mean[d]	Mean[e]	Mean[f]	Mean[g]	Mean[h]	Mean[i]	Mean
Owners - Black	%	8.6	8.6	9.1	9.0	8.9	8.9	8.9	8.8	8.2	7.8
Owners - Asian	%	3.8	3.9	3.7	3.7	3.7	3.6	3.5	3.7	3.0	3.8
Owners - White	%	80.9	80.9	80.7	80.9	81.0	81.1	81.0	81.0	83.1	82.3
Owners - Other races	%	6.7	6.6	6.6	6.5	6.4	6.4	6.6	6.6	5.7	6.2
Education (>Bachelor)	%	56.1	56.4	56.3	56.6	56.7	56.4	56.1	56.5	57.0	59.7
Male	%	67.8	67.8	68.2	67.7	67.7	67.6	67.8	67.9	68.2	72.8
Born in the US	%	88.8	89.3	89.2	89.6	89.7	89.5	89.6	89.5	91.1	88.7
Age		44.3	44.5	44.5	44.5	44.5	44.5	44.5	44.5	44.8	44.8
Serial entrepreneur	%	40.5	40.3	40.6	41.2	40.8	40.9	41.2	41.2	41.0	41.2
Work experience (years)		11.4	11.4	11.4	11.4	11.4	11.4	11.4	11.4	11.4	12.4
Hours worked		41.1	41.0	40.9	40.8	41.0	40.9	41.0	41.1	40.6	40.5
Owner -Employee	%	47.0	46.4	46.5	46.3	46.6	47.3	46.9	47.1	46.5	48.2
Number of Owners											
1	%	70.2	70.2	70.5	70.4	70.4	70.3	70.6	70.3	70.3	69.9
2	%	24.1	23.9	23.7	24.0	24.1	24.1	23.9	24.2	24.0	23.7
3	%	4.0	4.1	4.2	4.0	4.0	4.2	4.0	3.9	4.1	4.4
4	%	1.3	1.5	1.3	1.2	1.3	1.3	1.3	1.3	1.3	1.5
5+	%	0.4	0.4	0.3	0.3	0.2	0.2	0.2	0.3	0.2	0.5
Number of employees		1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5
Employment size											
0	%	34.9	35.3	35.3	35.3	35.3	34.7	34.9	34.8	34.8	34.1
1	%	25.5	25.4	25.7	25.4	24.8	25.2	25.1	25.2	25.8	25.8
2	%	14.8	14.9	14.6	14.7	15.0	14.9	15.0	14.9	14.9	15.0
3	%	6.6	6.4	6.6	6.7	6.9	6.9	6.8	6.7	6.4	6.6
4+	%	18.3	18.0	17.8	17.9	18.0	18.3	18.2	18.5	18.0	18.6
Location											
-Home based business	%	49.3	49.3	49.2	49.4	49.1	49.3	49.4	49.2	50.5	50.6
-Non-home based business	%	50.7	50.7	50.8	50.6	50.9	50.8	50.6	50.8	49.5	49.4
Legal status											
-Sole proprietorship	%	35.8	35.7	35.9	36.0	35.9	35.9	36.0	35.9	35.7	33.2
-Other	%	64.2	64.3	64.2	64.0	64.1	64.1	64.0	64.1	64.3	66.8

Table 8 - continued		Weighted									Unweighted
Characteristics		Mean[a]	Mean[b]	Mean[c]	Mean[d]	Mean[e]	Mean[f]	Mean[g]	Mean[h]	Mean[i]	Mean
Provide a service	%	86.1	86.1	85.9	85.8	85.5	85.5	85.7	85.7	85.6	85.3
Provide a product	%	51.4	51.4	51.3	51.5	51.5	51.6	51.6	52.0	51.2	51.7
Competitive advantage	%	62.8	63.0	62.6	63.1	62.6	63.0	62.8	62.9	63.4	64.6
Have a patent	%	2.2	2.2	2.3	2.4	2.4	2.4	2.3	2.3	2.4	3.8
Have a copyright	%	8.7	8.7	8.8	8.8	8.9	8.8	8.7	8.9	8.6	9.9
Have a trademark	%	13.5	13.3	13.4	13.9	13.4	13.7	13.4	13.7	13.2	14.7
Have a R&D	%	18.1	18.3	18.2	18.0	18.1	18.2	18.2	18.2	17.5	21.4
Total revenue											
-Less than $10000	%	55.1	55.0	54.7	54.7	54.5	54.4	54.8	54.5	53.6	54.5
-$10,000 to $100,000	%	27.9	28.0	28.4	28.3	28.3	28.2	28.3	28.3	29.6	27.7
-$100,000 or more	%	17.1	17.0	16.9	17.0	17.3	17.4	16.9	17.2	16.9	17.9
Total assets											
-Less than $10000	%	40.4	40.5	40.8	41.2	41.2	40.7	40.7	40.4	41.0	40.9
-$10,000 to $100,000	%	38.9	39.0	39.0	38.3	38.4	38.7	39.1	39.1	39.2	38.3
-$100,000 or more	%	20.6	20.6	20.2	20.4	20.4	20.6	20.2	20.5	19.8	20.8
Total debt											
-Less than $10000	%	68.1	68.1	68.5	68.4	68.1	68.3	68.0	67.5	68.3	69.4
-$10,000 to $100,000	%	21.2	21.3	20.9	21.1	21.2	20.7	21.1	21.4	21.2	20.4
-$100,000 or more	%	10.7	10.7	10.6	10.4	10.7	11.0	10.9	11.1	10.5	10.2
Total equity											
-Less than $10000	%	57.5	57.5	57.4	57.7	57.5	57.2	56.9	57.0	57.2	57.9
-$10,000 to $100,000	%	33.4	33.4	33.5	33.2	33.3	33.4	33.8	33.8	34.1	32.4
-$100,000 or more	%	9.1	9.1	9.1	9.1	9.2	9.4	9.3	9.2	8.7	9.7
High tech	%	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	14.3
Medium tech	%	13.2	13.2	13.2	13.2	13.2	13.2	13.2	13.2	13.2	27.0
Non tech	%	85.0	85.0	85.0	85.0	85.0	85.0	85.0	85.0	85.0	58.7

[a] Calculated using wgt_final_0
[b] Calculated using wgt_final_1
[c] Calculated using wgt_final_f2_2
[d] Calculated using wgt_final_f3_3
[e] Calculated using wgt_final_f4_4
[f] Calculated using wgt_final_f5_5
[g] Calculated using wgt_final_f6_6
[h] Calculated using wgt_final_f7_7
[i] Calculated using wgt_final_f7_long_7

1.4.3. The Response Pattern and Weights

For a better understanding of the relation between responding to a particular follow-up and the cross-sectional and longitudinal weights, Table 9 and Table 10 report the response patterns in the KFS and the weights assigned for each pattern using cross-sectional and longitudinal weights. The response pattern column depicts the actual response patterns in the KFS (1 for response and 0 for non-response) from the baseline to the seventh follow-up. For example, a 11101011 pattern shows that 14 businesses responded to the baseline, first, second, fourth, sixth and seventh follow-up surveys, but they did not respond to the third and fifth follow-up surveys. Those businesses are cross-sectional cases in the baseline, first, second, fourth, sixth and seventh follow-up surveys and longitudinal cases only in first and second follow-up surveys.

Table 9 and Table 10 present some of the basic features of cross-sectional and longitudinal weights. First, one notes that a very longitudinal business in a given follow-up will be a cross-sectional business in all the previous follow-ups, and second, the longitudinal sample at time *t* is a subset of the longitudinal sample at time *t-1*.

Data analysts must face the fact that receiving a response (being a complete case) to a follow-up survey does not mean that the respondent will answer all the key survey questions chosen for analysis. Thus, even when weights are assigned to complete cases, the number of cases that will contribute in the parameters estimates will be far less than the number of cases that have been assigned weights. Given that the KFS weights incorporating a survey non-response adjustment, only the effect of item non-response (missing data) needs to be considered.

In the event that item non-response constitutes a small percentage of the variable under analysis, the target population parameters estimates would be reasonably accurate. However, if the item non-response rate is high, then the target population parameters estimates might not necessarily be representative of the target population.

Table 9

Response Patterns	n	Cross-sectional weights(N) by follow-up surveys							
		0[a]	1[b]	2[c]	3[d]	4[e]	5[f]	6[g]	7[h]
10000000	124	1,902	-	-	-	-	-	-	-
10000001	21	353	-	-	-	-	-	-	429
10000010	2	14	-	-	-	-	-	16	-
10000011	8	105	-	-	-	-	-	127	127
10000100	4	24	-	-	-	-	29	-	-
10000101	4	59	-	-	-	-	88	-	74
10000110	4	77	-	-	-	-	92	94	-
10000111	29	448	-	-	-	-	552	549	513
10001000	5	56	-	-	-	67	-	-	-
10001001	1	19	-	-	-	24	-	-	21
10001011	2	44	-	-	-	52	-	49	51
10001100	2	23	-	-	-	29	26	-	-
10001101	1	3	-	-	-	3	3	-	3
10001110	1	19	-	-	-	21	19	21	-
10001111	45	724	-	-	-	910	861	877	845
10010000	7	145	-	-	194	-	-	-	-
10010001	3	33	-	-	41	-	-	-	37
10010010	3	72	-	-	96	-	-	85	-
10010100	1	8	-	-	11	-	9	-	-
10010111	6	86	-	-	108	-	100	101	106
10011000	2	33	-	-	40	44	-	-	-
10011001	1	21	-	-	26	29	-	-	25
10011010	1	4	-	-	10	5	-	7	-
10011011	2	54	-	-	67	72	-	66	68
10011100	3	63	-	-	87	74	74	-	-
10011101	1	7	-	-	8	8	8	-	8
10011110	1	2	-	-	3	3	2	2	-
10011111	75	1,124	-	-	1,453	1,473	1,381	1,452	1,386
10100000	7	56	-	71	-	-	-	-	-
10100001	2	52	-	56	-	-	-	-	61
10100010	1	20	-	24	-	-	-	25	-
10100011	3	38	-	45	-	-	-	51	47

Table 9 - continued

Response Patterns	n	Cross-sectional weights(N) by follow-up surveys							
		0[a]	1[b]	2[c]	3[d]	4[e]	5[f]	6[g]	7[h]
10100100	1	29	-	38	-	-	32	-	-
10100101	3	51	-	59	-	-	62	-	63
10100110	2	15	-	19	-	-	18	17	-
10100111	6	66	-	79	-	-	71	87	71
10101000	1	2	-	4	-	3	-	-	-
10101001	1	21	-	32	-	26	-	-	30
10101011	1	2	-	3	-	2	-	3	2
10101100	1	31	-	36	-	41	43	-	-
10101111	10	162	-	203	-	192	201	204	180
10110000	3	47	-	56	58	-	-	-	-
10110011	3	16	-	19	17	-	-	20	19
10110101	2	36	-	47	42	-	58	-	45
10110111	4	56	-	71	83	-	67	66	65
10111000	1	32	-	32	34	34	-	-	-
10111001	1	2	-	2	3	2	-	-	2
10111011	1	21	-	24	25	33	-	22	24
10111100	3	28	-	38	37	34	33	-	-
10111101	3	47	-	63	59	55	64	-	56
10111110	4	62	-	72	72	82	77	76	-
10111111	138	2,010	-	2,570	2,519	2,508	2,395	2,461	2,419
11000000	74	965	1,101	-	-	-	-	-	-
11000001	27	446	538	-	-	-	-	-	546
11000010	2	32	40	-	-	-	-	37	-
11000011	9	122	139	-	-	-	-	141	135
11000100	3	34	39	-	-	-	37	-	-
11000101	6	89	97	-	-	-	108	-	107
11000110	3	62	78	-	-	-	72	77	-
11000111	24	409	468	-	-	-	499	492	496
11001000	5	78	87	-	-	92	-	-	-
11001011	3	74	80	-	-	90	-	83	88
11001101	2	11	12	-	-	13	13	-	13

Table 9 - continued

Response Patterns	n	Cross-sectional weights(N) by follow-up surveys							
		0[a]	1[b]	2[c]	3[d]	4[e]	5[f]	6[g]	7[h]
11001110	3	73	86	-	-	97	87	83	-
11001111	42	583	680	-	-	713	700	683	692
11001100	5	75	82	-	-	89	94	-	-
11010000	18	298	329	-	358	-	-	-	-
11010001	3	37	40	-	44	-	-	-	40
11010010	1	9	13	-	12	-	-	10	-
11010011	4	68	74	-	88	-	-	77	81
11010100	1	6	6	-	7	-	10	-	-
11010110	1	30	33	-	34	-	36	33	-
11010111	9	147	160	-	175	-	173	179	165
11011000	4	75	88	-	93	91	-	-	-
11011011	2	37	38	-	41	40	-	41	39
11011100	6	103	124	-	128	124	127	-	-
11011101	5	62	68	-	70	77	72	-	71
11011110	3	56	61	-	64	65	63	67	-
11011111	119	1,915	2,161	-	2,364	2,341	2,247	2,300	2,234
11100000	71	1,203	1,372	1,428	-	-	-	-	-
11100001	12	181	212	223	-	-	-	-	219
11100010	3	48	54	58	-	-	-	56	-
11100011	22	326	368	390	-	-	-	423	385
11100100	7	116	132	136	-	-	140	-	-
11100101	2	26	28	32	-	-	30	-	30
11100110	5	84	101	117	-	-	120	105	-
11100111	34	493	575	601	-	-	581	594	590
11101000	8	147	178	178	-	187	-	-	-
11101001	6	102	127	124	-	126	-	-	118
11101010	1	18	19	19	-	21	-	20	-
11101011	14	186	220	226	-	221	-	217	214
11101100	8	109	126	128	-	139	131	-	-
11101101	6	90	104	100	-	104	101	-	101
11110001	27	365	409	445	451	-	-	-	439

Table 9 - continued

Response Patterns	n	Cross-sectional weights(N) by follow-up surveys							
		0[a]	1[b]	2[c]	3[d]	4[e]	5[f]	6[g]	7[h]
11110010	10	206	236	253	254	-	-	254	-
11110011	29	442	500	525	530	-	-	550	534
11110100	8	101	121	118	140	-	118	-	-
11110101	6	83	90	96	97	-	96	-	102
11101110	4	43	47	50	-	54	53	51	-
11101111	122	1,956	2,208	2,392	-	2,383	2,365	2,319	2,287
11110000	62	885	1,024	1,024	1,065	-	-	-	-
11110110	4	71	77	80	84	-	82	75	-
11110111	76	1,174	1,346	1,384	1,414	-	1,417	1,403	1,379
11111000	44	691	791	805	842	824	-	-	-
11111001	28	403	457	466	469	495	-	-	487
11111010	7	101	122	122	127	143	-	116	-
11111011	40	560	645	648	675	672	-	673	652
11111100	57	776	881	927	955	916	918	-	-
11111101	56	680	795	814	813	813	803	-	803
11111110	64	1,032	1,241	1,227	1,284	1,256	1,275	1,233	-
11111111	3,140	46,262	51,950	54,478	55,506	55,266	54,344	54,406	53,455
n		4,928	4,367	4,185	4,103	4,112	4,185	4,152	4,252
N		73,278	73,278	73,278	73,278	73,278	73,278	73,278	73,278

[a] Calculated using wgt_final_0
[b] Calculated using wgt_final_1
[c] Calculated using wgt_final_f2_2
[d] Calculated using wgt_final_f3_3
[e] Calculated using wgt_final_f4_4
[f] Calculated using wgt_final_f5_5
[g] Calculated using wgt_final_f6_6
[h] Calculated using wgt_final_f7_7

Table 10

Response Patterns	n	Longitudinal weights(N) by follow-up surveys						
		1[a]	2[b]	3[c]	4[d]	5[e]	6[f]	7[g]
11000000	74	1,101	-	-	-	-	-	-
11000001	27	538	-	-	-	-	-	-
11000010	2	40	-	-	-	-	-	-
11000011	9	139	-	-	-	-	-	-
11000100	3	39	-	-	-	-	-	-
11000101	6	97	-	-	-	-	-	-
11000110	3	78	-	-	-	-	-	-
11000111	24	468	-	-	-	-	-	-
11001000	5	87	-	-	-	-	-	-
11001011	3	80	-	-	-	-	-	-
11001100	5	82	-	-	-	-	-	-
11001101	2	12	-	-	-	-	-	-
11001110	3	86	-	-	-	-	-	-
11001111	42	680	-	-	-	-	-	-
11010000	18	329	-	-	-	-	-	-
11010001	3	40	-	-	-	-	-	-
11010010	1	13	-	-	-	-	-	-
11010011	4	74	-	-	-	-	-	-
11010100	1	6	-	-	-	-	-	-
11010110	1	33	-	-	-	-	-	-
11010111	9	160	-	-	-	-	-	-
11011000	4	88	-	-	-	-	-	-
11011011	2	38	-	-	-	-	-	-
11011100	6	124	-	-	-	-	-	-
11011101	5	68	-	-	-	-	-	-
11011110	3	61	-	-	-	-	-	-
11011111	119	2,161	-	-	-	-	-	-
11100000	71	1,372	1,504	-	-	-	-	-
11100001	12	212	293	-	-	-	-	-
11100010	3	54	59	-	-	-	-	-

Table 10 - continued

Response Patterns	n	Longitudinal weights(N) by follow-up surveys						
		1[a]	2[b]	3[c]	4[d]	5[e]	6[f]	7[g]
11100011	22	368	423	-	-	-	-	-
11100100	7	132	135	-	-	-	-	-
11100101	2	28	34	-	-	-	-	-
11100110	5	101	140	-	-	-	-	-
11100111	34	575	622	-	-	-	-	-
11101000	8	178	193	-	-	-	-	-
11101001	6	127	130	-	-	-	-	-
11101010	1	19	19	-	-	-	-	-
11101011	14	220	237	-	-	-	-	-
11101100	8	126	137	-	-	-	-	-
11101101	6	104	102	-	-	-	-	-
11101110	4	47	53	-	-	-	-	-
11101111	122	2,208	2,615	-	-	-	-	-
11110000	62	1,024	1,090	1,211	-	-	-	-
11110001	27	409	466	548	-	-	-	-
11110010	10	236	257	291	-	-	-	-
11110011	29	500	534	592	-	-	-	-
11110100	8	121	145	167	-	-	-	-
11110101	6	90	102	104	-	-	-	-
11110110	4	77	89	94	-	-	-	-
11110111	76	1,346	1,470	1,597	-	-	-	-
11111000	44	791	858	944	1,030	-	-	-
11111001	28	457	490	538	573	-	-	-
11111010	7	122	134	161	156	-	-	-
11111011	40	645	686	754	827	-	-	-
11111100	57	881	999	1,064	1,100	1,177	-	-
11111101	56	795	855	944	1,000	1,041	-	-
11111110	64	1,241	1,271	1,485	1,533	1,751	1,698	-
11111111	3,140	51,950	57,139	62,784	67,058	69,309	71,581	73,278
n		4,367	3,983	3,658	3,436	3,317	3,204	3,140
N		73,278	73,278	73,278	73,278	73,278	73,278	73,278

[a] wgt_final_1. [b] wgt_final_f12_long_2, [c] wgt_final_f123_long_3, [d] wgt_final_f1234_long_4, [e] wgt_final_f5_long_5, [f] wgt_final_f6_long_6, [g] wgt_final_f7_long_7

1.5. Complex Sample Design Effects

The KFS was constructed using complex survey sample designs wherein the population of interest is stratified, both explicitly and implicitly, based on industrial technology level and gender and oversampled in high- and medium-tech industries. Thus, weights are only one component of the KFS complex sample design. All features of complex sample design will influence the size of variance for survey estimates.

Complex samples design effects are usually understood in comparison to a simple random sample (SRS) of the same size. A simple random sample consists of independent, identically distributed observations selected with replacements (SRSWR) and with an equal probability of selection from an infinite population; thus, standard inferential statistical methods allow us to make valid inferences about the target population from the sample.

However, a complex sample design generates sampled observations that are not independent, are not identically distributed, are selected without replacement (SRSWOR) with an unequal probability of selection, and are not selected from an infinite population; thus, standard inferential statistical methods must account for the complex design to allow for valid inferences about the target population estimators and their variances.

1.5.1. The Finite Population Correction

Because the size of the target population affects the sampling variance, accounting for the finite nature of the target population is necessary in some special circumstances. Consider a sample of n size sampled from a population that is of finite size N; as the sample size increases ($n \to N$), the sampling variance decreases (e.g., in census, $n = N$, the sampling variance is zero). For a SRSWR, the variance of the sample mean is $\frac{s^2}{n}$, where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Meanwhile, for the SRSWOR, the variance of the sample mean needs to be adjusted because the sampled observations are not independent. Define $\frac{n}{N}$ as the sampling fraction (sampling rate) and $1 - \frac{n}{N}$ as the finite population correction (*fpc*) factor, and the variance of the sample mean from SRSWOR is $\frac{s^2}{n}(1 - \frac{n}{N})$(Cochran, 1977; Kish, 1965; Lohr, 2010).

The finite population correction factor measures the reduction in sampling variance of survey estimates due to sampling without a replacement from a finite population compared to sampling with a replacement from the same population. When the sample size is small compared to the population ($fpc\ factor \approx 1$), the *fpc* factor can be ignored. According to Cochran (1977), the *fpc* factor can be ignored when the sample size is less than 5% of the population size (fpc exceed 95%). In most surveys, the size of the population is quite large and the *fpc* factor is close to one, and

consequently, statisticians choose to ignore the *fpc* factors in favor of conservative estimates of variance.

Table 11 shows the *fpc* factors for the longitudinal as well as for cross sectional follow-ups. While the *fpc* factor for the whole sample is close to 1, we can see that attrition increases the *fpc* factor by effectively decreasing the sample size. The *fpc* factors in Table 11 are calculated under the assumption that we are sampling from the entire population with the same sampling rate.

Table 11

Strata	**Sample (n)**	**Fpc factors (1-[n/N])**
Sample (n)		
Baseline Survey	4,928	0.933
First follow-up (cross sectional)	4,367	0.940
Second follow-up (cross sectional)	4,185	0.943
Third follow-up (cross sectional)	4,103	0.944
Fourth follow-up (cross sectional)	4,112	0.944
Fifth follow-up (cross sectional)	4,185	0.943
Sixth follow-up (cross sectional)	4,152	0.943
Seventh follow-up (cross sectional)	4,252	0.942
First follow-up (longitudinal)	4,367	0.940
Second follow-up (longitudinal)	3,983	0.946
Third follow-up (longitudinal)	3,658	0.950
Fourth follow-up (longitudinal)	3,436	0.953
Fifth follow-up (longitudinal)	3,317	0.955
Sixth follow-up (longitudinal)	3,204	0.956
Seventh follow-up (longitudinal)	3,140	0.957
Target Population (N)	73,278	

1.5.2. Stratification

With stratified sampling, the target population is divided into homogeneous, non-overlapping groups called strata, and then the final sampled observations are randomly selected from the different strata. For this reason, the stratified sample will have smaller standard errors (increased precision) for sample estimates (Cochran, 1977) relative to an SRS of equal size.[14]

Consider a population that is size N and is divided into H strata. Where N_h is the population size of stratum h and n_h is the number of observations sampled using SRS from each stratum, we must have $N = \sum_{h=1}^{H} N_h$ and $n = \sum_{h=1}^{H} n_h$ (Lohr, 2010).

The sample mean can be calculated as:

[14] Cochran (1977) explains why stratification can increase the precision of the estimates relative to SRS: "If each stratum is homogeneous, in that the measurements vary little from one unit to another, a precise estimate of any stratum mean can be obtained from a small sample in that stratum. These estimates can be combined in a precise estimate for the whole population."

$$\bar{y} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h \tag{3}$$

and the variance under SRSWOR is

$$V(\bar{y}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \langle \frac{N_h}{N} \rangle^2 \frac{{s_h}^2}{n_h} \tag{4}$$

Where h_i is the unit number within stratum h, $\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$, ${s_h}^2 = \sum_{i=1}^{n_h} \frac{\langle y_{hi} - \bar{y}_h \rangle^2}{n_h - 1}$ and $\left(1 - \frac{n_h}{N_h}\right)$ is the finite population correction (fpc) factor for stratum h.

Equation 4 shows that a stratified SRS is more efficient (has smaller variance) than an SRS because the variance of the sample estimate depended only on the within-stratum variances and there is no between-stratum variances component. In other words, given that total variance = within-variance + between-variance and because stratified sampling assumes that between-variance is zero, variance from a stratified SRS is always smaller than from an SRS. Equation 4 also suggests that the more homogeneous the strata are, the greater the gain in precision arising from stratification.

Equation 4 shows that with different sampling rates in different strata, the *fpc* factors may be very small, which cannot be ignored. In this case ignoring the *fpc* factors will lead to an overestimate of the variance in some strata.

The same results apply for complex sample design. The estimate of the mean is

$$\bar{y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}\, y_{hij} \Big/ \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} \tag{5}$$

and the estimated variance is[15]

$$V(y) = \sum_{h=1}^{H} \frac{n_h(1 - f_h)}{n_h - 1} \left[\sum_{i=1}^{n_h} \left(\sum_{j=1}^{m_{hi}} w_{hij}\, y_{hij} \right)^2 - \left(\left(\sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}\, y_{hij} \right)^2 \Big/ n_h \right) \right] \tag{6}$$

Where

$h = 1,2, \ldots H$ is the stratum number, with a total of H strata

$i = 1,2, \ldots n_h$ is the cluster number within stratum h, with a total of n_h clusters

[15] This notation is also applicable to other sample designs. For example, for a sample design without stratification, you can let H = 1; for a sample design without clusters, you can let m_{hi} = 1 for every h and i.

$j = 1,2, \ldots m_{hi}$ is the unit number within cluster i of stratum h , with a total of m_{hi} units
w_{hij} is the sampling weight for observation j j in cluster i is of stratum h
$f_h = n_h/N_h$ is the sampling rate i for stratum h
N_h is the population size of stratum h

In the case of the KFS, the high-technology stratum included only 3,869 businesses of the target population; to insure a large pool of these businesses for the longitudinal panel required a very high sampling rate for this stratum. Sampling a substantial fraction of a stratum results in decreased variability compared to a sample from an infinite population. Table 12 shows the *fpc* factors for each stratum. While the *fpc* factor for the whole sample is close to 1, this is not the case when we consider the *fpc* for each strata. With an *fpc* factor close to 50% in some strata, it is clear that ignoring the *fpc* factors within strata will overestimate the variance of high-tech stratum.

Table 12

Cross sectional	***fpc* factors for each stratum**					
Strata	Woman owned			Not woman owned		
	High tech	Medium tech	Non tech	High tech	Medium tech	Non tech
Baseline Survey	0.458	0.866	0.964	0.464	0.862	0.950
First follow-up	0.532	0.878	0.968	0.531	0.877	0.956
Second follow-up	0.527	0.885	0.969	0.549	0.880	0.958
Third follow-	0.521	0.883	0.970	0.564	0.884	0.959
Fourth follow-up	0.532	0.887	0.970	0.559	0.884	0.959
Fifth follow-up	0.532	0.884	0.970	0.559	0.880	0.958
Sixth follow-up	0.548	0.886	0.969	0.573	0.881	0.958
Seventh follow-up	0.532	0.881	0.969	0.550	0.878	0.958
Longitudinal						
Seventh follow-up	0.632	0.906	0.976	0.678	0.908	0.969

Researchers who are interested in studying high tech , medium tech or non-tech separately should avoid using the technology and gender ownership sampling strata variable that Mathematica used to select the KFS sample to split their sample.

The primary industry of the business confirmed or updated during every survey and as a results the sampling strata variable do not reflect the current primary industry classification for the business.

1.5.3. Variance Estimation

When complex sample design features are implemented in the analysis, most of the statistics of interest will not be simple linear functions of the observed data; thus, a more complex method for estimating the variances is required.[16] The two major approaches to compute the variance in a complex design sample are the use of a Taylor

[16] Even simple statistics, such as the mean, become non-linear in a complex survey.

series linearization (TSL) of the estimator or the use of repeated re-sampling variance estimation procedures. Among the re-sampling (replication) methods, the Jackknife repeated replication (JRR) method is the most used approach. For a detailed description of this method, see Lee and Forthofer (2005), Marsden and Wright (2010), and Lohr (2010).

1.6. Assessing the Loss or Gain in Precision: Design Effect

Measuring the impact of various complexities in the survey design on the variance of an estimate can be achieved by the design effect factor (DEFF). The design effect factor is defined as the ratio of the variance of an estimate under the actual survey complex design to the variance of that estimate with a simple random sample of the same size. For example, the interpretation of a value of the design effect factor (DEFF) of 1.3 is that the sample variance is 1.3 times bigger than it would be if the survey were based on SRS of equal size; alternatively, the sample would have to be 1.3 as large to yield the standard errors that would have been found based on SRS.

The impact of the design effect on the standard error (DEFT) of the statistic is measured by taking the square root of the design effect factor. As rules of thumb, the magnitude of the design effects are considered to be low if the DEFF is in the range of 1-1.3, medium if the DEFF is in the range of 1.4-1.9, and high for a DEFF of 2 or more (Aday and Llewellyn, 2006).

1.6.1. Descriptive Statistics

To illustrate the impact of the survey design features on the variance of a univariate estimates, Table 13 reports the standard errors for a range of owners and businesses characteristics of the survey population at the baseline using the TSL method for variance estimation. Meanwhile, Table 14 reports the same estimates using the JRR method for variance estimation.

As columns six and seven show, accounting for strata and *fpc* in addition to the weights reduces the estimate's standard errors relative to using the weights only. The point estimates (mean) are identical regardless of how many design features were implemented (stratification and finite population correction) or how the variances were computed. This is because weights affect point estimates and standard errors, while stratification and finite population correction only affect the standard errors.

The unweighted results show that the point estimates are biased relative to weighted results. On the other hand, implementing the design features into the calculations of the standard errors increases the size of the standard error relative to the unweighted standard error.

Both the DEFF and the DEFT are reported in columns eight and nine. The first observation regarding the design effect is not fixed across variables. The design effect

would be high for characteristics that are the same for all of the businesses (e.g., race) and would be low for characteristics that are different for different businesses (e.g., have a patent, work experience, age). For example, the design-based variance estimate for the percentage of white owners is about 42% larger than the estimate from the hypothetical SRS design including the full population. Meanwhile, the design-based variance estimate for the percentage of businesses having a patent is about 14% larger than the estimate from the hypothetical SRS design including the full population.

The second observation, the design effect factor (DEFF), for the stratification variable is below one, which indicates that stratification makes our sample more accurate than a simple random sample. The design-based variance estimate for the percentage of male owners is about 21% smaller than the estimate from the hypothetical SRS design including the full population. Also, we should recognize that stratification not only affects the design effect factor (DEFF) for the stratification variable, but also any other variables related to stratification variables.

Given that the total design effect (DEFF) components can be decomposed into the design effect components of clustering, the design effect components of weighting, and the design effect components of stratification, it is important to recognize that the DEFF for a particular characteristic will not be the same from one follow-up to another. Moreover, because attrition increases the sampling variance as well as the variance of the sample weights over time, the DEFF will be an increasing function of attrition.

Comparing the standard errors produced by the TSL method to the ones produced by the JRR shows that the standard errors produced by JRR are slightly higher than the ones calculated by the TSL method. Consequently, the JRR method provides larger confidence intervals and higher DEFF than TSL does.[17]

Researchers interested in studying subpopulation based on strata, or groups of strata, should consider estimating the design effect factors (DEFF) computed with respect to SRS within an individual strata or groups of strata. Statistically speaking, because the KFS is a stratified sample based on industrial technology level (high-tech, medium-tech, and non-tech) and gender, the sample of high-tech, medium-tech, or non-tech businesses standalone is equivalent to a stratified simple random sample (e.g., the high-tech sample is a stratified simple random sample based on gender).[18] Table 15 shows the DEFFs for a range of owners and business characteristics based on a subpopulation of groups of strata, namely, high-tech, medium-tech, and non-tech. As

[17] To four significant digits, the standard errors are almost the same under both TSL and JRR.

[18] In the KFS sampling process, businesses within each technology and woman-owned indicator sampling stratum were sorted by two control variables (implicit stratification): (1) D&B employee count categories, and (2) three-digit zip code; then, sampling selection was done using Chromy's sequential random sampling method.

expected, the DEFFs are lower when we consider DEFFs by a group of strata relative the DEFFs for the entire population.

In general, the DEFF and DEFT indicate that the impact of the various complexities in the KFS design on inflating the standard errors of the baseline survey characteristics ranges from low to medium; moreover, this impact is trivial for the estimates within strata or groups of strata.

Table 13

Design features Implemented	Un-weighted	Weights[a] & Strata & fpc	Un-weighted	Weights[a]	Weights[a] & Strata		Weights[a] & Strata & fpc	
Variance Estimation		**TSL**		**TSL**	**TSL**	**TSL**	**TSL**	**TSL**
Characteristics	Mean	Mean	Std. Err.	Std. Err.	Std. Err.	Std. Err.	DEFF	DEFT
White	82.3	80.9	0.005293	0.006459	0.006458	0.006294	1.417	1.150
Education (>Bachelor)	59.7	56.1	0.006991	0.008221	0.008177	0.007971	1.362	1.127
Male	72.8	67.8	0.005746	0.007293	0.005330	0.005195	0.790	0.859
Age	44.8	44.3	0.148201	0.169321	0.169308	0.164914	1.346	1.121
Work experience (years)	12.4	11.4	0.142811	0.160682	0.158284	0.154123	1.290	1.097
-Home based business	50.6	49.3	0.007126	0.008252	0.008215	0.008005	1.353	1.123
-Sole proprietorship	33.2	35.8	0.006708	0.007961	0.007888	0.007688	1.359	1.126
Number of employees	1.5	1.5	0.021256	0.024663	0.024558	0.023937	1.361	1.127
Competitive advantage	64.6	62.8	0.006861	0.008074	0.008069	0.007866	1.379	1.134
Have a patent	3.8	2.2	0.002739	0.002246	0.002242	0.002172	1.144	1.033
Total assets								
-Less than $10000	40.9	40.4	0.007014	0.008100	0.008062	0.007854	1.349	1.122
-$10,000 to $100,000	38.3	38.9	0.006936	0.008057	0.008059	0.007852	1.366	1.129
-$100,000 or more	20.8	20.6	0.005785	0.006702	0.006667	0.006496	1.358	1.125

[a] Calculated using wgt_final_0.

Table 14

Design features Implemented	Un-weighted	Weights[a] & Strata & fpc	Un-weighted	Weights[a]	Weights[a] & Strata		Weights[a] & Strata & fpc	
Variance Estimation		**JRR**		**JRR**	**JRR**	**JRR**	**JRR**	**JRR**
Characteristics	Mean	Mean	Std. Err.	Std. Err.	Std. Err.	Std. Err.	DEFF	DEFT
White	82.3	80.9	0.005293	0.006460	0.006459	0.006294	1.417	1.150
Education (>Bachelor)	59.7	56.1	0.006991	0.008222	0.008177	0.007971	1.362	1.127
Male	72.8	67.8	0.005746	0.007294	0.005330	0.005195	0.790	0.859
Age	44.8	44.3	0.148201	0.169337	0.169310	0.164916	1.346	1.121
Work experience (years)	12.4	11.4	0.142811	0.160697	0.158284	0.154124	1.290	1.097
-Home based business	50.6	49.3	0.007126	0.008253	0.008215	0.008005	1.353	1.123
-Sole proprietorship	33.2	35.8	0.006708	0.007962	0.007888	0.007688	1.359	1.126
Number of employees	1.5	1.5	0.021256	0.024666	0.024558	0.023938	1.361	1.127
Competitive advantage	64.6	62.8	0.006861	0.008075	0.008069	0.007866	1.379	1.134
Have a patent	3.8	2.2	0.002739	0.002247	0.002242	0.002172	1.144	1.033
Total assets								
-Less than $10000	40.9	40.4	0.007014	0.008100	0.008062	0.007855	1.349	1.122
-$10,000 to $100,000	38.3	38.9	0.006936	0.008058	0.008059	0.007852	1.366	1.129
-$100,000 or more	20.8	20.6	0.005785	0.006703	0.006667	0.006496	1.358	1.125

[a] Calculated using wgt_final_0.

Table 15

Group		High-Tech			Medium-Tech			Non-Tech		
Characteristics		Mean	DEFF	DEFT	Mean	DEFF	DEFT	Mean	DEFF	DEFT
Owners - Black	%	4.3	1.085	1.042	8.8	1.019	1.010	8.6	1.078	1.038
Owners - Asian	%	3.5	1.132	1.064	5.2	1.092	1.045	3.6	1.080	1.039
Owners - White	%	86.6	1.067	1.033	80.4	1.037	1.018	80.9	1.069	1.034
Owners - Other races	%	5.7	1.026	1.013	5.6	1.036	1.018	6.9	1.071	1.035
Education (>Bachelor)	%	60.9	1.028	1.014	75.0	1.011	1.006	53.1	1.032	1.016
Male	%	81.4	0.621	0.788	74.8	0.502	0.709	66.4	0.596	0.772
Born in the US	%	87.0	1.066	1.033	86.8	1.056	1.027	89.2	1.064	1.032
Age		46.8	1.028	1.014	45.0	1.006	1.003	44.1	1.026	1.013
Serial entrepreneur	%	50.4	1.029	1.015	37.7	1.026	1.013	40.7	1.026	1.013
Work experience (years)		14.0	1.034	1.017	14.5	1.009	1.004	10.9	0.996	0.998
Hours worked		41.0	1.017	1.009	38.5	1.015	1.007	41.5	1.024	1.012
Owner -Employee	%	49.7	1.023	1.012	51.7	1.019	1.010	46.2	1.031	1.015
Number of Owners										
1	%	61.9	1.027	1.013	74.1	1.042	1.021	69.8	1.030	1.015
2	%	27.1	1.018	1.009	19.9	1.046	1.023	24.7	1.032	1.016
3	%	7.2	1.054	1.027	4.5	1.059	1.029	3.9	1.030	1.015
4	%	2.7	1.014	1.007	1.2	1.026	1.013	1.3	1.026	1.013
5+	%	1.1	0.990	0.995	0.3	1.128	1.062	0.4	0.968	0.984
Number of employees		1.8	1.032	1.016	1.4	1.048	1.023	1.5	1.019	1.009
Employment size										
0	%	29.4	1.039	1.019	34.0	1.014	1.007	35.2	1.031	1.015
1	%	22.1	1.037	1.018	29.4	1.010	1.005	24.9	1.039	1.019
2	%	13.1	1.013	1.007	17.2	1.022	1.011	14.4	1.038	1.019
3	%	7.8	1.027	1.014	5.7	1.029	1.014	6.7	1.039	1.019
4+	%	27.6	1.026	1.013	13.8	1.073	1.036	18.8	1.014	1.007
Location										
-Home based business	%	37.7	1.020	1.010	64.9	1.033	1.016	47.2	1.033	1.016
-Non-home based business	%	62.3	1.020	1.010	35.2	1.033	1.016	52.8	1.033	1.016
Legal status										
-Sole proprietorship	%	22.7	1.053	1.026	31.0	0.989	0.994	36.8	1.022	1.011
-Limited liability company	%	77.4	1.053	1.026	69.0	0.989	0.994	63.2	1.022	1.011

Table 15 - continued

Group		High-Tech			Medium-Tech			Non-Tech		
Characteristics		Mean	DEFF	DEFT	Mean	DEFF	DEFT	Mean	DEFF	DEFT
Provide a service	%	67.1	1.041	1.020	94.4	1.039	1.020	85.3	1.039	1.019
Provide a product	%	76.1	1.042	1.021	34.8	1.020	1.010	53.5	1.031	1.015
Have a patent	%	13.2	1.056	1.028	3.7	1.063	1.031	1.8	1.033	1.017
Have a copyright	%	12.3	1.079	1.039	14.1	1.035	1.018	7.8	1.048	1.024
Have a trademark	%	22.9	1.050	1.024	14.0	1.030	1.015	13.2	1.035	1.017
Have a R&D	%	35.3	1.040	1.020	25.9	1.033	1.017	16.5	1.048	1.024
Total revenue										
-Less than $10000	%	53.1	1.022	1.011	55.0	1.016	1.008	55.1	1.027	1.013
-$10,000 to $100,000	%	20.6	1.023	1.012	30.5	1.012	1.006	27.6	1.029	1.014
-$100,000 or more	%	26.4	1.025	1.012	14.5	1.046	1.023	17.3	1.011	1.006
Total assets										
-Less than $10000	%	32.2	1.021	1.010	50.7	1.018	1.009	39.0	1.033	1.017
-$10,000 to $100,000	%	36.8	1.028	1.014	36.6	1.022	1.011	39.3	1.033	1.016
-$100,000 or more	%	31.0	1.022	1.011	12.7	1.044	1.022	21.6	1.009	1.004
Total debt										
-Less than $10000	%	59.9	1.034	1.017	82.2	1.041	1.020	66.1	1.025	1.013
-$10,000 to $100,000	%	25.2	1.031	1.016	13.5	1.042	1.021	22.3	1.024	1.012
-$100,000 or more	%	14.9	1.041	1.020	4.3	1.055	1.027	11.6	1.018	1.009
Total equity										
-Less than $10000	%	47.0	1.011	1.005	66.4	1.031	1.015	56.3	1.028	1.014
-$10,000 to $100,000	%	33.5	1.017	1.008	27.8	1.034	1.017	34.3	1.025	1.013
-$100,000 or more	%	19.4	1.038	1.019	5.9	1.091	1.045	9.4	1.040	1.020
Agriculture	%							3.4	0.965	0.983
Mining	%							0.2	0.992	0.996
Construction	%							12.6	1.078	1.038
Manufacturing	%	100.0			2.6	1.066	1.033	5.3	1.045	1.022
Transportation, and utilities	%							4.8	1.007	1.004
Wholesale trade	%							7.0	1.051	1.025
Retail trade	%							19.2	1.026	1.013
Finance, insurance & real estate	%							11.1	1.033	1.017
Services	%				97.4	1.066	1.033	36.7	1.029	1.014

1.6.2. Analytical Statistics

While it is clear that the survey design features can considerably affect point estimate, variance, and standard errors in descriptive statistics, this is not the case for analytical statistics.[19,20] Regardless of the controversy among theorists about role of sampling weights in the statistical analysis of survey data, survey statisticians who approach survey data from a design-based perspective incorporate the sample weights into every analysis.

Pfeffermann and Holmes (1985), Kott (1991), and Pfeffermann (1993) highlighted that the important role of sampling weights in the statistical analysis of survey data mainly protect against non-ignorable sampling designs, which could cause selection bias and protect against misspecification of the model holding in the population.

Incorporating the sample weights into analytical studies involves a bias-variance trade-off. While weights are essential for bias reduction, they do increase the variance.[21] Given that the MSE decomposes into a sum of the bias and variance of the estimator, both quantities are important and need to be as small as possible to achieve good estimation performance. It is common to trade-off some increase in variance for a larger decrease in the bias, and vice versa.

To illustrate the impact of the survey design features on analytical statistics, Table 16 reports the results of a binary logistic regression—the logistic regression predicting profitability in the start-up year (has profit in the baseline=1, has a loss in the baseline =0). Table 16 shows the results of logistic regression for the same baseline data, with and without the survey design taken into account. The results using both TLS and JRR are reported.

The estimated parameters changed when survey design was taken into account, and the estimated standard errors increased, as reflected in the DEFF. Despite the increased standard errors, the coefficients for race, gender, work experience, home-based business, and provides services are significantly different from 0. Meanwhile, the coefficient of having a patent is significantly different from 0 for the sample, but not for the target population.

Because of oversampling of the high-tech and medium-tech industries, the bias in the unweighted estimates of the characteristics that are more common in the oversampled industries are higher relative to the estimates for the less common characteristics.

---

[19] Descriptive statistics are used to provide information about the specific data being analyzed.
[20] Analytical statistics are used to draw conclusions about a population based on sample data.
[21] An alternative to weighting is to model the survey design in your statistical model; however, it is not possible to include all the design information in your model.

Table 16

Design features Implemented	Unweighted			Weights& Strata & fpc				
Variance Estimation: TSL	Coef.	Std. Err.	P-Value	Coef.	Std. Err.	P-Value	DEFF	DEFT
White	0.120	0.083	0.149	0.172	0.093	0.065	1.402	1.143
Male	0.164	0.077	0.034	0.233	0.086	0.007	1.472	1.172
Age	-0.005	0.003	0.140	-0.004	0.004	0.285	1.380	1.135
Work experience	0.024	0.003	0.000	0.026	0.004	0.000	1.360	1.126
Number of employees	-0.088	0.024	0.000	-0.094	0.027	0.001	1.384	1.136
Competitive advantage	0.078	0.064	0.227	0.091	0.073	0.211	1.374	1.132
Have a patent	-0.339	0.169	0.045	-0.357	0.231	0.122	1.172	1.045
Assets -$10,000 to $100,000	0.097	0.070	0.167	0.032	0.079	0.684	1.350	1.122
Assets -$100,000 or more	0.225	0.092	0.014	0.234	0.105	0.025	1.388	1.138
Home based business	0.194	0.066	0.003	0.273	0.074	0.000	1.377	1.133
Provide a service	0.543	0.090	0.000	0.476	0.107	0.000	1.436	1.157
Variance Estimation: JRR	Coef.	Std. Err.	P-Value	Coef.	Std. Err.	P-Value	DEFF	DEFT
White %	0.120	0.083	0.149	0.172	0.094	0.066	1.415	1.149
Male %	0.164	0.077	0.034	0.233	0.087	0.007	1.484	1.176
Age	-0.005	0.003	0.140	-0.004	0.004	0.288	1.392	1.139
Work experience	0.024	0.003	0.000	0.026	0.004	0.000	1.373	1.132
Number of employees	-0.088	0.024	0.000	-0.094	0.027	0.001	1.396	1.141
Competitive advantage	0.078	0.064	0.227	0.091	0.073	0.213	1.384	1.136
Have a patent	-0.339	0.169	0.045	-0.357	0.235	0.130	1.216	1.065
Assets -$10,000 to $100,000	0.097	0.070	0.167	0.032	0.079	0.685	1.360	1.126
Assets -$100,000 or more	0.225	0.092	0.014	0.234	0.105	0.026	1.400	1.143
Home based business	0.194	0.066	0.003	0.273	0.074	0.000	1.388	1.138
Provide a service	0.543	0.090	0.000	0.476	0.107	0.000	1.452	1.164

[a] Calculated using wgt_final_0.

1.6.3. Analysis of Subpopulations

Analysis of subpopulations (Cochran, 1977; Lohr, 2010); Kish, 1987) (also called domains analysis, subgroup analysis, subpopulation analysis, or subdomain analysis) refers to the computation of descriptive and analytical statistics for subpopulations, e.g., African-owned businesses, team-owned businesses, or home-based businesses.

A common mistake is the elimination of cases that do not belong to the subpopulation under study while carrying out the computation of descriptive and analytical statistics using the remainder cases. Given that the formation of a subpopulation is unrelated to the sample design, the subpopulation sample size is a random variable.[22] To correctly calculate the variance of an estimate for subpopulation, we should take the randomness in the subpopulation sample size into account by including all the data in the analysis (West, Berglund, and Heeringa, 2008). The implications of ignoring the randomness in the subpopulation sample size will lead to underestimated standard errors. Most of the advanced complex survey software packages (e.g., SAS®, SPSS®, and Stata®) have the capacity to correctly carry out domain analysis.

To illustrate the importance of performing subpopulation analyses, Table 17 reports the mean and the standard errors for a range of owner and business characteristics in the survey population at the baseline for a restricted sample consisting of white-owned businesses as well as a subpopulation analysis of white-owned businesses.

In Table 17, we note that eliminating cases that do not belong to the subpopulation resulted in smaller standard errors and fewer design degrees of freedom. For studies interested in descriptive and analytical analysis of subpopulations, it is highly recommended to conduct domains analysis rather than using restricted samples.

[22] In very rare cases where a stratum is the subpopulation (domain has a fixed sample size), eliminate cases are not a problem.

Table 17

Analyses	Restricted Sample			Subpopulation		
	Mean	Std. Err.	DEFF	Mean	Std. Err.	DEFF
Male	0.681	0.00676	0.774	0.681	0.00767	0.997
Hours worked	41.954	0.48595	1.322	41.954	0.48605	1.323
Work experience	9.680	0.15737	1.310	9.680	0.15780	1.317
Number of employees	1.543	0.03096	1.323	1.543	0.03099	1.326
Competitive advantage	0.628	0.01020	1.335	0.628	0.01020	1.335
Have a patent	0.021	0.00286	1.175	0.021	0.00286	1.176
Assets						
-Less than $10000	0.389	0.01019	1.310	0.389	0.01021	1.314
-$10,000 to $100,000	0.408	0.01035	1.329	0.408	0.01035	1.328
-$100,000 or more	0.202	0.00843	1.318	0.202	0.00844	1.323
Home based business	0.494	0.01048	1.318	0.494	0.01049	1.321
Provide a service	0.882	0.00689	1.369	0.882	0.00690	1.372
Number of obs.		2,999			4,607	
Population size		45,451			68,214	
Number of strata		6			6	
Sub Population no. obs.					2,999	
Sub Population size					45,451	
Design df		2,993			4,601	

1.7. Which Weight to Use?

For some KFS users, the selection of the correct weights to use in analysis may seem confusing. The general guideline to determine the correct weights to use is that studies that aim at providing estimates at cross-sectional levels, compare aggregates over time, net (macro-level) changes from one follow-up to another, snap shot of a population at one point in time, or investigate association utilizing the cross-sectional weights. On the other hand, studies aim at measuring gross (micro-level) changes over time or investigating causation utilize the longitudinal weights.

The KFS is a multipurpose survey and the best way to determine which weight to use is to briefly talk about the sample designs over time. Kish (1965) and Kish (2004) are the best sources for researchers who would like to understand sample design in depth. Following Kish (1965, 2004), the KFS can be used for the following purposes:

1. Measuring current levels (also known snap shot / static / point in time / one-shot/single-period estimate). In this case we will use one follow-up survey. For example, if we are only using information collected during the baseline survey then we would use the baseline cross-section weights. Similarly, if we are only using the fifth survey, then we would use the fifth cross-section weights. Measuring current levels can use either cross-sectional weight or longitudinal weight; but the cross-sectional weights include many more cases.
2. Measuring net change (also known as macro/mean/external change). Net change refers to the difference of means (ratios, totals, proportions, etc.) between two

periods ($d = \bar{x}_t - \bar{x}_{t-1}$ *or* $d = \bar{x}_{t+4} - \bar{x}_t$). Repeated cross-sectional survey with observations overlapping is used to measure net change. Repeated survey follow birth cohort(s) over time, whereas panel surveys follow individuals over time. Each cross-sectional survey of KFS has a partial overlaps (P), for example some firms responded to both surveys while other responded to one survey only. Treating the KFS as a repeated cross-sectional surveys with partial overlaps ($0 < P < 1$) allow us to measure net change (net effect of all the changes). If we want to infer about how firm characteristics have changes (net change) between the first and fifth survey, then we would use the first and fifth survey cross sectional weights.

3. Measuring gross change (also known as micro/ individual/ internal change). Gross change refers to individual changes between two periods ($d_i = x_{it} - x_{it-1}$). Because panel surveys follow the same individuals over time, we can use it to estimate both gross and net changes. More importantly, revealing the gross changes behind a net change can be done only using panel survey. The KFS provides the longitudinal weights for panel data, thus If we want to infer about how firm characteristics have changes (gross change) across the four years between baseline and the fourth survey, then we would use the fourth (or seventh) survey longitudinal weights. Similarly, if we want to infer about how firm characteristics have changes (gross change) across the eight years between baseline and the seventh survey, then we would use the seventh survey longitudinal weights.

To illustrate the use of cross-sectional and longitudinal weights in the KFS and understand gross vs. net change, we will focus on studying total employment by small firms established in 2004. In order to put our analysis into the context of cross-sectional and longitudinal studies, we need to distinguish between two employment measures: point-in-time measure that can allow us to compare aggregates over time (net change) and dynamic measure that can allow us study the gross change (at the individual level) and the various components of gross change. Table 18 reports total employment in each year since 2004. The cross-sectional weights in each year is used calculate total employment. The total employment was increasing up to 2005 than start to decline up to 2010. The change in the number in sample reflect the population (respondents) as it exists at the time of a particular follow-up of the survey. Thus, net change in employment reflects the change in employment inflow, employment outflow and the change in population composition (as firms goes out of business, merged or sold). Since that the net change has too many inseparable components, cross-sectional analysis does not allow for studying change at the individual level between two or more time points.

Table 18

Year	Total Employment	Net Change in Total Employment	Number of firms
2004[a]	183,585		4,928
2005[b]	261,049	77,464	3,998
2006[c]	246,352	-14,696	3,390
2007[d]	223,169	-23,183	2,915
2008[e]	198,459	-24,710	2,606
2009[f]	180,807	-17,653	2,408
2010[g]	166,757	-14,050	2,126
2011[h]	173,036	6,279	2,007

[a] Calculated using WGT_FINAL_0
[b] Calculated using WGT_FINAL_1
[c] Calculated using WGT_FINAL_F2_2
[d] Calculated using WGT_FINAL_F3_3
[e] Calculated using WGT_FINAL_F4_4
[f] Calculated using WGT_FINAL_F5_5
[g] Calculated using WGT_FINAL_F5_5
[h] Calculated using WGT_FINAL_F5_5

Table 19 reports total employment in each year since 2004. The longitudinal weight for the seventh survey is used to calculate total employment. The total employment was increasing up to 2005 than start to decline. Table 19 shows the dynamics of annual change (gross) in employment, for example the creation of 67,769 net new jobs during 2005 was a result of generating 105,790 new jobs by expanding firms, termination of 21,987 jobs by contracting firms and a loss of 16,035 jobs by firms that closed (or sold).

Meanwhile the gross change of 13,863 job losses during 2009 was a result of generating 29,814 new jobs by expanding firms, termination of 32,280 jobs by contracting firms and a loss of 11,397 jobs by firms that closed (or sold).

Table 19

Year	Total Employment	Gross Change in Total Employment	Number of firms	Expanding	Contracting	Closing
2004	173,884		3,140			
2005	241,653	67,769	2,837	105,790	-21,987	-16,035
2006	228,872	-12,781	2,554	53,283	-36,928	-29,136
2007	212,230	-16,642	2,330	44,281	-35,659	-25,264
2008	184,898	-27,332	2,092	31,201	-39,835	-18,699
2009	171,035	-13,863	1,928	29,814	-32,280	-11,397
2010	167,731	-3,303	1,775	26,125	-22,659	-6,769
2011	163,990	-3,742	1,644	28,940	-21,464	-11,217

Further analysis could address questions like, what industries were the most jobs creators before and after the recent recession, what industries were creating new jobs during the recent recession, what determine the employment growth trajectory of

firms, and did the stimulus bills help create jobs by small businesses. These type of questions that involve studying change can only answered by using the panel (longitudinal weight) data.

Comparing the results in table 18 and 19 give us a clear picture how gross change using panel data revealing results undiscovered by net change using cross sectional data.

We will discuss many more examples of using the correct weights in chapters four, five and six.

1.8. Conclusion

Any data analysis of the KFS data should consider the design effects of sample weighting adjustments, finite population corrections, and stratification and their impact on the estimates and standard errors. If ignored, our estimates and their standard errors are likely to be wrong.

The previous analysis shows that stratification and finite population corrections affect the standard errors, confidence intervals, and significance tests, but they do not affect the points and coefficient estimates. Stratification and finite population corrections usually make standard errors smaller, thus ignoring them gives us a more conservative estimates of standard errors, which could be ok.

On the other hand, weighting affects population estimates, standard errors, confidence intervals, and significance tests.

It is important to point out that standard statistical procedures in software packages (e.g., SAS®, SPSS®, and Stata®) do not account for complex sample design, non-response adjustments, and other adjustments. The consequence of using standard statistical procedures (assume simple random sampling) will be a bias in both of the estimates and their variances, which can lead to incorrect inferences for all types of analyses. In order to estimate the variance and to conduct valid statistical inference from complex sample design, a specialized statistical procedure needs to be considered when performing the statistical analysis. Most current software packages (e.g., SAS®, SPSS®, and Stata®) have special procedures or modules to analyze complex sample survey data.

2.1. Preparing the KFS Data for Complex Sample Survey Analysis

Like with most survey data, the first requirements before starting an analysis of the KFS data is to become familiar with the KFS questionnaire underlying structure, data contents, variable definitions, data formats, simple skip patterns, complex skip patterns, section skip patterns, data missingness, and the richness of the data.

In this chapter, we will focus on the underlying structure of the KFS questionnaire, questionnaire skip patterns, and how to prepare the KFS data for complex sample survey analysis.

2.2. The KFS Questionnaire

Understanding the underlying structure of the KFS questionnaire helps to distinguish data missing due to item non-response (refused to answer) from data missing due to unit non-response (dropouts) and legitimate missing values. Item non-response occurs when certain questions in a survey are not answered by a respondent. Unit non-response takes place when an eligible business cannot be contacted or refuses to participate in a survey. Legitimate missing value occurs when a business is sold, is merged, has temporarily stopped operations, or has permanently went out of businesses.

The KFS questionnaire was designed by more than 20 technical advisors who have interest, expertise, and scholarship related to entrepreneurship. The KFS questionnaire is organized in major sections that provide information about businesses and up to ten active-owner-operators. In the KFS, an active owner-operator is defined as an owner who provides regular assistance or advice with day-to-day operations of the business, rather than providing only money or occasional operating assistance. The questionnaire has seven parts discussed below.

2.5.1. Section A: Introduction

This section was devoted to verify information about the business and to confirm if the business is still in operation. In section A, businesses that reported that they permanently stopped operations, temporarily stopped operations, merged, or were sold did not complete the survey. For businesses no longer in operation, the reasons for ceasing operations were collected.

2.5.2. Section B: Eligibility Screening

As discussed earlier, to ensure that a business qualified as a unit of the target population, inclusion and exclusion criteria must be used as business eligibility screeners. This section was included in the baseline survey as screening criteria to determine business eligibility to the KFS target population.

2.5.3. Section C: Business Characteristics

The business characteristics section covers information about a business's legal status, a business's primary industry, number of full- and part-time employees, number of owners, number of owner-operators, and the primary location of the business. In addition, this section collects the names of up to ten active owner-operators.

2.5.4. Section D: Strategy and Innovation

The strategy and innovation section collects information regarding whether the business is offering a product or a service, if the business has a competitive advantage, if the business owns or licenses any intellectual properties (patents, copyrights, or trademarks), if the business has any sales and the percentage distribution of such sales to individuals, government, and other businesses.

2.5.5. Section E: Business Organization and Human Resource Benefits

The business organization and human resources benefits section provides information about the number of employees involved in various functions (human resources, sales, marketing, executive, R&D, production, general administration and financial administration) in a business. Also, this section covers the types of benefits (health insurance, retirement plan, stock options, bonus plan, tuition reimbursement, paid vacation, paid sick days, flex time, etc.) offered to both full- and part-time employees.

2.5.6. Section F: Business Finances

The business finances section is the most comprehensive part of the survey; it covers all aspects of business financing at the owner and business level as well as the business financial statements. At the owner level, this section provides information about the sources (owners, spouses, parents, other individuals, venture capitalists, companies, government, etc.), amount of equity financing, information about the sources (personal credit cards, personal loans from a bank, business credit cards, personal loans from any family or friends, etc.), and amount of personal debt financing.

At the business level, the business finances section provides information about sources (credit cards, loans from a commercial bank, line of credit, loans from a non-bank financial institution, loans from any family or friends, and loans from any other individuals) and the amount of business debt financing, trade financing, revenue, expenses, net income, assets (cash, accounts receivable, inventory, equipment or machinery, land, buildings, and other structures, vehicles, other business owned property, and other assets), liabilities (accounts payable, pension benefits, and other liabilities), R&D spending, the purchase of new or used machinery or equipment, and the rental or lease of buildings and rental or lease of machinery or equipment.

2.5.7. Section G: Work Behaviors and Demographics of Owner(S)

The work behaviors and demographics of owner(s) section provides information about work behavior, ownership, and demographics of up to ten owner-operators of the business. The data in this section includes if the active owner-operators are paid employees of the business; average weekly hours worked by each active owner-operator in the business; years of experience that each active owner-operator had in the industry of the new business; number of new businesses that each active owner-operator had started; and age, gender, race, education level, and U.S. citizenship for each active owner-operator.

2.3. Skip Logic

The skip pattern in the KFS comes into two major types; question skip logic and a section skip logic. Question skip logic involves conditionally asking/skipping questions based upon responses to prior question(s). Meanwhile, section skip logic involves asking / skipping a whole section in the survey questionnaire based upon responses to prior question(s).

The primary goal of understanding the skip patterns and the underlying structure of the questionnaire is to recognize non-applicable questions responses (hard missing or legitimate missing) and differentiate them from responses that are missing (soft missing) due to refusal or didn't know response. While it is common among most survey data to use user-defined missing values to identify why items are missing (e.g. for “refused” we assign a ".r" or for "not applicable" we assign a "999999"), this is not the case for the KFS. All missing responses are assigned a "." regardless of the reasons of missingness (soft or hard missing).

Having the missing values clearly defined (soft vs. hard missing) helps us to determine the correct number of observations for each variable, explains why number of observations varying from variable to variable, correctly construct aggregate variables, and correctly impute soft missing data.

2.4. Logical Imputation (Data Editing)

Logical (or deductive) imputation refers to any method that uniquely identified the true value of the missing value with certainty from within the dataset.

There are two major types of missing data in KFS: unit nonresponse and item nonresponse. Unit nonresponse occurs when a business refuse to participate in the survey. In the KFS the unit nonresponses are dealt with through weighting adjustments, thus it is not subject for logical imputation or any other kind of imputation. Meanwhile, item nonresponse occurs when certain questions in a survey are not answered by a respondent.

The single-cohort panel structure of the KFS offers many possibilities for logical imputation. To better understand what a deductive imputation procedure does, some examples are discussed:

For example, if the race information of an active-owner-operator was missing in follow-up t-1 then we keep asking about this missing information in the following surveys until we have a valid response. Thus, missing data for follow-up t-1 can be directly filled from other portions of an individual's record.

Another example of how to use logical imputation will be number of employees. During the first, second and third follow-ups, the respondents were asked:

C5. Not counting owner(s), on December 31, 2005, how many people worked for [NAME BUSINESS]? Please include all full- and part-time employees, but exclude contract workers who work for the business either full- or part-time but are not on the business' official payroll.

|__|__|__| Number Of Employees On December 31, 2005
Don't Know ..... "."
Refused ..... "."

C5b. Was this change an increase, a decrease, or no change in the number of people who worked for [NAME BUSINESS] on December 31, 2005 compared to December 31, 2004?

Increase 01
Decrease 02
No Change 03
Don't Know ..... "."
Refused ..... "." → GO TO C6

C5c. And what was the (increase/decrease) in the number of people who worked for [NAME BUSINESS] on December 31, 2005 compared to December 31, 2004? Your best estimate is fine.

|__|__|__| Change In Number Of Employees
Don't Know ..... "."
Refused ..... "."

If all the information for C5, C5b, and C5c are available for the 2005 follow-up and C5 is missing for 2004, then logical imputation can help determine the missing value for C5 in 2004.

2.5. Recoding Soft and Hard Missing values using Stata®

In this section, we will track the skip patterns for the core set of questions in each section of the KFS questionnaire based on the first follow-up questionnaire. A complete data set file that was subject to recoding missing values into soft and hard missing values and implementing logical imputations for the baseline and follow-up surveys is

available at NORC under the name "KFS8_LI.dta." To better understand what a deductive imputation procedure does, some examples are discussed in the following subsections.

Variables in the KFS data file have a suffix that corresponds to a round of data collection. Baseline variables contain a _0 suffix, First Follow-Up variables contain a _1 suffix, Second Follow-Up variables contain a _2 suffix, Third Follow-Up variables contain a _3 suffix, Fourth Follow-Up variables contain a _4 suffix, Fifth Follow-Up variables contain a _5 suffix, Sixth Follow-Up variables contain a _6 suffix, and Seventh Follow-Up variables contain a _7 suffix. In following section, we drop the suffix from the variables names for the core set of questions asked by all businesses in every follow-up survey.

To determine legitimate non-responses, the KFS data has a variable—final_status_code—that allows us to determine if the business is still in operation (complete the survey), dropped out, permanently stopped operations, temporarily stopped operations, merged, or was sold. Based on the final_status_code variable, we created a new variable—class_f—for every follow-up; the class_f variable has the following values and labels:

Value	Label
0	Located : No data was collected (due to skip logic)
1	Dropout (unit non-response) during follow-up t.
2	Missing because the business closed, sold or merged in the previous follow-ups.
3	Permanently stopped operations during follow-up t.
4	Merged, or sold during follow-up t.
5	Temporarily stopped operations during follow-up t.
6	Survival business(complete the survey)during follow-up t.

```
use KFS8_LI,clear
forvalues s = 0/7 {
tab classf_`s'
}

/* Cross sectional */
     Classification of Business |
              Status ( numeric ) |      Freq.     Percent        Cum.
--------------------------------+-----------------------------------
                        Complete |      4,928      100.00      100.00
--------------------------------+-----------------------------------
                           Total |      4,928      100.00
     Classification of Business |
              Status ( numeric ) |      Freq.     Percent        Cum.
--------------------------------+-----------------------------------
                         Refusal |        561       11.38       11.38
                Out of Business |        260        5.28       16.66
                  Merged Or Sold |         43        0.87       17.53
              Temporarily Stopped |         66        1.34       18.87
                        Complete |      3,998       81.13      100.00
--------------------------------+-----------------------------------
```

```
                          Total |      4,928      100.00
     Classification of Business |
             Status ( numeric ) |      Freq.     Percent        Cum.
--------------------------------+-----------------------------------
                        Refusal |        743       15.08       15.08
              Hard Missing Value |        303        6.15       21.23
                 Out of Business |        321        6.51       27.74
                  Merged Or Sold |         47        0.95       28.69
              Temporarily Stopped |        124        2.52       31.21
                        Complete |      3,390       68.79      100.00
--------------------------------+-----------------------------------
                          Total |      4,928      100.00



     Classification of Business |
             Status ( numeric ) |      Freq.     Percent        Cum.
--------------------------------+-----------------------------------
Located : No data was collected |         75        1.52        1.52
                        Refusal |        825       16.74       18.26
              Hard Missing Value |        671       13.62       31.88
                 Out of Business |        299        6.07       37.95
                  Merged Or Sold |         45        0.91       38.86
              Temporarily Stopped |         98        1.99       40.85
                        Complete |      2,915       59.15      100.00
--------------------------------+-----------------------------------
                          Total |      4,928      100.00
     Classification of Business |
             Status ( numeric ) |      Freq.     Percent        Cum.
--------------------------------+-----------------------------------
Located : No data was collected |         49        0.99        0.99
                        Refusal |        816       16.56       17.55
              Hard Missing Value |      1,015       20.60       38.15
                 Out of Business |        344        6.98       45.13
                  Merged Or Sold |         40        0.81       45.94
              Temporarily Stopped |         58        1.18       47.12
                        Complete |      2,606       52.88      100.00
--------------------------------+-----------------------------------
                          Total |      4,928      100.00
     Classification of Business |
             Status ( numeric ) |      Freq.     Percent        Cum.
--------------------------------+-----------------------------------
Located : No data was collected |         51        1.03        1.03
                        Refusal |        743       15.08       16.11
              Hard Missing Value |      1,399       28.39       44.50
                 Out of Business |        250        5.07       49.57
                  Merged Or Sold |         36        0.73       50.30
              Temporarily Stopped |         41        0.83       51.14
                        Complete |      2,408       48.86      100.00
--------------------------------+-----------------------------------
                          Total |      4,928      100.00
     Classification of Business |
             Status ( numeric ) |      Freq.     Percent        Cum.
--------------------------------+-----------------------------------
Located : No data was collected |         40        0.81        0.81
                        Refusal |        776       15.75       16.56
              Hard Missing Value |      1,685       34.19       50.75
                 Out of Business |        218        4.42       55.17
                  Merged Or Sold |         38        0.77       55.95
              Temporarily Stopped |         45        0.91       56.86
                        Complete |      2,126       43.14      100.00
```

```
--------------------------------+----------------------------------
                          Total |      4,928      100.00

     Classification of Business |
             Status ( numeric ) |      Freq.     Percent        Cum.
--------------------------------+----------------------------------
Located : No data was collected |         25        0.51        0.51
                        Refusal |        676       13.72       14.22
             Hard Missing Value |      1,941       39.39       53.61
                Out of Business |        209        4.24       57.85
                 Merged Or Sold |         40        0.81       58.66
              Temporarily Stopped |         30        0.61       59.27
                        Complete |      2,007       40.73      100.00
--------------------------------+----------------------------------
                          Total |      4,928      100.00


stset  Duration, failure(event==1)

     failure event:  event == 1
obs. time interval:  (0, Duration]
 exit on or before:  failure

------------------------------------------------------------------------------
     4928  total obs.
        0  exclusions
------------------------------------------------------------------------------
     4928  obs. remaining, representing
     2190  failures in single record/single failure data
    27568  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =         8

ltable  Duration event, survival

                Beg.                                    Std.
   Interval     Total   Deaths   Lost    Survival    Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
    1     2      4928      303    124     0.9377    0.0035     0.9306    0.9442
    2     3      4501      368     74     0.8604    0.0050     0.8503    0.8699
    3     4      4059      344     78     0.7868    0.0059     0.7749    0.7982
    4     5      3637      384     90     0.7027    0.0067     0.6894    0.7156
    5     6      3163      286     70     0.6384    0.0071     0.6244    0.6521
    6     7      2807      256    110     0.5790    0.0073     0.5646    0.5932
    7     8      2441      249    160     0.5180    0.0075     0.5032    0.5326
    8     9      2032        0   2032     0.5180    0.0075     0.5032    0.5326
```

2.7.1. Renaming, Recoding and Creating New Variables

To use loops and reshape data efficiently, the names of the variables need to be the same across all the follow-ups. To insure consistency of the variable names across all years, the following variables were renamed:

Old	New
fstatus_f2_2	fstatus_2
fstatus_f3_3	fstatus_3
fstatus_f4_4	fstatus_4
fstatus_f5_5	fstatus_5
fstatus_f6_6	fstatus_6
fstatus_f7_7	fstatus_7
b2a_legal_status_0	c1z2_legal_status_0
f2_owner_amt_eq_invest_allyrs_15	f2_ownr_amt_eqinvest_allyrs_15_5
f3a_xxxxxxxx	f3_xxxxxxxx
f3b_xxxxxxxx	f3_xxxxxxxx
f3c_xxxxxxxx	f3_xxxxxxxx
f3d_xxxxxxxx	f3_xxxxxxxx
f3e_xxxxxxxx	f3_xxxxxxxx
f3f_xxxxxxxx	f3_xxxxxxxx
f3g_xxxxxxxx	f3_xxxxxxxx
xx_2004_xx	xx_xx
xx_2005_xx	xx_xx
xx_2006_xx	xx_xx
xx_2007_xx	xx_xx
xx_2008_xx	xx_xx
xx_2009_xx	xx_xx
xx_2010_xx	xx_xx
xx_2011_xx	xx_xx
cswgt_final_0	wgt_final_0
cswgt_final_1	wgt_final_1
cswgt_final_2	wgt_final_f2_2
cswgt_final_3	wgt_final_f3_3
cswgt_final_4	wgt_final_f4_4
cswgt_final_5	wgt_final_f5_5
cswgt_final_6	wgt_final_f6_6
cswgt_final_7	wgt_final_f7_7
wgt_1_long	wgt_final_1
wgt_2_long	wgt_final_f12_long_2
wgt_3_long	wgt_final_f123_long_3
wgt_4_long	wgt_final_f1234_long_4
wgt_5_long	wgt_final_f5_long_5
wgt_6_long	wgt_final_f6_long_6
wgt_7_long	wgt_final_f7_long_7

For the race category questions, respondents were allowed to report multiracial or mixed-race. Thus, there is a question for each race. Because it is more practical for analysis purposes to have one variable with race coded as a categorical variable, we

created a new race variable "g6_race_group_xx_y" having the following codes/values:

American Indian or Alaska Native 01
Native Hawaiian or Other Pacific Islander 02
Asian 03
Black or African American 04
White 05
Other Races or Mixed-Race 06

The g6 questions about race are still in the data, but they have the values 1/0 (yes/no). For owners reporting multiracial or mixed-race, we recoded them as "Other Races or Mixed-Race."

The following table shows the variables that were subject to values recoding, as well as the old and new values.

Variable	New Values	Old Values
g3a_oth_bus_owner	5	6 to ∞
g6_race_amind_owner	0,1	1
g6_race_nathaw_owner	0,1	2
g6_race_asian_owner	0,1	3
g6_race_black_owner	0,1	4
g6_race_white_owner	0,1	5
g6_race_other_owner	0,1	6
g10_gender_owner	0	2
f23_profit_or_loss	0	2
f7b_pers_other_numused	5	6 to ∞
f9b_pers_other_numused	5	6 to ∞
f11b_bus_other_numused	5	6 to ∞
f7b_bus_credcard_numused	5	6 to ∞
f9b_bus_credcard_numused	5	6 to ∞
f7b_pers_credcard_numused	5	6 to ∞
f7b_pers_loan_fam_numused	5	6 to ∞
f9b_pers_credcard_numused	5	6 to ∞
f9b_pers_loan_fam_numused	5	6 to ∞
f11b_bus_credcard_numused	5	6 to ∞
f7b_pers_loan_bank_numused	5	6 to ∞
f9b_pers_loan_bank_numused	5	6 to ∞
f11b_bus_cred_line_numused	5	6 to ∞
f11b_bus_loans_emp_numused	5	6 to ∞
f11b_bus_loans_fam_numused	5	6 to ∞
f7b_pers_loan_other_numused	5	6 to ∞
f9b_pers_loan_other_numused	5	6 to ∞
f11b_bus_loans_bank_numused	5	6 to ∞
f11b_bus_loans_govt_numused	5	6 to ∞
f11b_bus_loans_owner_numused	5	6 to ∞
f11b_bus_loans_nonbank_numused	5	6 to ∞
f11b_busloans_otherind_numused	5	6 to ∞
f11a_busloans_otherbus_numused	5	6 to ∞

To use loops efficiently among all surveys, we created the following variables and we set their values to hard missing:

Variable Name	Added to	Variable Name	Added to
c10_morelocations	0,1	f12f_business_equip_veh	0,1,2,3,4
c11_num_locations	0,1	f12f_business_sec_dep	0,1,2,3,4
c12a_sba	0,1,2,3,5,6,7	f12f_intellectual_prop	0,1,2,3,4
c12b_fed_gov	0,1,2,3,5,6,7	f12f_inventory_acctrec	0,1,2,3,4
c12c_statelocal_gov	0,1,2,3,5,6,7	f12f_other	0,1,2,3,4
c12d_non_profit	0,1,2,3,5,6,7	f12f_other_pers_assets	0,1,2,3,4
c12e_college_univ	0,1,2,3,5,6,7	f12f_pers_real_estate	0,1,2,3,4
c12f_chamber_of_comm	0,1,2,3,5,6,7	f14d_new_loans	0,1,2
c12g_for_profit_org	0,1,2,3,5,6,7	f14e_approved_denied	0,1,2
c12h_other	0,1,2,3,5,6,7	f14f_bus_credit_hist	0,1,2
c9_loc_change_reason	0	f14f_inadeq_doc	0,1,2
d1_a_new_product	0,1,2,3,4	f14f_insuff_coll	0,1,2
d1_b_new_to_market	0,1,2,3,4	f14f_loan_toolarge	0,1,2
d1c_a_regional	0,1,2,3,4	f14f_new_bus	0,1,2
d1c_b_national	0,1,2,3,4	f14f_other	0,1,2
d1c_c_international	0,1,2,3,4	f14f_pers_credit_hist	0,1,2
d1d_new_processes	0,1,2,3,4	f14f_restr_on_lending	0,1,2,3
d2a_compadv_comp_reason	0,1,2	f14g_didnotapply	0,1,2
d2a_compadv_govlab_reason	0,1,2	f14h_loan_guarantees	0,1,2,3
d2a_compadv_patents_reason	0,1,2	f14i_economy_effect	0,1,2,3,5,6,7
d2a_compadv_univ_reason	0,1,2	f14j_most_challenging	0,1,2,3
d2b_compadv_comp_strength	0,1,2	f19a_res_dev_amt	0,1,2
d2b_compadv_govlab_strength	0,1,2	f19b_a_design	0,1,2,3
d2b_compadv_patents_strength	0,1,2	f19b_b_investments	0,1,2,3
d2b_compadv_univ_strength	0,1,2	f19b_c_brand_dev	0,1,2,3
d2c_compadv_cost_reason	0,1,2,3,4,6,7	f19b_d_org_dev	0,1,2,3
d2c_compadv_design_reason	0,1,2,3,4,6,7	f19b_e_worker_training	0,1,2,3
d2c_compadv_expertise_reason	0,1,2,3,4,6,7	f19b_f_other	0,1,2,3
d2c_compadv_marketing_reason	0,1,2,3,4,6,7	f19c_a_design_amt	0,1,2,3,4
d2c_compadv_price_reason	0,1,2,3,4,6,7	f19c_b_investments_amt	0,1,2,3,4
d2c_compadv_reputation_reason	0,1,2,3,4,6,7	f19c_c_brand_dev_amt	0,1,2,3,4
d2c_compadv_speed_reason	0,1,2,3,4,6,7	f19c_d_org_dev_amt	0,1,2,3,4
d2d_compadv_cost_strength	0,1,2,3,4,6,7	f19c_e_worker_training_amt	0,1,2,3,4
d2d_compadv_design_strength	0,1,2,3,4,6,7	f19c_f_other_amt	0,1,2,3,4
d2d_compadv_expertise_strength	0,1,2,3,4,6,7	f19c_intangassets_amt	0,1,2,3,5,6,7
d2d_compadv_marketing_strength	0,1,2,3,4,6,7	f32_chap11_bankruptcy	0,1,2,3
d2d_compadv_price_strength	0,1,2,3,4,6,7	f33_expected_growth	0,1,2,3,5,6,7
d2d_compadv_reput_strength	0,1,2,3,4,6,7	f34_future_revenue	0,1,2,3,5,6,7
d2d_compadv_speed_strength	0,1,2,3,4,6,7	f5a_seek_equity	0,1,2,3,4
d5a_founded_newprod	0,1,2,3,4,6,7	f6z_family_owned	0,1,2,3,5,6,7
d5b_a_personaluse	0,1,2,3,4,6,7	g10b_marital_status	0,1,2,3
d5b_b_previousjob	0,1,2,3,4,6,7	g1b2_reasonfor_business	0,1,2,3,4,5,6

Variable Name	Added to	Variable Name	Added to
d5b_c_startingbus	0,1,2,3,4,6,7	g10d_personal_outlook	0,1,2,3,5,6,7
d8_customer_locations	0,1,2	g10c_net_worth	0,1,2,3
d8a_international_sales	0,1,2	d9a_perc_internet_sales	0,1,2
d8b_perc_international_sales	0,1,2	f12e_collateral	0,1,2,3,4
d9_internet_sales	0,1,2		

For ease of running loops at owner level variables, all survey rounds have owner level variables for 15 owners.

2.7.2. Section C: Business Characteristics

As of the first follow-up, the KFS questionnaire asks the respondent to confirm legal status of the business, and then the legal status of the business is recorded under the "c1z2_legal_status" variable. In the baseline survey, legal status of the business is recorded under "b2a_legal_status_0." If the respondent confirms the legal status to be the same as in the previous year, the legal status of the business is copied from the previous year. Meanwhile if the respondent confirms that the legal status is not the same as the previous year, the respondent was asked to provide the new legal status of the business.

These types of questions do not have missing values due to skip logic but they have missing values due to legitimate missing values (business is sold, merged, temporarily stopped operations, or permanently out of businesses) as well as missing values due to item non-response.

C. BUSINESS CHARACTERISTICS

All of the following questions I'm going to ask are about [NAME BUSINESS]. Some of the questions will ask to confirm information about your business which you provided to us previously. As we go through the interview, please tell me if any of the information about your business is incorrect and needs to be updated.

C1z. Our records show that [NAME BUSINESS] had a legal status of [Read from file]. As of December 31, YYYY, is that still the legal status of [NAME BUSINESS]?

Yes .....1 → GO TO C1a
No.....0
Don't know ....."."
Refused ....."."

[c1z_confirm_legal_status]

C1z2. I'm going to read you a list of some different forms of legal status a business can have. As of December 31, YYYY, which form of legal status did [NAME BUSINESS] have? Was it a . . .

Sole Proprietorship, ... 1
Limited Liability Company, ... 2
Subchapter S-Corporation, ... 3
C-Corporation, ... 4
General Partnership, ... 5
Limited Partnership, or 6
Something else? (SPECIFY) 7

[c1z2_legal_status]

```
/* Rename  b2a_legal_status_0 to c1z2_legal_status_0
To use loops efficiently and to go over all the owners variables , we need to have
the names of the variables to be the same across all the follow-ups*/

use Kfs8_enclave_14oct13,clear
rename b2a_legal_status_0 c1z2_legal_status_0
misstable  sum c1z2*
```

Variable	Obs=.	Obs>.	Obs<.	Unique values	Min	Max
c1z2_legal~1	930		3,998	7	1	7
c1z2_legal~2	1,459		3,469	7	1	7
c1z2_legal~3	1,957		2,971	7	1	7
c1z2_legal~4	2,274		2,654	7	1	7
c1z2_legal~5	2,478		2,450	7	1	7
c1z2_legal~6	2,769		2,159	7	1	7
c1z2_legal~7	2,885		2,043	7	1	7

The misstable command makes a tables that helps to understand the missing values in our data. In the output table above, the labels refer to the following types of missing data:

"Obs=." are counts of system missing values (soft missing)

"Obs>." are counts of extended missing values (hard missing)

"Obs<." are counts of nonmissing values

"Unique values" specifies the number of unique values of the variables

Next, we need to recode the "c1z2_legal_status" variable into hard and soft missing values. In Stata® there are 26 different missing values—namely, .a, .b, ... , .z—that can be used to designate hard missing values ("Obs>.").[1] We will be using ".a" for all hard missing values in the KFS.

[1] In Stata® missing values are ordered as nonmissing < . < .a < .b < .. < .z

```
/* Loops are used to do repetitive tasks. Stata has commands that allow looping
over sequences of various types of lists */
gen c1z_confirm_legal_status_0=1
global  suffix "_0 _1 _2 _3 _4 _5 _6 _7"
foreach fup in $suffix{
/* Recode legitimate missing values */
replace c1z2_legal_status`fup'       =.a if classf`fup'<6
replace c1z_confirm_legal_status`fup'=.a if classf`fup'<6
}
misstable  sum c1z2*
                                                              Obs<.

                                          +------------------------------
              |                           | Unique
     Variable |     Obs=.     Obs>.     Obs<.  | values        Min         Max
  ------------+---------------------------------+------------------------------
  c1z2_legal~1 |                 930     3,998  |      7          1           7
  c1z2_legal~2 |               1,538     3,390  |      7          1           7
  c1z2_legal~3 |         6     2,013     2,909  |      7          1           7
  c1z2_legal~4 |         4     2,322     2,602  |      7          1           7
  c1z2_legal~5 |         9     2,520     2,399  |      7          1           7
  c1z2_legal~6 |         7     2,802     2,119  |      7          1           7
  c1z2_legal~7 |         6     2,921     2,001  |      7          1           7
  ------------------------------------------------------------------------------
```

Recoding missing values into soft missing (".") and hard missing (".a") discloses that very few observations represent true missing values.

In longitudinal surveys, last observation carried forward (LOCF) and last observation carried backward (LOCB) imputation methods can be utilized. Both LOCF and LOCB methods can be used in longitudinal research designs, but they require the strong assumption of stability. LOCF takes into account the individual's previous observed value on a given variable. If an observation at a certain data collection wave is missing, the last observed value is then used as an estimate for this missing observation. A related method, LOCB, works according to the same approach, but imputes a newer observation in the case of a missing earlier observation of the same individual.

For items that are stable over time, the value of a nonmissing item is substituted from one time period to another where the same item is missing. Direct substitution can be a highly accurate form of imputation in some situations.

```
KFS_Stability_Test c1z2_legal_status

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
c1z2_lega~01 |      3998    .9614807    .1924702          0          1
c1z2_lega~12 |      3282    .9667885    .1792156          0          1
c1z2_lega~23 |      2731    .9740022    .1591578          0          1
c1z2_lega~34 |      2413    .9776212    .1479427          0          1
c1z2_lega~45 |      2233    .9793999    .1420731          0          1
c1z2_lega~56 |      2012    .9811133    .1361587          0          1
c1z2_lega~67 |      1843    .9869778    .1134004          0          1
```

The stability test indicates that direct substitution can be used for legal status variables.

C1a. As of December 31, 2004, our records indicate the principal activity of the business was [D&B NAICS CODE DESCRIPTION OR LAST YEAR'S OTHER SPECIFY TO THIS QUESTION]. Was that still the principal activity of the business as of December 31, YYYY?

Yes ..... 1 → GO TO C2
No 0
Don't know ..... "."
Refused ..... "."

[c1a_naics_verification]

C2. As of December 31, YYYY, how many individuals or entities owned [NAME BUSINESS]? Please include all individuals or entities who owned shares in the business.

___ Number of owners
Don't know ..... "."
Refused ..... "."

[c2_owners]

C3. Of the [NUMBER OF OWNERS FROM C2] owners as of December 31, YYYY, how many owners actively helped to run [NAME BUSINESS]? By helped to run the business we mean that they provided regular assistance or advice with day-to-day operations of the business, rather than providing only money or occasional operating assistance.

___ Number of owner/operators
Don't know ..... "."
Refused ..... "."

[c3a_owner_ operators]

IF ONE OWNER/OPERATOR REPORTED AT C3, GO TO C5. IF MORE THAN ONE OWNER/OPERATOR REPORTED AT C3, ASK:

C4. FOR EACH OWNER/OPERATOR WHO IS NOT THE RESPONDENT, ASK:

I want to record with you the first and last names of these owners.

INTERVIEWER: ENTER FIRST AND LAST NAMES.
C4Confirm.

NOTE: UP TO 10 OWNER/OPERATORS WILL BE INCLUDED.
[c4_numowners_confirm]

```
foreach fup in $suffix{
/* Recode legitimate missing values */
replace c1a_naics_verification`fup'=.a if classf`fup'< 6
replace c2_owners`fup'              =.a if classf`fup'< 6
replace c3a_owner_operators`fup'    =.a if classf`fup'< 6
replace c4_numowners_confirm`fup'   =.a if classf`fup'< 6
}
misstable  sum c2_* c3a_* c4_*
/* Output omitted */
```

In section C, the KFS questionnaire asks about the total number of employees, excluding owner(s), who are paid employees of the business. Because of skip logic, data for "c6_num_ft_employees" and "c7_num_pt_employees" has to be recoded to zero if "c5_num_employees" is zero.

C5. Not counting owner(s), on December 31, YYYY, how many people worked for [NAME BUSINESS]? Please include all full- and part-time employees, but exclude contract workers who work for the business either full- or part-time but are not on the business' official payroll.

___ Number of December 31th YYYY
Don't know ..... "."
Refused ..... "."

[c5_num_employees]

IF BUSINESS REPORTED "0" EMPLOYEES AT C5, GO TO C8.

C6. . . . And of those [NUMBER FROM C5], how many were full-time? (IF NEEDED: Full-time is considered 35 hours or more per week)

___ Number of December 31th YYYY
Don't know ..... "."
Refused ..... "."

[c6_num_ft_employees]

C7. . . . And how many were part-time? (IF NEEDED: Part-time is considered less than 35 hours per week)

___ Number of December 31th YYYY
Don't know ..... "."
Refused ..... "."

[c7_num_pt_employees]

```
foreach fup in $suffix{
/* Recode due to Skip Patterns */
replace c6_num_ft_employees`fup'=0 if c5_num_employees`fup'==0
replace c7_num_pt_employees`fup'=0 if c5_num_employees`fup'==0
/* Recode legitimate missing values */
replace c5_num_employees`fup'   =.a if classf`fup'<6
replace c6_num_ft_employees`fup'=.a if classf`fup'<6
replace c7_num_pt_employees`fup'=.a if classf`fup'<6
}
misstable sum  c5_*  c6_* c7_*
/* Output omitted */
```

C8z. Our records show that the primary location where [NAME BUSINESS] operates is [Primary Location]. Is that correct?

Yes	01	→ GO TO D1
No	00	
Don't know	"."	→ GO TO D1
Refused	"."	

[c8z_primary_loc_confirm]

C8. How would you describe the primary location where [NAME BUSINESS] operates? Is it . . .

A residence such as a home or garage,	01
A rented or leased space,	02
Space the business purchased,	03
A site where a client is located, or	04
Some other location? (SPECIFY)	05
Don't know	"."
Refused	"."

[c8_primary_loc]

```
foreach fup in $suffix{
/* Recode legitimate missing values */
replace c8_primary_loc`fup'=.a if classf`fup'<6
}
misstable sum  c8_*
/* Output omitted */
```

2.7.3. Section D: Strategy and Innovation

All variables in section D should be recoded for legitimate missing values. An examination of the questions' skip logic indicates that the following recoding is needed:

1. "d3_a_num_patent," "d3_b_num_copyright" and "d3_c_num_trademark" need to be recoded to zero if the answers for "d3_a_have_patent", "d3_b_have_copyright," or "d3_c_have_trademark" is no, respectively.

2. "d4_a_lic_out_patent," "d4_b_lic_out_copyright," and "d4_c_lic_out_trademark" need to be recoded to zero if the answer for "d3_a_have_patent," "d3_b_have_copyright," or "d3_c_have_trademark" is No, respectively.

3. "d7_perc_sales_indiv," "d7_perc_sales_bus," and "d7_perc_sales_govt" need to be recoded to zero if the answer for "d6_have_sales" is No.

D. STRATEGY AND INNOVATION

D1. Does [NAME BUSINESS] provide (READ ITEM)?

	Yes	No	Don't Know	Refused
a. A service ..... [d1a_provide_service]	1	0	.	.
b. A product..... [d1a_provide_product]	1	0	.	.

D2. Businesses often have to compete with other businesses. A competitive advantage is something unique or distinctive a business provides that gives it an advantage compared to competitors. In calendar year YYYY, did [NAME BUSINESS] have a competitive advantage over its competitors?

Yes..... 01
No..... 00
Don't know ..... "."
Refused ..... "."

[d2_comp_advantage]

D3. Whether assigned by an owner or obtained in some other way, does [NAME BUSINESS] have any of the following? (READ LIST)

FOR EACH "YES," ASK: How many (READ ITEM) does [NAME BUSINESS] have?

INTERVIEWERS IF NEEDED:

Patent: A patent is a right given by the government to preclude others from making and selling an invention for 20 years from the date of application in return for disclosure of how the invention operates.

Copyright: The legal right granted to authors, composers, artists and publishers to protect their thoughts and ideas for exclusive publication, reproduction, sale and distribution of their works.

Trademark: Words, names, symbols or devices, or any combination of these used to identify the goods of a business and to distinguish these goods from the goods of others.

	Yes	No	Don't Know	Refused	NUMBER BUSINESS HAS
a. Patents ..... [d3_a_have_patent]	1	0	.	.	___ [d3_a_num_patent]
b. Copyrights ..... [d3_b_have_copyright]	1	0	.	.	___ [d3_b_num_copyright]
c. Trademarks..... [d3_c_have_trademark]	1	0	.	.	___ [d3_c_num_trademark]

D4. "Licensing out" is licensing patents, copyrights, or trademarks owned by the business to other parties under a licensing agreement. In calendar year YYYY, did [NAME BUSINESS] license out any (READ ITEM)?

	Yes	No	Don't Know	Refused
a. Patents ..... [d4_a_lic_out_patent]	1	0	.	.
b. Copyrights ..... [d4_b_lic_out_copyright]	1	0	.	.
c. Trademarks..... [d4_c_lic_out_trademark]	1	0	.	.

D5."Licensing in" is acquiring the right to use intellectual property such as patents, copyrights, or trademarks created by someone outside the business through a licensing agreement. In calendar year YYYY, did [NAME BUSINESS] license in any (READ ITEM)?

	Yes	No	Don't Know	Refused
a. Patents [d5_a_lic_in_patent]	1	0	.	.
b. Copyrights [d5_b_lic_in_copyright]	1	0	.	.
c. Trademarks [d5_c_lic_in_trademark]	1	0	.	.

D6.Did [NAME BUSINESS] have any customers or sales in calendar year YYYY?

Yes 01
No .....00 → GO TO E1
Don't know ....."."
Refused ....."."

[d6_have_sales]

D7. I'd like to learn more about the type of customers that [NAME BUSINESS] had during calendar year YYYY. I am going to ask you to estimate the percent of the business' sales that were made to individuals, businesses, and government agencies. The total should equal 100%.

a. During calendar year YYYY, what percentages of the business' sales were to private individuals?

___ Percentage
Don't know ....."."
Refused ....."."

[d7_perc_sales_indiv]

b. What percentages of the business' sales were to other businesses? [IF NEEDED: Please include sales to for-profit and not-for-profit business.]

___ Percentage
Don't know ....."."
Refused ....."."

[d7_perc_sales_bus]

c. What percentages of the business' sales were to government agencies?

___ Percentage
Don't know ....."."
Refused ....."."

[d7_perc_sales_govt]

```
foreach fup in $suffix{
replace d3_a_have_patent`fup'       =.a if classf`fup'<6
replace d3_a_num_patent`fup'        =.a if classf`fup'<6
replace d4_a_lic_out_patent`fup'    =.a if classf`fup'<6
replace d5_a_lic_in_patent`fup'     =.a if classf`fup'<6
}
foreach fup in $suffix{
replace d3_a_num_patent`fup'         = 0 if d3_a_have_patent`fup'  ==0
replace d4_a_lic_out_patent`fup'     = 0 if d3_a_have_patent`fup'  ==0
}
```

2.7.4. Section E: Business Organization and Human Resource Benefits

All variables in section E should be recoded for legitimate missing values. Section E has both question skip logic and section skip logic. The entire section was skipped for businesses that have one owner (c2_owners) and reported zero employees (c5_num_employees). Meanwhile, the part-time employee benefits questions were skipped for businesses that reported the number of part-time employees is zero.

E. BUSINESS ORGANIZATION AND HR BENEFITS

IF ONE OWNER REPORTED AT C2 AND BUSINESS REPORTED "0" EMPLOYEES AT C5, GO TO F1.

Next, I'd like to ask about how [NAME BUSINESS] is organized and about the benefits that are offered to employees.

E1. On December 31, YYYY, how many employees or owners, if any, did [NAME BUSINESS] have who were primarily responsible for (READ ITEM)? Please include only full- and part-time employees, but not contract workers who work for the business but are not on the business' official payroll.

	Number Employees Or Owners	Don't Know	Refused
a. Human resources such as employee benefits, recruitment, or hiring ..... [e1_a_num_human_res]	___	.	.
b. Sales or Marketing such as sales, market research, customer analysis, or promotional activities ..... [e1_b_num_sales]	___	.	.
c. Executive administration functions such as strategic planning, competitive analysis, shareholder relations, or general management ..... [e1_c_num_exec_admin]	___	.	.
d. Research and development on new products or services [e1_d_num_resdev]	___	.	.
e. Production or manufacturing such as producing materials or products, production planning, production control, quality control, or storage..... [e1_e_num_prod_manu]	___	.	.
f. General administration such as office management, responding to maintenance requests, purchase supplies, or training employees in office procedures [e1_f_num_gen_admin]	___	.	.
g. Financial administration such as accounting procedures, budgeting, financial analysis, or investment activities [e1_g_num_fin_admin]	___	.	.
h. Does [NAME BUSINESS] have employees with any other key responsibilities? (Specify) [e1_h_num_other]	___	.	.

E2a. As of December 31, YYYY, did [NAME BUSINESS] offer full-time employees or owners (READ ITEM):

	Yes	No	Don't Know	Refused
a. A health insurance plan either through the business or an association..... [e2a_ft_emp_hlth_plan]	1	0	.	.
b. A retirement plan such as profit sharing, pension, including 401K, annuity, Keogh, etc. ..... [e2a_ft_emp_retire_plan]	1	0	.	.
c. Stock options or other stock ownership..... [e2a_ft_emp_stock_own]	1	0	.	.
d. A bonus plan ..... [e2a_ft_emp_bonus_plan]	1	0	.	.
e. Tuition reimbursement..... [e2a_ft_emp_tuit_reim]	1	0	.	.
f. Paid vacation ..... [e2a_ft_emp_paid_vaca]	1	0	.	.
g. Paid sick days ..... [e2a_ft_emp_paid_sick]	1	0	.	.
h. Alternative work schedules such as flex time or job sharing..... [e2a_ft_emp_flex_time]	1	0	.	.
i. Any other benefits? (SPECIFY)..... [e2a_ft_emp_other]	1	0	.	.

IF ZERO PART-TIME EMPLOYEES AT C7, GO TO F1.

E2b. As of December 31, YYYY, did [NAME BUSINESS] offer part-time employees (READ ITEM):

	Yes	No	Don't Know	Refused
a. A health insurance plan either through the business or an association [e2b_pt_emp_hlth_plan]	1	0	.	.
b. A retirement plan such as profit sharing, pension, including 401K, annuity, Keogh, etc. [e2b_pt_emp_retire_plan]	1	0	.	.
c. Stock options or other stock ownership [e2b_pt_emp_stock_own]	1	0	.	.
d. A bonus plan [e2b_pt_emp_bonus_plan]	1	0	.	.
e. Tuition reimbursement [e2b_pt_emp_tuit_reim]	1	0	.	.
f. Paid vacation [e2b_pt_emp_paid_vaca]	1	0	.	.
g. Paid sick days [e2b_pt_emp_paid_sick]	1	0	.	.
h. Alternative work schedules such as flex time or job sharing [e2b_pt_emp_flex_time]	1	0	.	.
i. Any other benefits? (Specify) [e2b_pt_emp_other]	1	0	.	.

```
foreach fup in $suffix{
gen skip_e`fup'=1 if c2_owners`fup'== 1  & c5_num_employees`fup'== 0
replace e1_a_num_human_res`fup'    =.a     if skip_e`fup'==1
replace e2a_ft_emp_bonus_plan`fup' =.a     if skip_e`fup'==1
replace e2b_pt_emp_bonus_plan`fup' =.a     if c7_num_pt_employees`fup'  == 0
replace e1_a_num_human_res`fup'    =.a     if classf`fup'   <6
replace e2a_ft_emp_bonus_plan`fup' =.a     if classf`fup'   <6
replace e2b_pt_emp_bonus_plan`fup' =.a     if classf`fup'   <6
}
```

2.7.5. Section F: Business Finances

Section F deals with the major sources of financing—namely, equity, debt, and other financial information of the business. Because KFS collects data for up to ten active owner-operators, each owner was assigned a number. For all variables that are related to active owner-operators, the number ("_0,_1,_2,_3,_4,_5,_6,_7") prior to the suffixes indicates the number assigned to the owner. For example, "f2_owner_amt_eq_invest_02_1," refers to the equity injection by owner number two in the first follow-up, while "f2_owner_amt_eq_invest_09_1" refers to the equity injections by owner number nine in the first follow-up. Unless indicated, the variables in this document are listed without a suffix if the variable name is the same across all rounds.

In the baseline survey, the respondent was always owner number one. Because the respondent could change from one follow-up to the next, starting from the first follow-up the variable "respondent" contains the number of the owner who responded for the business in a particular follow-up.

While the KFS collects data for up to ten active owner-operators, the number assigned to the owner can be more than ten. Because some owners who used to be active (non-active) in one follow-up could be non-active (active) in another, the number assigned to the owner can be more than ten.

Starting from the first follow-up and to identify active owner-operators in each follow-up survey, a variable called "owner_active_(owner-number)" was created to ensure that users could see which owner was still an active owner-operator in the business.

For all the financial variables in the KFS, if the respondent did not provide the exact amount of the variable in dollars, the respondent was asked to provide a range of the amount instead. The range interval classes were standard across all the financial variables in the KFS. The interval classes were:

$0.....00
$500 or less,.....01
$501 to $1,000,.....02
$1,001 to $3,000,.....03
$3,001 to $5,000,.....04
$5,001 to $10,000,.....05
$10,001 to $25,000,.....06
$25,001 to $100,000,.....07
$100,001 to $1,000,000,.....08
$1,000,001 or more?.....09
Don't Know ..... "."
Refused..... "."

Based on the exact amount of the variable or the range of the amount provided by the respondent, new variables were constructed for ease of analysis. The constructed variables represent the financial variables in terms of range interval classes by translating the exact amount into a range. To distinguish these variables from the range variables, the term "_r_" was included into the variable's name.

2.5.5.1. Equity Injections by the Active-Owner-Operators

In every survey, the respondents were asked about their equity injections into the business in that year (indicator question) and the amount that was injected, if any. Starting from the first follow-up, respondents were asked to provide how much equity they injected into the business in all years.

For businesses that had more than one owner, equity injections by other active owner-operators, up to nine of them, were collected through the respondents. The respondents were asked about the equity injections into the business by each active owner-operator in that year (yearly inflow) and the amount that was injected, if any. Starting from the first follow-up, respondents were asked to provide how much equity each of other active owner-operators injected and if they obtained equity financing during follow-up into the business in in all years.

The amount of equity injections needed to be recoded to zero if the active owner-operator(s) stated that he/they did not inject equity into the business in that follow-up.

Recoding hard missing values is required for two types of missing values. First, all variables in section F should be recoded for legitimate missing values. Second, the variables for non-active owner-operators should be recoded to hard missing values.

F. BUSINESS FINANCES

F1. Now I'd like to ask about [NAME BUSINESS]'s financing. Businesses can get money from the savings or investments of the owner(s), money from spouses, family or other individuals, from companies, borrowing in an owner's name, venture funds, or by borrowing in the name of the business. Some of the funds must be paid back and other funds represent an equity stake or share of the business. We will ask some questions about what happened during calendar year YYYY, some questions about what has happened since the business began, and other questions about balances as of December 31, YYYY.

F1a. First, in calendar year YYYY, did you put any of your own money into [NAME BUSINESS] in return for an ownership share of the business? Please do not include any money borrowed from others or from credit cards.
IF NEEDED: This would include all additional money invested by [you/OWNER NAME] in the business during calendar year YYYY.

Yes 01
No.....00
Don't know....."."
Refused....."."
[f2_owner_eq_ invest]

F2a. IF YES: How much of [your/her/his] own money did [you/he/she] put into the business during calendar year YYYY? IF NEEDED: Your best estimate is fine.

OWNER A..... $ ___
Don't know....."."
Refused....."."
[f2_owner_amt_eq_invest]

IF DON'T KNOW OR REFUSED, PROBE WITH RANGES.

F2a1. Counting all years, how much of [your/OWNER B-J] own money did [you/she/he] put into [NAME BUSINESS] as of December 31, YYYY? → GO TO F2a1
IF NEEDED: This includes all money [you/she/he] invested in the business as of December 31, YYYY.

$ ___ Total Equity
Don't know....."."
Refused....."."
[f2_ownr_amt_eqinvest_allyrs]

PROBE: IF DON'T KNOW OR REFUSED, PROBE WITH RANGES.
[f2_ownr_eqinvst_allyrs_range]

IF F2a IS GREATER THAN F2a1: I may have made a mistake. The amount invested in YYYY is greater than the amount invested in all years combined. Is there an error?

F2b. What percentage of the business did [you/OWNER B-J] own on December 31, YYYY?

___ Percentage Of Business
Don't know....."."
Refused....."."
[f2_owner_perc_own_01]

IF MORE THAN ONE OWNER/OPERATOR AT C4, ASK F1a-F2b FOR EACH WNER/OPERATOR. OTHERWISE, GO TO F3.

F2 series asked of up to 10 owner-operators.

```
global owners_1_15 "01 02 03 04 05 06 07 08 09 10 11 12 13 14 15"

forvalues i = 0/7 {
foreach ow in $owners_1_15  {
replace f2_owner_eq_invest_`ow'_`i'     =.a if owner_active_`ow'_`i'    !=1
}
}
```

```
forvalues i = 0/7 {
foreach ow in $owners_1_15  {
replace f2_owner_amt_eq_invest_`ow'_`i' =0  if f2_owner_eq_invest_`ow'_`i'==0
}
}
```

```
forvalues i = 0/7 {
foreach ow in $owners_1_15  {
replace f2_owner_eq_invest_`ow'_`i'         =.a      if classf_`i'<6
replace f2_owner_amt_eq_invest_`ow'_`i'     =.a      if classf_`i'<6
}
}
```

2.5.5.2. Equity Injections by Other Owners

In every survey, the respondents were asked if the business obtained equity financing from owners who were not actively involved in operating the business, non-operator-owners, and the amount that was obtained, if any. The balance of each source of funding that was used during follow-up t was collected in every follow-up survey. Data collections for equity financing obtained from non-operator-owners were at the aggregate level.

The amount of equity injections needs to be recoded to zero if the non-active-owner-operator(s) state that he/they did not inject equity into the business in that follow-up.

Because the equity injection by other owners is not applicable for the businesses that reported the legal status as sole proprietorship, the variables for sole proprietorship need to recode to hard missing values.

IF SOLE PROPRIETORSHIP AT QUESTION C1z OR C1z2, GO TO F6b.

F3. Equity investment is money received in return for some portion of ownership, and it is another way to fund business expenses. During calendar year YYYY, did the business obtain equity financing from any of the following sources?

	Yes	No	Don't Know	REFUSED
a. Spouses or life partners of owners of the business. This does not include spouses or life partners already named as owners [f3a_eq_invest_spouse]	1	0	.	.
b. Parents, in-laws or children of owners of the business [f3b_eq_invest_parents]	1	0	.	.
c. Individuals who are not spouses or life partners, parents, in-laws or children of the owners, excluding venture capitalists [f3c_eq_invest_angels]	1	0	.	.
d. Other companies [f3d_eq_invest_companies]	1	0	.	.
e. Government agencies [f3e_eq_invest_govt]	1	0	.	.
f. Venture capitalists [f3f_eq_invest_vent_cap]	1	0	.	.
g. Any other sources? (SPECIFY) [f3g_eq_invest_other]	1	0	.	.

F4. FOR EACH EQUITY FINANCING OPTION REPORTED AS "YES" ABOVE, ASK:

In calendar year YYYY, how much money did [NAME BUSINESS] receive from [EQUITY OPTION]?

$ ___ Amount From Equity Option
Don't know ..... "."
Refused ..... "."

[f4_eq_amt_spouse]
[f4_eq_amt_parents]
[f4_eq_amt_angels]
[f4_eq_amt_companies]
[f4_eq_amt_govt
,[f4_eq_amt_vent_cap]
[f4_eq_amt_other]

IF DON'T KNOW OR REFUSED, PROBE WITH RANGES.

F4a. Counting all years, how much did [EQUITY OPTION] put into [NAME BUSINESS] as of December 31, YYYY?

PROBE: This includes all money invested by [EQUITY OPTION] in all years.
$ ___ Total Equity
Don't know ..... "."
Refused ..... "."
[f4_eq_amt_spouse_allyrs]
[f4_eq_amt_parents_allyrs]
[f4_eq_amt_angels_allyrs]
[f4_eq_amt_companies_allyrs]
[f4_eq_amt_govt_allyrs]
[f4_eq_amt_vent_cap_allyrs]
[f4_eq_amt_other_allyrs]

PROBE: IF DON'T KNOW OR REFUSED, PROBE WITH RANGES.
IF F4 IS GREATER THAN F4a:
I may have made a mistake. The amount invested in YYYY is greater than the amount invested in all years combined. Is there an error?

F5. FOR EACH EQUITY FINANCING OPTION REPORTED AS "YES" ABOVE, ASK:

What was the total percentage of the business owned by the [EQUITY OPTION] who invested money in the business as of December 31, YYYY?
___ Percent
Don't know ..... "."
Refused ..... "."
[f5_perc_owned_spouse] [f5_perc_owned_parents] [f5_perc_owned_angels] [f5_perc_owned_companies] [f5_perc_owned_govt] [f5_perc_owned_vent_cap] [f5_perc_owned_other]

```
global List1 "spouse parents angels companies govt vent_cap other"

forvalues i = 0/7 {
foreach name in $List1  {
replace f3_eq_invest_`name'_`i'     =.a   if   c1z2_legal_status_`i'==1
replace f4_eq_amt_`name'_`i'        =.a   if   c1z2_legal_status_`i'==1
replace f4_eq_amt_`name'_`i'        = 0   if   f3_eq_invest_`name'_`i'==0
replace f3_eq_invest_`name'_`i'     =.a   if       classf_`i'<6
replace f4_eq_amt_`name'_`i'        =.a   if       classf_`i'<6
}
}
```

2.5.5.3. Cash Withdrawals by Owners

In every follow-up survey, respondents were asked if any of the owners' withdrew money from the business for personal use and, if yes, how much was withdrawn. This does not include salaries paid to owners who are full-time employee of the business. For the baseline, this question was asked in the first follow-up.

The amount withdrawn needs to be recoded to zero if any of the owners state that they did not withdrawn money from the business in that follow-up.

INTERVIEWER CHECK BOX: CHECK ANSWER FROM F2b AND F5 FOR TOTAL PERCENTAGE OF BUSINESS ACCOUNTED FOR.
[f6check]
IF TOTAL PERCENTAGE EQUALS 100%, GO TO F6a
IF TOTAL EQUALS LESS OR MORE THAN 100%
F6.So far, you've given me the following information on who owns [NAME BUSINESS]: [LIST EQUITY INVESTORS FROM F2b AND F5]. Can we review this list?

REVIEW LIST OF OWNERS AND PERCENTAGES WITH RESPONDENT. MAKE CHANGES AS NEEDED, ADDING NEW OWNERS AND/OR PERCENTAGES AS NECESSARY.

Don't know ..... "."
Refused ..... "."

[f6_perc_owned_owner]

F6a. Have you or other owner's withdrawn money from the business for personal use in YYYY?
INTERVIEWER: IF NEEDED: This does not include owner salaries.

Yes 01
No 00
Don't know ..... "." → GO TO F7a
Refused ..... "."

[f6a_personal_use]

F6b. IF YES: Now, thinking about calendar year YYYY, how much money, if any, did you and other owners withdraw from the business for personal use? This includes any dividends paid.

$ ___ Total Drawings YYYY
Don't know ..... "."
Refused ..... "."

[f6b_personal_use_amt]

IF DON'T KNOW OR REFUSED, PROBE WITH RANGES.

```
/* For the baseline: owners withdrawn money from the business for personal use in
2004 were asked the question in 2005*/
gen     f6a_personal_use_0=.
replace f6a_personal_use_0=1 if   tot_personal_use_r_0<.
replace f6a_personal_use_0=0 if   tot_personal_use_r_0==0
replace f6a_personal_use_0=0 if   f6a_personal_use_1  ==0
forvalues i = 0/7 {
replace f6a_personal_use_`i'      =.a     if classf_`i'< 6
replace f6b_personal_use_amt_`i'=.a     if classf_`i'< 6
replace tot_personal_use_r_`i'  =.a     if classf_`i'< 6
replace f6b_personal_use_amt_`i'=0 if f6a_personal_use_`i'==0
replace tot_personal_use_r_`i'  =0 if f6a_personal_use_`i'==0
}
```

2.5.5.4. Personal Debt Obtained by the Respondent

Respondents were asked about all types of personal debt that was obtained in their names on behalf of the business and how much was obtained, if any, during follow-up t. In addition to the amount obtained every year for each type of personal debt, the amount owed for each type of personal debt used in follow-up t was collected.

The amount of personal debt needs to be recoded to zero if the respondent states that they did not use that source of funding. Also, the number of personal debt used needs to be recoded to zero if the respondent states that they did not use that source of funding.

F7a. Another way to finance a business is debt financing. Debt is money borrowed that has to be paid back with or without interest.

We will be talking about categories of debt based on who is responsible for paying it back. For each category, I'll ask you about several sources of debt business owners or businesses can use to fund operations. We want to make sure that any business-related debt is reported in the right category, and is reported only once. I will identify each category and remind you when I change categories. Here is the first category.
I'm going to ask you about some different types of debt financing you may have borrowed in your name on behalf of [NAME BUSINESS]. For each, please tell me if you used this type at any time during calendar year YYYY. Did you use [NAME FINANCING OPTION FROM LIST]?

F7b. IN BELOW LIST, *FOR EACH DEBT FINANCING OPTION BUSINESS REPORTED, ASK*: How many [NAME DEBT FINANCING OPTION] did you use to finance the operation of the business during calendar year YYYY?

	Yes	No	Don't Know	Refused	Number Used
a. Personal credit cards for business-related purposes..... [f7a_pers_credcard]	1	0	.	.	___ [f7b_pers_credcard_numused]
b. Personal loans from a bank or other financial institution, such as a mortgage or home equity loan used for the business ..... [f7a_pers_loan_bank]	1	0	.	.	___ [f7b_pers_loan_ bank_numused]
c. Business or corporate credit cards issued in your name..... [f7a_bus_credcard]	1	0	.	.	___ [f7b_bus_credcard_numused]
d. Personal loans from any family or friends ..... [f7a_pers_loan_fam]	1	0	.	.	___ [f7b_pers_loan_fam_numused]
e. Personal loans from any other individuals not associated with the management of the business..... [f7a_pers_loan_other]	1	0	.	.	___ [f7b_pers_loan_other_numused]
f. Any other sources? (SPECIFY)..... [f7a_pers_other]	1	0	.	.	___ [f7b_pers_other_numused]

F8a. *IF ANSWERED "YES" TO F7a ITEMS a, c, ASK*: As of December 31, YYYY, what was the maximum credit line on the [NAME DEBT FINANCING OPTION]?

$_____ December 31, YYYY Credit Line

Don't know..... "."

Refused..... "."

[f8a_pers_credcard_line], [f8a_bus_credcard_line]

IF DON'T KNOW OR REFUSED, PROBE WITH RANGES.

F8b. *IF ANSWERED "YES" TO F7a ITEMS a, c, ASK*: As of December 31, YYYY, what was the outstanding balance on the [NAME DEBT FINANCING OPTION]?

$ ___ December 31, YYYY Outstanding Credit Card Balance
Don't know ....."."
Refused ....."."

[f8b_pers_credcard_bal]
[f8b_bus_credcard_bal]

IF DON'T KNOW OR REFUSED, PROBE WITH RANGES.

IF F8b IS GREATER THAN F8a: Perhaps I made a mistake. The amount I recorded as the balance outstanding is greater than the amount reported as the maximum credit limit.

F8c. *IF ANSWERED "YES" TO F7a ITEMS b, d, e, f, ASK*: In calendar year YYYY, how much was obtained from the [NAME DEBT FINANCING OPTION]?

$ ___ Calendar Year YYYY Debt Amount
Don't know ....."."
Refused ....."."

[f8c_pers_loan_bank_amt]
[f8c_pers_loan_fam_amt]
[f8c_pers_loan_other_amt]
[f8c_pers_other_amt]

IF DON'T KNOW OR REFUSED, PROBE WITH RANGES

F8d. As of December 31, YYYY, what was the estimated amount of the [NAME DEBT FINANCING OPTION] owed by you on behalf of [NAME BUSINESS]?

$ _____ Debt Financing Value As Of December 31, YYYY
Don't know ....."."
Refused ....."."

[f8d_pers_loan_bank_owed]
[f8d_pers_loan_fam_owed]
[f8d_pers_loan_other_owed]
[f8d_pers_other_owed]

IF DON'T KNOW OR REFUSED, PROBE WITH RANGES.

```
forvalues i = 0/7 {

replace f7b_pers_loan_bank_numused_`i' =0  if f7a_pers_loan_bank_`i' ==0
replace f8c_pers_loan_bank_amt_`i'     =0  if f7a_pers_loan_bank_`i' ==0
replace f7a_pers_loan_bank_`i'         =.a if classf_`i'<6
replace f7b_pers_loan_bank_numused_`i' =.a if classf_`i'<6
}
```

2.5.5.5. Personal Debt Obtained by the Other Owners

For businesses that have more than one active owner-operator, the respondents were asked to report all types of personal debt obtained by all other owners on behalf of the business and how much was obtained, if any. Unlike equity financing by other owners, personal debt by other owners was collected at the aggregate level for active owner-operators only. In addition to the amount of personal debt obtained by active owner-operators in every year for each type of personal debt, the amount owed for each type of personal debt used in follow-up t was collected.

Because the personal debt obtained by the other owners is not applicable for the businesses that reported having one active owner-operator, the variables for businesses that reported having one active owner-operator need to be recoded to hard missing values.

The amount of personal debt obtained by the other owners needs to be recoded to zero if the respondent states that they did not use that source of funding. Also, the number of personal debt used by the other owners needs to be recoded to zero if the respondent states that they did not use that source of funding

IF MORE THAN ONE OWNER/OPERATOR AT C4, ASK F9a. OTHERWISE, GO TO F11a.
F9a. Here is the next debt category. I'm going to ask you about some different types of debt financing that other owners may have borrowed on behalf of [NAME BUSINESS]. This debt does not include amounts already reported in the previous section about your debt. For each, please tell me if other owners used this type at any time during calendar year YYYY. Did other owners use [NAME DEBT FINANCING OPTION FROM LIST]?

F9b.IN BELOW LIST, *FOR EACH DEBT FINANCING OPTION BUSINESS REPORTED, ASK*: How many [NAME DEBT FINANCING OPTION] did other owners use to finance the operation of the business during calendar year YYYY?

	Yes	No	Don't Know	Refused	Number Used
a. Personal credit cards for business-related purposes..... [f9a_pers_credcard]	1	0	.	.	__ [f9b_pers_credcard_numused]
b. Personal loans from a bank or other financial institution, such as a mortgage or home equity loan used for the business..... [f9a_pers_loan_bank]	1	0	.	.	__ [f9b_pers_loan_bank_numused]
c. Business or corporate credit cards issued in the other owner's name(s) ..... [f9a_bus_credcard]	1	0	.	.	__ [f9b_bus_credcard_numused]
d. Personal loans from any family or friends..... [f9a_pers_loan_fam]	1	0	.	.	__ [f9b_pers_loan_fam_numused]
e. Personal loans from any other individuals not associated with the management of the business ..... [f9a_pers_loan_other]	1	0	.	.	__ [f9b_pers_loan_other_numused]
f. Any other sources? (Specify)..... [f9a_pers_other]	1	0	.	.	__ [f9b_pers_other_numused]

F10a. *IF ANSWERED "YES" TO F9a ITEMS a, c, ASK*: As of December 31, YYYY, what was the maximum credit line on the [NAME DEBT FINANCING OPTION] of (one of) the other owner(s)?

$ ___ December 31, YYYY Credit Line
Don't know ..... "."
Refused ..... "."

[f10a_pers_credcard_line] [f10a_bus_credcard_line]

IF DON'T KNOW OR REFUSED, PROBE WITH RANGES.

F10b. IF ANSWERED "YES" TO F9a ITEMS a, c, ASK: As of December 31, YYYY, what was the outstanding balance on the [NAME DEBT FINANCING OPTION] used by (one of) the other owner(s)?

$ ___ December 31, YYYY Outstanding Credit Card Balance
Don't know ..... "."
Refused ..... "."

[f10b_pers_credcard_bal] [f10b_bus_credcard_bal]
IF DON'T KNOW OR REFUSED, PROBE WITH RANGES.
IF F10b IS GREATER THAN F10a:
Perhaps I made a mistake. The amount I recorded as the balance outstanding is greater than the amount reported as the maximum credit limit.

F10c.*IF ANSWERED "YES" TO F9a, ITEMS b, d, e, f, ASK*: In calendar year YYYY, how much was obtained from the [NAME DEBT FINANCING OPTION] other owners used?

$ ___ Calendar Year YYYY Debt Amount
Don't know....."."
Refused....."."

[f10c_pers_loan_bank_amt] [f10c_pers_loan_fam_amt] [f10c_pers_loan_other_amt] [f10c_pers_other_amt]

IF DON'T KNOW OR REFUSED, PROBE WITH RANGES.

F10d. As of December 31, YYYY, what was the estimated amount of the [NAME DEBT FINANCING OPTION] owed by other owners on behalf of [NAME BUSINESS]?

$ ___ Debt Amount As Of December 31, YYYY
Don't know....."."
Refused....."."

[f10d_pers_loan_bank_owed] [f10d_pers_loan_fam_owed] [f10d_pers_loan_other_owed]
[f10d_pers_other_owed]

IF DON'T KNOW OR REFUSED, PROBE WITH RANGES.

```
forvalues i = 0/7 {
replace f9a_pers_loan_bank_`i'        =.a if c4_numowners_confirm_`i'<2
replace f9b_pers_loan_bank_numused_`i'=.a if c4_numowners_confirm_`i'<2
replace f10c_pers_loan_bank_amt_`i'   =.a if c4_numowners_confirm_`i'<2
replace f10c_pers_loan_bank_amt_`i'   =0  if f9a_pers_loan_bank_`i' ==0
replace f9b_pers_loan_bank_numused_`i'=0  if f9a_pers_loan_bank_`i' ==0
replace f9a_pers_loan_bank_`i'        =.a if classf_`i'<6
replace f9b_pers_loan_bank_numused_`i'=.a if classf_`i'<6
replace f10c_pers_loan_bank_amt_`i'   =.a if classf_`i'<6
}
```

2.5.5.6. Debt Obtained by the Business

In addition to personal debt financing, the KFS collects data about different types of debt financing obtained in the name of the business during baseline and each follow-up survey.

The amount of debt obtained by the business needs to be recoded to zero if the respondent states that the business did not use that source of funding. Also, the number of business debt used needs to be recoded to zero if the respondent states that the business did not use that source of funding.

F11a.We are once again switching to another debt category. Now I'm going to ask you about some different types of debt financing that may have been obtained in the name of the business during calendar year YYYY. This debt does not include amounts already reported in the previous sections about your debt or the debt of other owners. During calendar year YYYY, did [NAME BUSINESS] use [NAME DEBT FINANCING OPTION FROM LIST]?

F11b.IN BELOW LIST, *FOR EACH DEBT FINANCING OPTION BUSINESS REPORTED, ASK*: How many [NAME DEBT FINANCING OPTION] did the business use to finance the operation or the business during calendar year YYYY?

	Yes	No	Don't Know	Refused	Number Used
a. Business or corporate credit cards issued in the name of the business [f11a_bus_credcard]	1	0	.	.	__ [f11b_bus_credcard_numused]
b. Business loans from a commercial bank [f11a_bus_loans_bank]	1	0	.	.	__ [f11b_bus_loans_bank_numused]
c. Business line of credit (READ IF NEEDED: a business line of credit is when a business has an agreement with a bank or other financial institution to borrow up to a certain amount of funds) [f11a_bus_cred_line]	1	0	.	.	__ [f11b_bus_cred_line_numused]
d. Business loans from a non-bank financial institution [f11a_bus_loans_nonbank]	1	0	.	.	__ [f11b_bus_loans_nonbank_numused]
e. Business loans from any family or friends of the owners [f11a_bus_loans_fam]	1	0	.	.	__ [f11b_bus_loans_fam_numused]
f. Business loans from another owner of the business or a partner	1	0	.	.	__ [f11b_bus_loans_owner_numused]

[f11a_bus_loans_owner]					
g. [*IF HAVE EMPLOYEES AT C5*] Loans to the business from employees that are not owners of the business	1	0	.	.	___
[f11a_bus_loans_emp]					[f11b_bus_loans_emp_numused]
h. Loans from government agencies	1	0	.	.	___
[f11a_bus_loans_govt]					[f11b_bus_loans_govt_numused]
i. Loans from other businesses			.	.	
[F11a_Bus_Loans_Other_Bus_1]	1	0			___
[f11a_bus_loans_other_bus]					[f11a_busloans_otherbus_numused]
j. Business loans from any other individuals not associated with the management of the business	1	0	.	.	___
[f11a_bus_loans_other_ind]					[f11b_busloans_otherind_numused]
k. Any other sources? (SPECIFY)	1	0	.	.	___
[f11a_bus_other]					[f11b_bus_other_numused]

F12a. *IF ANSWERED "YES" TO F11a ITEMS a, c, ASK*: As of December 31, YYYY, what was the maximum credit line on the [NAME DEBT FINANCING OPTION]?

$ ____ December 31, YYYY Credit Line
Don't know ..... "."
Refused ..... "."

[f12a_bus_credcard_line] [f12a_bus_cred_line]

IF DON'T KNOW OR REFUSED, PROBE WITH RANGES.

F12b. *IF ANSWERED "YES" TO F11a ITEMS a, c, ASK*: As of December 31, YYYY, what was the outstanding balance on the [NAME DEBT FINANCING OPTION]?

$ _____ December 31, YYYY Outstanding Credit Balance
Don't know ..... "."
Refused ..... "."

[f12b_bus_credcard_bal] [f12b_bus_cred_line_bal]

IF DON'T KNOW OR REFUSED, PROBE WITH RANGES.

IF F12a IS GREATER THAN F12b:
Perhaps I made a mistake. The amount I recorded as the balance outstanding is greater than the amount reported as the maximum credit limit.

F12c. *IF ANSWERED "YES" TO F11a ITEMS b, d-k, ASK*: In calendar year YYYY, how much was the amount obtained from [NAME DEBT FINANCING OPTION] used by [NAME BUSINESS]?

$ ____ Calendar Year YYYY Debt Amount
Don't know ..... "."
Refused ..... "."

[f12c_bus_loans_bank_amt],[f12c_bus_loans_nonbank_amt],[f12c_bus_loans_fam_amt],[f12c_bus_loans_owner_amt],[f12c_bus_loans_emp_amt],[f12c_bus_loans_govt_amt],[f12c_bus_loans_bus_amt], [f12c_bus_loans_other_ind_amt], [f12c_bus_other_amt]
IF DON'T KNOW OR REFUSED, PROBE WITH RANGES.

F12d. As of December 31, YYYY, what was the estimated amount of the [NAME DEBT FINANCING OPTION] owed by [NAME BUSINESS]?

$ ___ Debt Amount As Of December 31, YYYY
Don't know..... "."
Refused..... "."

[f12d_bus_loans_bank_owed],[f12d_bus_loans_nonbank_owed],[f12d_bus_loans_fam_owed],[f12d_bus_loans_owner_owed],[f12d_bus_loans_emp_owed],[f12d_bus_loans_govt_owed],[f12d_bus_loans_bus_owed],[f12d_bus_loans_other_ind_owed],[f12d_bus_other_owed]

IF DON'T KNOW OR REFUSED, PROBE WITH RANGES

```
forvalues i = 0/7 {
replace f11a_bus_loans_bank_`i'         =.a if classf_`i'<6
replace f12c_bus_loans_bank_amt_`i'     =.a if classf_`i'<6
replace f11b_bus_loans_bank_numused_`i'=.a if classf_`i'<6
replace f11b_bus_loans_bank_numused_`i'=0  if f11a_bus_loans_bank_`i'==0
replace f12c_bus_loans_bank_amt_`i'     =0  if f11a_bus_loans_bank_`i'==0
replace f11a_bus_loans_emp_`i'          =.a if c5_num_employees_`i'==0
replace f11a_bus_loans_owner_`i'        =.a if c2_owners_`i'==1
}
```

2.5.5.7. Other Financial Information

In addition to the sources of financing, the KFS collects other financial information from the balance sheet and income statement as well as financial information regarding the existence of R&D and rental or lease.

F13. Trade financing is where a business has an arrangement with a supplier to make purchases on account. In calendar year YYYY, did [NAME BUSINESS] make any purchases through trade financing?

Yes 01
No 00
Don't know ..... "."
Refused ..... "."

→ GO TO F15

[f13_trade_fin]

F14. *IF YES*: In calendar year YYYY, what was the amount of purchases made through trade financing?

$ ___ Calendar year YYYY amount of trade purchases
Don't know ..... "."
Refused ..... "."

[f14a_trade_fin_amt]

IF DON'T KNOW OR REFUSED, PROBE WITH RANGES.

F15. In calendar year YYYY, did [NAME BUSINESS] receive any revenue (money), from the sales of goods, services, or intellectual property? [IF SOLE PROPRIETORSHIP, ADD: This would be gross receipts reported on a Schedule C or C-EZ with your personal income tax return.]

Yes 01
No 00
Don't know ..... "."
Refused ..... "."

→ GO TO F17

[f15_revenue]

F16. What was [NAME BUSINESS]'s total revenue for calendar year YYYY? [IF SOLE PROPRIETORSHIP, ADD: This would be gross receipts reported on a Schedule C or C-EZ with your personal income tax return.]

$ ___ Total revenue YYYY
Don't know ..... "."
Refused ..... "."

[f16a_rev_yyyy_amt]
IF DON'T KNOW OR REFUSED, PROBE WITH RANGES.

F17. Now I'm going to ask about the expenses the business paid. Expenses are the costs paid for the operation of the business, including wages, salaries, interest on loans, capital leases, materials, etc. How much, if any, did [NAME BUSINESS] pay in expenses during calendar year YYYY?

$ ___ Total expenses in calendar year YYYY
Don't know ....."."
Refused ....."."

[f17a_total_exp_yyyy_amt]

IF DON'T KNOW OR REFUSED, PROBE WITH RANGES.

F18. *IF BUSINESS REPORTED "0" EMPLOYEES AT C5, GO TO F19.*
How much, if any, did [NAME BUSINESS] pay in wages, salaries, and benefits to full-and part-time employees in calendar year YYYY? Please do not include wages, salaries, and benefits to contract workers who work for the business but are not on the business' official payroll.

$ ___ Total payroll expenses in calendar year YYYY
Don't know ....."."
Refused ....."."

[f18a_wage_exp_yyyy_amt]

IF DON'T KNOW OR REFUSED, PROBE WITH RANGES.

F19. Did [NAME BUSINESS] spend any money on research and development of new products and services during calendar year YYYY?

Yes 01
No.....00
Don't know ....."."
Refused ....."."

[f19_res_dev]

F20. Did [NAME BUSINESS] spend any money on the purchase of new or used machinery or equipment during calendar year YYYY?

Yes 01
No.....00
Don't know ....."."
Refused ....."."

[f20_mach]

F21. Did [NAME BUSINESS] spend any money on rental or lease payments for buildings or other structures during calendar year YYYY?

Yes 01
No 00
Don't know ..... "."
Refused ..... "."

[f21_land_rent]

F22. Did [NAME BUSINESS] spend any money on rental or lease payments for machinery or equipment during calendar year YYYY?

Yes 01
No 00
Don't know ..... "."
Refused ..... "."

[f22_mach_rent]

F23. Profit is the business' income after all expenses and taxes have been deducted. What was [NAME BUSINESS]'s total profit or loss for calendar year YYYY?

Profit ..... 1 → GO TO F24
Loss ..... 2 → GO TO F26
Don't know ..... "."
Refused ..... "."

[f23_profit_or_loss]

F24. ENTER PROFIT AMOUNT

$ ___ Total profit in calendar year YYYY
Don't k now ..... "."
Refused ..... "."

[f24_profit_amt]

F25. IF DON'T KNOW OR REFUSED, PROBE WITH RANGES

F26. ENTER LOSS AMOUNT

$ ___ Total loss in calendar year YYYY
Don't know ..... "."
Refused ..... "."

[f26_loss_amt]

F27. IF DON'T KNOW OR REFUSED, PROBE WITH RANGES

```
forvalues i = 0/7 {
/* Recode to Zero - due to Skip Patterns */
replace f14a_trade_fin_amt_`i'=0 if f13_trade_fin_`i'   ==0
replace f16a_rev_amt_`i'      =0 if f15_revenue_`i'     ==0
/* Recode legitimate missing values */
replace f13_trade_fin_`i'     =.a if classf_`i'<6
replace f14a_trade_fin_amt_`i'=.a if classf_`i'<6
replace f15_revenue_`i'       =.a if classf_`i'<6
replace f16a_rev_amt_`i'      =.a if classf_`i'<6
replace f17a_total_exp_amt_`i'=.a if classf_`i'<6
replace tot_trade_finan_r_`i' =.a if classf_`i'<6
replace tot_revenue_r_`i'     =.a if classf_`i'<6
replace tot_expenses_r_`i'    =.a if classf_`i'<6
replace tot_wages_r_`i'       =.a if classf_`i'<6

replace f18a_wage_exp_amt_`i'=.a if classf_`i'<6
replace f19_res_dev_`i'      =.a if classf_`i'<6
replace f20_mach_`i'         =.a if classf_`i'<6
replace f21_land_rent_`i'    =.a if classf_`i'<6
replace f22_mach_rent_`i'    =.a if classf_`i'<6
}
# delimit ;
misstable sum f14* f15* f16* f17* f18* f19* f20*
f21* f22* tot_revenue_r* tot_trade_finan_r* tot_expenses_r* tot_wages_r*;
/* Output omitted */
# delimit cr
forvalues i = 0/7 {
/* Recode legitimate missing values */
replace f24_profit_amt_`i'=.a if f23_profit_or_loss_`i' ==2
replace tot_profit_r_`i'  =.a if f23_profit_or_loss_`i' ==2
replace f26_loss_amt_`i'  =.a if f23_profit_or_loss_`i' ==1
replace tot_loss_r_`i'    =.a if f23_profit_or_loss_`i' ==1
/* Recode legitimate missing values */
replace f23_profit_or_loss_`i'=.a if classf_`i'<6
replace f24_profit_amt_`i'    =.a if classf_`i'<6
replace tot_profit_r_`i'      =.a if classf_`i'<6
replace f26_loss_amt_`i'      =.a if classf_`i'<6
replace tot_loss_r_`i'        =.a if classf_`i'<6
}
misstable sum f23* f24*  f26* tot_loss_r*  tot_profit_r*
/* Output omitted */
```

F28. Assets are what the business owns. As of December 31, YYYY, did [NAME BUSINESS]'s assets include [NAME ASSET FROM LIST]?

	Yes	No	Don't Know	Refused
a. Cash on hand in checking, savings, money market accounts, certificates of deposit and other time deposits..... [f28a_asset_cash]	1	0	.	.
b. Accounts receivable..... [f28b_asset_acct_rec]	1	0	.	.
c. Product inventory ..... [f28c_asset_inv]	1	0	.	.
d. Equipment or machinery..... [f28d_asset_equip]	1	0	.	.
e. Land, buildings, and other structures..... [f28e_asset_landbuild]	1	0	.	.
f. Vehicles ..... [f28f_asset_veh]	1	0	.	.
g. Any other business owned property ..... [f28g_other_bus_prop]	1	0	.	.
h. Any other assets? (Specify) ..... [f28h_other_assets]	1	0	.	.

F29. *FOR EACH ASSET BUSINESS REPORTED, ASK*:

As of December 31, YYYY, what was the estimated value of the [NAME OF ASSET] owned by [NAME BUSINESS]?

$ ___ Asset value as of December 31, YYYY
Don't know..... "."
Refused..... "."

[f29_assetval_cash],[f29_assetval_acctrec],[f29_assetval_inv]
[f29_assetval_equip],[f29_assetval_landbuild],[f29_assetval_veh]
[f29_assetval_othbusprop],[f29_assetval_other]

IF DON'T KNOW OR REFUSED, PROBE WITH RANGES.

F30. Liabilities are what the business owes. Other than the loans and the financial debt we've already talked about, did [NAME BUSINESS]'s liabilities as of December 31, YYYY include [NAME LIABILITY FROM LIST]?

	Yes	No	Don't Know	Refused
a. Accounts Payable ..... [f30a_liab_acctpay]	1	0	.	.
b. Pension and post-retirement benefits ..... [f30b_liab_pension]	1	0	.	.
c. Any other liabilities? (SPECIFY) ..... [f30c_liab_other]	1	0	.	.

F31. *FOR EACH LIABILITY BUSINESS HAS, ASK*: As of December 31, YYYY, what was the estimated value of [NAME BUSINESS]'s [NAME OF LIABILITY]?

$ ___ Liability value as of December 31, YYYY
Don't know ..... "."
Refused ..... "."

[f31_value_acctpay],[f31_value_pension],[f31_value_other]

IF DON'T KNOW OR REFUSED, PROBE WITH RANGES.

```
forvalues i = 0/6 {
/* Recode to Zero - due to Skip Patterns */
replace f29_assetval_cash_`i'      =0 if f28a_asset_cash_`i'     ==0
replace f29_assetval_acctrec_`i'   =0 if f28b_asset_acct_rec_`i' ==0
replace f29_assetval_inv_`i'       =0 if f28c_asset_inv_`i'      ==0
replace f29_assetval_equip_`i'     =0 if f28d_asset_equip_`i'    ==0
replace f29_assetval_landbuild_`i' =0 if f28e_asset_landbuild_`i'==0
replace f29_assetval_veh_`i'        =0 if f28f_asset_veh_`i'      ==0
replace f29_assetval_othbusprop_`i'=0 if f28g_other_bus_prop_`i' ==0
replace f29_assetval_other_`i'     =0 if f28h_other_assets_`i'   ==0
/* Recode legitimate missing values */
replace f28a_asset_cash_`i'     =.a  if classf_`i'< 6
replace f28b_asset_acct_rec_`i' =.a  if classf_`i'< 6
replace f28c_asset_inv_`i'      =.a  if classf_`i'< 6
replace f28c_asset_inv_`i'      =.a  if classf_`i'< 6
replace f28d_asset_equip_`i'    =.a  if classf_`i'< 6
replace f28e_asset_landbuild_`i'=.a  if classf_`i'< 6
replace f28f_asset_veh_`i'      =.a  if classf_`i'< 6
replace f28g_other_bus_prop_`i' =.a  if classf_`i'< 6
replace f28h_other_assets_`i'   =.a  if classf_`i'< 6
/* Recode legitimate missing values */
replace f29_assetval_cash_`i'      =.a if classf_`i'< 6
replace f29_assetval_acctrec_`i'   =.a if classf_`i'< 6
replace f29_assetval_inv_`i'       =.a if classf_`i'< 6
replace f29_assetval_equip_`i'     =.a if classf_`i'< 6
replace f29_assetval_landbuild_`i' =.a if classf_`i'< 6
replace f29_assetval_veh_`i'       =.a if classf_`i'< 6
```

```
replace f29_assetval_othbusprop_`i'=.a if classf_`i'< 6
replace f29_assetval_other_`i'     =.a if classf_`i'< 6
/* Recode to Zero - due to Skip Patterns */
replace f31_value_acctpay_`i'=0 if f30a_liab_acctpay_`i'==0
replace f31_value_pension_`i'=0 if f30b_liab_pension_`i'==0
replace f31_value_other_`i'  =0 if f30c_liab_other_`i'  ==0
/* Recode legitimate missing values */
replace f31_value_acctpay_`i'=.a if classf_`i'< 6
replace f31_value_pension_`i'=.a if classf_`i'< 6
replace f31_value_other_`i'  =.a if classf_`i'< 6
/* Recode legitimate missing values */
replace f30a_liab_acctpay_`i'=.a if classf_`i'< 6
replace f30b_liab_pension_`i'=.a if classf_`i'< 6
replace f30c_liab_other_`i'  =.a if classf_`i'< 6
}
```

2.7.6. Section G: Work Behaviors and Demographics of Active-Owner-Operators

Information regarding work behaviors by active owner-operators was collected in the baseline as well as in every follow-up survey. Meanwhile, demographics information was collected once. If the demographics information of an active owner-operator was collected during follow-up t-1, then no demographics information will be collected during follow-up t. If the demographics information of an active owner-operator was missing in follow-up t-1, then we continue asking about this missing information in the follow-up surveys until a valid response is received. During each follow-up, the work behaviors and demographics of new active owner-operators were collected.

The work behaviors and demographics variables for non-active owner-operators should be recoded to hard missing values.

G. WORK BEHAVIORS AND DEMOGRAPHICS OF OWNER/OPERATOR(S)

The last section contains questions for classification purposes only.

C4 LISTING OF OWNER/OPERATORS SHOULD BE ASKED THIS SERIES IN THE FOLLOWING ORDER:

RESPONDENT FIRST, THEN OTHER OWNER/OPERATORS, THEN NEW OWNER/OPERATORS.
NO QUESTIONS WILL BE ASKED ABOUT OWNER/OPERATORS WHO HAVE LEFT.

FOR ALL OWNER/OPERATORS IN C4, ASK *BLOCK bSectionG1*

FOR ALL NEW OWNER/OPERATORS, ASK *BLOCK bSectionG2*

BLOCK bSectionG1—

G1a. (Are/Is) (You/ [OWNER B-J]) also a paid employee at [NAME BUSINESS]?
Yes..... 01
No..... 00
Don't know ..... "."
Refused ..... "."
[g1a_emp_owner]

G1b. During the time [NAME BUSINESS] was in business during YYYY, how many hours in an average week did (you/[OWNER B-J]) spend working at [NAME BUSINESS]?
___ Hours worked in average week
Don't know ..... "."
Refused ..... "."
[g1b1_hours_owner]

IF DON'T KNOW OR REFUSED PROBE: Would you say it was . . .
Less than 20 hours, ... 01
20 hours to 35 hours, ... 02
36 hours to 45 hours, ... 03
46 hours to 55 hours, ... 04
56 hours to 65 hours, ... 05
66 hours or more? 06
Don't know ..... "."
Refused ..... "."

ANY DEMOGRAPHIC QUESTION G1d-G10a NOT ANSWERED IN PREVIOUS FOLLOW UP WILL BE ASKED AGAIN.

G1a. (Are/Is) (you/[OWNER B-J]) also a paid employee at [NAME BUSINESS]?
Yes..... 01
No..... 00
Don't know ..... "."
Refused ..... "."
[g1a_emp_owner]

G1b. During the time [NAME BUSINESS] was in business during YYYY, how many hours in an average week did (you [OWNER B-J] spend working at [NAME BUSINESS]?
___ Hours worked in average week
Don't know ..... "."
Refused ..... "."
[g1b1_hours_owner]

IF DON'T KNOW OR REFUSED, PROBE: Would you say it was . . .

Less than 20 hours,.....01
20 hours to 35 hours,.....02
36 hours to 45 hours,.....03
46 hours to 55 hours,.....04
56 hours to 65 hours,.....05
66 hours or more?.....06
Don't know.....“.”
Refused.....“.”

```
global owners_1_15 "01 02 03 04 05 06 07 08 09 10 11 12 13 14 15"


forvalues i = 0/7 {
foreach ow in $owners_1_15  {
/* Recode legitimate missing values */
replace g1a_emp_owner_`ow'_`i'        =.a if owner_active_`ow'_`i'!=1
replace g1b1_hours_owner_`ow'_`i'     =.a if owner_active_`ow'_`i'!=1
replace g1a_emp_owner_`ow'_`i'        =.a if classf_`i'< 6
replace g1b1_hours_owner_`ow'_`i'     =.a if classf_`i'< 6
}
}
```

G2. How many years of work experience (have/has) (you/[OWNER B-J]) had in this industry—the one in which [NAME BUSINESS] competes?

___ Years
Don't know.....“.”
Refused.....“.”

[g2_work_exp_owner]

G3a. How many other new businesses (have/has) (you/[OWNER B-J]) started besides [NAME BUSINESS]?

___ Number of businesses (Enter “0” for none)
Don't know.....“.”
Refused.....“.”

[g3a_oth_bus_owner]

IF ZERO NEW BUSINESSES AT G3a, GO TO G4.

G3b. (Was this/Were any of the) business(es) in the same industry as [NAME BUSINESS]?

Yes..... 01
No..... 00
Don't know ..... "."
Refused ..... "."

[g3b_bus_same_ind_owner]

G4. How old will (you/[OWNER B-J]) be on (your/his/her) next birthday?

___ Owner A
Don't know ..... "."
Refused ..... "."

[g4_age_owner]

IF DON'T KNOW OR REFUSED, PROBE WITH RANGES: Would you say . . .

18-24,..... 01
25-34,..... 02
35-44,..... 03
45-54,..... 04
55-64,..... 05
65-74,..... 06
75 or older? 07
Don't know ..... "."
Refused ..... "."

Now I have a few questions about race and ethnicity.

G5. (Are/Is) (you/[OWNER B-J]) of Hispanic or Latino origin?

Yes..... 01
No..... 00
Don't know ..... "."
Refused ..... "."

[g5_hisp_origin_owner]

```
/*  If  G2-G10 is missing in wave t read from wave t+1 */
forvalues i = 0/7 {
foreach ow in $owners_1_15  {
/* Recode to Zero - due to Skip Patterns */
replace      g3b_bus_same_ind_owner_`ow'_`i'=  0  if  g3a_oth_bus_owner_`ow'_`i'==0

}
}
local g2  "g2_work_exp_owner"
local g3  "g3a_oth_bus_owner g3b_bus_same_ind_owner"
local g5  "g5_hisp_origin_owner"
local g235 "`g2' `g3' `g5'"
```

```
foreach Q of local g235{
foreach ow in $owners_1_15{
/*last observation carried backward  */
replace `Q'_`ow'_0=`Q'_`ow'_1 if `Q'_`ow'_0==.
replace `Q'_`ow'_0=`Q'_`ow'_2 if `Q'_`ow'_0==.
replace `Q'_`ow'_0=`Q'_`ow'_3 if `Q'_`ow'_0==.
replace `Q'_`ow'_0=`Q'_`ow'_4 if `Q'_`ow'_0==.
replace `Q'_`ow'_0=`Q'_`ow'_5 if `Q'_`ow'_0==.
replace `Q'_`ow'_0=`Q'_`ow'_6 if `Q'_`ow'_0==.
}
}
foreach Q of local g235{
foreach ow in $owners_1_15{
replace `Q'_`ow'_1 =`Q'_`ow'_0 if `Q'_`ow'_1==.
replace `Q'_`ow'_2 =`Q'_`ow'_0 if `Q'_`ow'_2==.
replace `Q'_`ow'_3 =`Q'_`ow'_0 if `Q'_`ow'_3==.
replace `Q'_`ow'_4 =`Q'_`ow'_0 if `Q'_`ow'_4==.
replace `Q'_`ow'_5 =`Q'_`ow'_0 if `Q'_`ow'_5==.
replace `Q'_`ow'_6 =`Q'_`ow'_0 if `Q'_`ow'_6==.
}
}
forvalues i = 0/7 {
foreach Q of local g235 {
foreach ow in $owners_1_15 {
replace `Q'_`ow'_`i'=.a if owner_active_`ow'_`i'!=1
replace `Q'_`ow'_`i'=.a if classf_`i'< 6
}
}
}
foreach ow in $owners_1_15{
/*last observation carried backward  */
replace g4_age_owner_`ow'_0=g4_age_owner_`ow'_1-1    if g4_age_owner_`ow'_0==.
replace g4_age_owner_`ow'_0=g4_age_owner_`ow'_2-2    if g4_age_owner_`ow'_0==.
replace g4_age_owner_`ow'_0=g4_age_owner_`ow'_3-3    if g4_age_owner_`ow'_0==.
replace g4_age_owner_`ow'_0=g4_age_owner_`ow'_4-4    if g4_age_owner_`ow'_0==.
replace g4_age_owner_`ow'_0=g4_age_owner_`ow'_5-5    if g4_age_owner_`ow'_0==.
replace g4_age_owner_`ow'_0=g4_age_owner_`ow'_6-6    if g4_age_owner_`ow'_0==.
}
foreach ow in $owners_1_15{
replace g4_age_owner_`ow'_1 =g4_age_owner_`ow'_0+1 if g4_age_owner_`ow'_1==.
replace g4_age_owner_`ow'_2 =g4_age_owner_`ow'_0+2 if g4_age_owner_`ow'_2==.
replace g4_age_owner_`ow'_3 =g4_age_owner_`ow'_0+3 if g4_age_owner_`ow'_3==.
replace g4_age_owner_`ow'_4 =g4_age_owner_`ow'_0+4 if g4_age_owner_`ow'_4==.
replace g4_age_owner_`ow'_5 =g4_age_owner_`ow'_0+5 if g4_age_owner_`ow'_5==.
replace g4_age_owner_`ow'_6 =g4_age_owner_`ow'_0+6 if g4_age_owner_`ow'_6==.
}

forvalues i = 0/7 {
foreach ow in $owners_1_15  {
/* Recode legitimate missing values */
replace      g4_age_owner_`ow'_`i'=.a if owner_active_`ow'_`i'!=1
replace      g4_age_owner_`ow'_`i'=.a if classf_`i'< 6
}
}
```

For the race category questions, respondents were allowed to report multiracial or mixed-race. Thus, there is a question for each race. Because it is more practical for analysis purposes to have one variable with race coded as a categorical variable, we created a new race variable "g6_race_group_" having the following codes:

American Indian or Alaska Native 01
Native Hawaiian or Other Pacific Islander 02
Asian 03
Black or African American 04
White 05
Other Races or Mixed Race 06

Owners who reported multiracial or mixed-race were recoded as "Other Races or Mixed Race."

G6. I am going to read a list of race categories. Please choose one or more that best describes (your/[OWNER B-J]'s) race. Are (you [OWNER B-J]) American Indian or Alaska Native, Native Hawaiian or other Pacific Islander, Asian, Black or African American, or White?

American Indian Or Alaska Native 01
Native Hawaiian Or Other Pacific Islander 02
Asian 03
Black Or African American 04
White 05
Other (Specify) 06
Don't Know ..... "."
Refused ..... "."

[g6_race_amind_ owner]
[g6_race_nathaw_owner]
[g6_race_asian_owner]
[g6_race_black_owner]
[g6_race_white_owner]
[g6_race_other_owner]

```
local r1 "g6_race_amind_owner"
local r2 "g6_race_nathaw_owner"
local r3 "g6_race_asian_owner"
local r4 "g6_race_black_owner"
local r5 "g6_race_white_owner"
local r6 "g6_race_other_owner"
local race "`r1' `r2' `r3' `r4' `r5' `r6'"
foreach Q of local race{
foreach ow in $owners_1_15{
/*last observation carried backward  */
replace `Q'_`ow'_0=`Q'_`ow'_1 if `Q'_`ow'_0==.
replace `Q'_`ow'_0=`Q'_`ow'_2 if `Q'_`ow'_0==.
replace `Q'_`ow'_0=`Q'_`ow'_3 if `Q'_`ow'_0==.
```

```
replace `Q'_`ow'_0=`Q'_`ow'_4 if `Q'_`ow'_0==.
replace `Q'_`ow'_0=`Q'_`ow'_5 if `Q'_`ow'_0==.
replace `Q'_`ow'_0=`Q'_`ow'_6 if `Q'_`ow'_0==.
}
}
foreach Q of local race{
foreach ow in $owners_1_15{
replace `Q'_`ow'_1 =`Q'_`ow'_0 if `Q'_`ow'_1==.
replace `Q'_`ow'_2 =`Q'_`ow'_0 if `Q'_`ow'_2==.
replace `Q'_`ow'_3 =`Q'_`ow'_0 if `Q'_`ow'_3==.
replace `Q'_`ow'_4 =`Q'_`ow'_0 if `Q'_`ow'_4==.
replace `Q'_`ow'_5 =`Q'_`ow'_0 if `Q'_`ow'_5==.
replace `Q'_`ow'_6 =`Q'_`ow'_0 if `Q'_`ow'_6==.
}
}
```

G7. (Were/Was) (you/[OWNER B-J]) born in the United States?

Yes	01	GO TO G9
No	00	
Don't know	"."	
Refused	"."	

[g7_native_born_owner]

G8. (Are/Is) (you/[OWNER B-J]) a U.S. citizen?

Yes	01
No	00
Don't know	"."
Refused	"."

[g8_us_cit_owner]

G9. What is the highest level of education (you/[OWNER B-J]) (have/has) completed so far? Would you say . . .

Less than 9th grade,	01
Some high school, but no diploma,	02
High school graduate (diploma or equivalent diploma GED)	03
Technical, trade or vocational degree,	04
Some college, but no degree,	05
Associate's degree,	06
Bachelor's degree,	07
Some graduate school but no degree,	08
Master's degree, or	09
Professional school or doctorate?	10
Don't know	"."
Refused	"."

[g9_education_owner]
BY OBSERVATION:
G10a. (Are/Is) (you/[OWNER B-J]) male or female?
Male.....01
Female.....02
Don't know....."."
Refused....."."
[g10_gender_owner]

ENDBLOCK bSectionG2
SECTION G ARRAYED UP TO 10 TIMES, ONCE FOR EACH NEW OWNER.

```
local g7  "g7_native_born_owner"
local g8  "g8_us_cit_owner"
local g9  "g9_education_owner"
local g10 "g10_gender_owner"
local g78910 "`g7' `g8' `g9' `g10'"
forvalues i = 0/7 {
foreach ow in $owners_1_15  {
/* Recode to one - due to Skip Patterns */
replace      g8_us_cit_owner_`ow'_`i'= 1 if g7_native_born_owner_`ow'_`i'==1

}
}
```

2.6. Other Type of Data in the KFS Database

In addition to the fixed core set of questions asked by all businesses in every follow-up survey, over the years, some new questions were added to the survey questionnaires and some of these questions were dropped later.

Appendix A summarizes all the questions that were added to KFS Questionnaires during the survey period as well as in which year the questions were dropped, if they were dropped.

In addition to the data collected by the surveys, the KFS data files have many other variables. Those variables include weighting and non-response adjustment variables, survey management variables, variables provided with D&B sampling frame, and some other variables. A summary of these variables is provided in Appendix B. Appendix B shows the variable names as well as the definition of each variable.

2.7. Single Imputation

2.7.1. Last Observation Carried Forward (LOCF) And Last Observation Carried Backward (LOCB).

Both LOCF and LOCB methods can be used in longitudinal research designs, but they require the strong assumption of stability. LOCF takes into account the

individual's previous observed value on a given variable. If an observation at a certain data collection wave is missing, the last observed value is used as an estimate for this missing observation.

A related method, LOCB, works according to the same approach but imputes a newer observation in the case of a missing earlier observation of the same individual.

- LOCF was used to impute the legal status variable and business location.
- To determine the missing values for d1a_provide_service and d1b_provide_product variables, we use the NAICS description to determine if the business is operating under the services or manufacturing industries. For observations in which we couldn't determine whether they provide a service or product, we used the LOCF method of imputation.

2.7.2. Internal Consistency: Using Information from Related Observations

- For c2_owners, missing and zero values were replaced by total owners.
- The "c4_numowners_confirm," was replaced by the sum of "owner-active," if "c4_numowners_confirm," is not equal to the sum of the owner-active.
- Replace "c1z2_legal_status=7," if "c1z2_legal_status=1," and "c4_numowners_confirm_1>1."

2.7.3. Other Single Imputations

- For d7_perc_sales_xxxx variables, if one variable has a missing value, we impute the missing value using 100-sum (non-missing d7_perc_sales _xxxx).
- For f8xxx_line_y, f8xxx_ bal_y, f8xxx_ owed _y, f10xxx_line_y, f10xxx_ bal_y, f10xxx_ owed _y ,f12xxx_line_y, f12xxx_ bal_y f12xxx_ owed _y, and f4_eq_amt_xxx_allyrs, the missing values are set to zero if the business never used these sources of funding.

```
 replace  f10d_pers_loan_fam_owed_2  =0     if     f9a_pers_loan_fam_0  ==0  &  ///
f9a_pers_loan_fam_1 ==0 & f9a_pers_loan_fam_2 ==0
```

2.8. The KFS Data File after Data Editing (Logical imputation)

A data set file that was subject to recoding of missing values into soft and hard missing values and implemented logical imputations for the baseline and follow-ups surveys is available (in the wide format) at NORC under the name "KFS8_LI.dta".

2.9. Appendix A

The question appear in follow-up								Question	Variable Name
0	1	2	3	4	5	6	7		
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	In what calendar year did [NAME BUSINESS] close?	a11_year_closed
No	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Did [NAME BUSINESS] file for bankruptcy?	a11a_bankruptcy
No	**No**	**No**	**No**	**Yes**	**No**	**No**	**No**	How much did the nation's recent financial problems, which became highly visible in YYYY, affect [NAME BUSINESS] during calendar year YYYY? Would you say . . .	a11b_economy_effect
No	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Our records show that [NAME BUSINESS] had a legal status of [LEGAL STATUS]. As of December 31, YYYY, is that still the legal status of [NAME BUSINESS]?	c1z_confirm_legal_status
No	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	I'm going to read you a list of some different forms of legal status a business can have. As of December 31, YYYY, which form of legal status did [NAME BUSINESS] have? Was it a . . . Something else? (SPECIFY)	c1z2otherspecify
No	**Yes**	**Yes**	**Yes**	**No**	**No**	**No**	**No**	Was this change an increase, a decrease, or no change in the number of people who worked for [NAME BUSINESS] on December 31, YYYY compared to December 31, YYYY?	c5b_num_employees_change
No	**Yes**	**Yes**	**Yes**	**No**	**No**	**No**	**No**	And what was the (increase/decrease) in the number of people who worked for [NAME BUSINESS] on December 31, YYYY compared to December 31, YYYY? Your best estimate is fine	c5c_num_employees_change_amt
No	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Our records show that the primary location where [NAME BUSINESS] operates is [PRIMARY LOCATION]. Is that correct?	c8z_primary_loc_confirm
No	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	What was the main reason for the change of location?	c9_loc_change_reason

Appendix A – Continued

The question appear in follow-up								Question	Variable Name
0	1	2	3	4	5	6	7		
No	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	What was the main reason for the change of location? OTHER (SPECIFY)	c9otherspecify
No	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	As of December 31, YYYY, did [NAME BUSINESS] operate in more than one location?	c10_morelocations
No	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	And as of December 31, YYYY, how many locations did [NAME BUSINESS] operate in?	c11_num_locations
No	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	In what month and year did you open your second location?, Month	c11a_2ndopening_month
No	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	In what month and year did you open your second location?, Year	c11a_2ndopening_year
No	**No**	**No**	**No**	**Yes**	**No**	**No**	**No**	There are many programs available to help new businesses. I am going to read some possible sources of training and assistance that may have been used to help [NAME BUSINESS]. Have you (or any of the other owners) ever received any business training, mentoring, or technical assistance sponsored by (READ ITEM) to help [NAME BUSINESS]? - The Small Business Administration or SBA	c12a_sba
No	**No**	**No**	**No**	**Yes**	**No**	**No**	**No**	There are many programs available to help new businesses. I am going to read some possible sources of training and assistance that may have been used to help [NAME BUSINESS]. Have you (or any of the other owners) ever received any business training, mentoring, or technical assistance sponsored by (READ ITEM) to help [NAME BUSINESS]? - A Federal government agency other than SBA	c12b_fed_gov

Appendix A – Continued

The question appear in follow-up								Question	Variable Name
0	1	2	3	4	5	6	7		
No	**No**	**No**	**No**	**Yes**	**No**	**No**	**No**	There are many programs available to help new businesses. I am going to read some possible sources of training and assistance that may have been used to help [NAME BUSINESS]. Have you (or any of the other owners) ever received any business training, mentoring, or technical assistance sponsored by (READ ITEM) to help [NAME BUSINESS]? - A state or local government	c12c_statelocal_gov
No	**No**	**No**	**No**	**Yes**	**No**	**No**	**No**	There are many programs available to help new businesses. I am going to read some possible sources of training and assistance that may have been used to help [NAME BUSINESS]. Have you (or any of the other owners) ever received any business training, mentoring, or technical assistance sponsored by (READ ITEM) to help [NAME BUSINESS]? - A non-profit association for small businesses such as SCORE	c12d_non_profit
No	**No**	**No**	**No**	**Yes**	**No**	**No**	**No**	There are many programs available to help new businesses. I am going to read some possible sources of training and assistance that may have been used to help [NAME BUSINESS]. Have you (or any of the other owners) ever received any business training, mentoring, or technical assistance sponsored by (READ ITEM) to help [NAME BUSINESS]? - A community college or university	c12e_college_univ
No	**No**	**No**	**No**	**Yes**	**No**	**No**	**No**	There are many programs available to help new businesses. I am going to read some possible sources of training and assistance that may have been used to help [NAME BUSINESS]. Have you (or any of the other owners) ever received any business training, mentoring, or technical assistance sponsored by (READ ITEM) to help [NAME BUSINESS]? - A chamber of commerce	c12f_chamber_of_comm

Appendix A – Continued

The question appear in follow-up								Question	Variable Name
0	1	2	3	4	5	6	7		
No	**No**	**No**	**No**	**Yes**	**No**	**No**	**No**	There are many programs available to help new businesses. I am going to read some possible sources of training and assistance that may have been used to help [NAME BUSINESS]. Have you (or any of the other owners) ever received any business training, mentoring, or technical assistance sponsored by (READ ITEM) to help [NAME BUSINESS]? - A for-profit organization such as an accounting firm	c12g_for_profit_org
No	**No**	**No**	**No**	**Yes**	**No**	**No**	**No**	There are many programs available to help new businesses. I am going to read some possible sources of training and assistance that may have been used to help [NAME BUSINESS]. Have you (or any of the other owners) ever received any business training, mentoring, or technical assistance sponsored by (READ ITEM) to help [NAME BUSINESS]? - Another Source	c12h_other
No	**No**	**No**	**No**	**Yes**	**No**	**No**	**No**	There are many programs available to help new businesses. I am going to read some possible sources of training and assistance that may have been used to help [NAME BUSINESS]. Have you (or any of the other owners) ever received any business training, mentoring, or technical assistance sponsored by (READ ITEM) to help [NAME BUSINESS]? - Other source specify	c12other_specify
No	**No**	**No**	**No**	**No**	**Yes**	**No**	**No**	As of December 31, YYYY, what state is [NAME BUSINESS] chartered in?	c1z3_state_chartered
No	**No**	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	First, during calendar year YYYY, did (BUSINESS NAME) introduce any products or services that were new or significantly improved?	d1_a_new_product
No	**No**	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	During calendar year YYYY, were any of the products or services new to any market or markets [NAME BUSINESS] competes in?	d1_b_new_to_market
No	**No**	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	Were any of the new or significantly improved products or services introduced in YYYY new to [ITEM]? a) A regional market such as nearby cities or counties	d1c_a_regional

Appendix A – Continued

The question appear in follow-up								Question	Variable Name
0	1	2	3	4	5	6	7		
No	**No**	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	Were any of the new or significantly improved products or services introduced in YYYY new to [ITEM]? b) A national-wide market	d1c_b_national
No	**No**	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	Were any of the new or significantly improved products or services introduced in YYYY new to [ITEM]? c) An international market	d1c_c_international
No	**No**	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	During calendar year YYYY, did [BUSINESS NAME] introduce any new or significantly improved processes in the production of goods or providing services? Please include any new or improved processes, even if [NAME BUSINESS] was not the first to introduce it.	d1d_new_processes
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Was the competitive advantage [NAME BUSINESS] had in calendar year YYYY related in any way to teaming up with another company?	d2a_compadv_comp_reason
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Was the competitive advantage [NAME BUSINESS] had in calendar year YYYY related in any way to teaming up with a government lab or research center?	d2a_compadv_govlab_reason
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Was the competitive advantage [NAME BUSINESS] had in calendar year YYYY related in any way to patents that [NAME BUSINESS] owns, has applied for, or licensed?	d2a_compadv_patents_reason
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Was the competitive advantage [NAME BUSINESS] had in calendar year YYYY related in any way to teaming up with a college or university?	d2a_compadv_univ_reason
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Do you consider this to have given [NAME BUSINESS] a major or a minor competitive advantage in calendar year YYYY?	d2b_compadv_comp_strength

Appendix A – Continued

The question appear in follow-up								Question	Variable Name
0	1	2	3	4	5	6	7		
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Do you consider this to have given [NAME BUSINESS] a major or a minor competitive advantage in calendar year YYYY?	d2b_compadv_govlab_strength
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Do you consider this to have given [NAME BUSINESS] a major or a minor competitive advantage in calendar year YYYY?	d2b_compadv_patents_strength
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Do you consider this to have given [NAME BUSINESS] a major or a minor competitive advantage in calendar year YYYY?	d2b_compadv_univ_strength
No	**No**	**No**	**No**	**No**	**Yes**	**No**	**No**	Was the competitive advantage [NAME BUSINESS] had in calendar year YYYY related in any way to [ITEM]? e) Cost advantages	d2c_compadv_cost_reason
No	**No**	**No**	**No**	**No**	**Yes**	**No**	**No**	Was the competitive advantage [NAME BUSINESS] had in calendar year YYYY related in any way to [ITEM]? f) Product or service design or quality	d2c_compadv_design_reason
No	**No**	**No**	**No**	**No**	**Yes**	**No**	**No**	Was the competitive advantage [NAME BUSINESS] had in calendar year YYYY related in any way to [ITEM]? g) Specialized or range of expertise, products or service	d2c_compadv_expertise_reason
No	**No**	**No**	**No**	**No**	**Yes**	**No**	**No**	Was the competitive advantage [NAME BUSINESS] had in calendar year YYYY related in any way to [ITEM]? b) Marketing and promotion	d2c_compadv_marketing_reason
No	**No**	**No**	**No**	**No**	**Yes**	**No**	**No**	Was the competitive advantage [NAME BUSINESS] had in calendar year YYYY related in any way to [ITEM]? a) Price	d2c_compadv_price_reason
No	**No**	**No**	**No**	**No**	**Yes**	**No**	**No**	Was the competitive advantage [NAME BUSINESS] had in calendar year YYYY related in any way to [ITEM]? d) Established reputation	d2c_compadv_reputation_reason

Appendix A – Continued

The question appear in follow-up								Question	Variable Name
0	1	2	3	4	5	6	7		
No	**No**	**No**	**No**	**No**	**Yes**	**No**	**No**	Was the competitive advantage [NAME BUSINESS] had in calendar year YYYY related in any way to [ITEM]? c) Speed of service	d2c_compadv_speed_reason
No	**No**	**No**	**No**	**No**	**Yes**	**No**	**No**	Do you consider this to have given [NAME BUSINESS] a major or a minor competitive advantage in calendar year YYYY? e) Cost advantages	d2d_compadv_cost_strength
No	**No**	**No**	**No**	**No**	**Yes**	**No**	**No**	Do you consider this to have given [NAME BUSINESS] a major or a minor competitive advantage in calendar year YYYY? f) Product or service design or quality	d2d_compadv_design_strength
No	**No**	**No**	**No**	**No**	**Yes**	**No**	**No**	Do you consider this to have given [NAME BUSINESS] a major or a minor competitive advantage in calendar year YYYY? g) Specialized or range of expertise, products or service	d2d_compadv_expertise_strength
No	**No**	**No**	**No**	**No**	**Yes**	**No**	**No**	Do you consider this to have given [NAME BUSINESS] a major or a minor competitive advantage in calendar year YYYY? b) Marketing and promotion	d2d_compadv_marketing_strength
No	**No**	**No**	**No**	**No**	**Yes**	**No**	**No**	Do you consider this to have given [NAME BUSINESS] a major or a minor competitive advantage in calendar year YYYY? a) Price	d2d_compadv_price_strength
No	**No**	**No**	**No**	**No**	**Yes**	**No**	**No**	Do you consider this to have given [NAME BUSINESS] a major or a minor competitive advantage in calendar year YYYY? d) Established reputation	d2d_compadv_reput_strength
No	**No**	**No**	**No**	**No**	**Yes**	**No**	**No**	Do you consider this to have given [NAME BUSINESS] a major or a minor competitive advantage in calendar year YYYY? c) Speed service	d2d_compadv_speed_strength
No	**No**	**No**	**No**	**No**	**Yes**	**No**	**No**	Was [NAME BUSINESS] founded around a new or customized product or service that was created by you or one of the founders of the business?	d5a_founded_newprod

Appendix A – Continued

The question appear in follow-up								Question	Variable Name
0	1	2	3	4	5	6	7		
No	**No**	**No**	**No**	**No**	**Yes**	**No**	**No**	Thinking about the new or customized product or service, around which [NAME BUSINESS] was founded, why was it originally developed? Was it because… a) You or one of the founders needed it for personal use?	d5b_a_personaluse
No	**No**	**No**	**No**	**No**	**Yes**	**No**	**No**	Thinking about the new or customized product or service, around which [NAME BUSINESS] was founded, why was it originally developed? Was it because… b) You or one of the founders needed it for use at a previous job or business?	d5b_b_previousjob
No	**No**	**No**	**No**	**No**	**Yes**	**No**	**No**	Thinking about the new or customized product or service, around which [NAME BUSINESS] was founded, why was it originally developed? Was it because… c) You or one of the founders thought about starting a business based on it or to sell it to someone else?	d5b_c_startingbus
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Now, I will read you a list of customer locations. When I am done reading, please select only one answer choice for your response. During calendar year YYYY, where were most of [NAME BUSINESS]'s customers located? Would you say . . .	d8_customer_locations
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	During calendar year YYYY, were any of [NAME BUSINESS]'s sales made to individuals, businesses, or governments outside the United States?	d8a_international_sales
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	What percent of [NAME BUSINESS]'s total sales were to individuals, businesses, or governments outside of the United States? Would you say . . .	d8b_perc_international_sales

Appendix A – Continued

The question appear in follow-up								Question	Variable Name
0	1	2	3	4	5	6	7		
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	During calendar year YYYY, were any of [NAME BUSINESS]'s sales made to customers through the internet, such as through the business' website or an online retailer site?	d9_internet_sales
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	What percent of [NAME BUSINESS]'s total sales were sales made to customers through the internet? Would you say . . .	d9a_perc_internet_sales
No	**No**	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	During calendar year YYYY, did [BUSINESS NAME] actively seek but not obtain equity from companies, government agencies, venture capitalists, angel investors, or any other individuals who are not spouses, life partners, parents, in-laws, or children of the owners?	f5a_seek_equity
No	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Have you (or other owners) withdrawn money from the business for personal use in calendar year YYYY?	f6a_personal_use
No	**No**	**No**	**No**	**Yes**	**No**	**No**	**No**	#N/A	f6z_family_owned
No	**No**	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	Was collateral required to obtain any of the debt financing options that were used in calendar year YYYY? That is, were you or was [NAME BUSINESS] required to pledge as security any personal or business assets that can be taken should the business fail to repay the debt?	f12e_collateral
No	**No**	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	What collateral was required? Was it... e) Business real estate?	f12f_bus_real_estate
No	**No**	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	What collateral was required? Was it... b) Business equipment or vehicles?	f12f_business_equip_veh
No	**No**	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	What collateral was required? Was it... c) Business securities or deposits?	f12f_business_sec_dep
No	**No**	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	What collateral was required? Was it... d) Patents, copyrights or trademarks?	f12f_intellectual_prop

Appendix A – Continued

The question appear in follow-up								Question	Variable Name
0	1	2	3	4	5	6	7		
No	**No**	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	What collateral was required? Was it... a) Inventory or Accounts receivable?	f12f_inventory_acctrec
No	**No**	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	What collateral was required? Was it... h) Some other type of collateral?	f12f_other
No	**No**	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	What collateral was required? Was it... g) Other personal assets?	f12f_other_pers_assets
No	**No**	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	What collateral was required? Was it... h) Some other type of collateral? (SPECIFY)	f12f_otherspecify
No	**No**	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	What collateral was required? Was it... f) Personal real estate?	f12f_pers_real_estate
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Did [NAME BUSINESS] make any applications for new or renewed loans or lines of credit in calendar year YYYY?	f14d_new_loans
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Were these applications always approved, sometimes approved and sometimes denied, or always denied?	f14e_approved_denied
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Consider the most recent time [NAME BUSINESS]'s credit application was denied. Officially, was the application denied because of business credit history?	f14f_bus_credit_hist
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Consider the most recent time [NAME BUSINESS]'s credit application was denied. Officially, was the application denied because of inadequate documentation provided?	f14f_inadeq_doc
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Consider the most recent time [NAME BUSINESS]'s credit application was denied. Officially, was the application denied because of insufficient collateral?	f14f_insuff_coll
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Consider the most recent time [NAME BUSINESS]'s credit application was denied. Officially, was the application denied because of the loan requested was too large?	f14f_loan_toolarge

Appendix A – Continued

The question appear in follow-up								Question	Variable Name
0	1	2	3	4	5	6	7		
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Consider the most recent time [NAME BUSINESS]'s credit application was denied. Officially, was the application denied because of not being in business long enough?	f14f_new_bus
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Consider the most recent time [NAME BUSINESS]'s credit application was denied. Officially, was the application denied because of other reason?	f14f_other
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Consider the most recent time [NAME BUSINESS]'s credit application was denied. Officially, was the application denied because of other reason (specify)?	f14f_otherspecify
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Consider the most recent time [NAME BUSINESS]'s credit application was denied. Officially, was the application denied because of personal credit history?	f14f_pers_credit_hist
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Consider the most recent time [NAME BUSINESS]'s credit application was denied. Officially, was the application denied because of banks putting stricter restrictions on lending?	f14g_didnotapply
No	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	During calendar year YYYY, was there any time when [NAME BUSINESS] needed credit, but did not apply because you or others associated with [NAME BUSINESS] thought the application would be denied?	f14g_didnotapply
No	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	In calendar year YYYY, did [NAME BUSINESS] have any loan guarantees from a federal government agency, such as the Small Business Administration, or any state or local government agencies?	f14h_loan_guarantees
No	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	What was the most challenging problem your business faced in calendar year YYYY? Was it . . .	f14j_most_challenging

Appendix A – Continued

The question appear in follow-up								Question	Variable Name
0	1	2	3	4	5	6	7		
No	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	What was the most challenging problem your business faced in calendar year YYYY? Other specify ...	f14j_other_specify
No	**No**	**No**	**No**	**Yes**	**No**	**No**	**No**	#N/A	f14i_economy_effect
No	**Yes**	**Yes**	**Yes**	**No**	**No**	**No**	**No**	Was this an increase, a decrease, or no change in the amount of revenue for [NAME BUSINESS] in YYYY compared to YYYY?	f16b_rev_YYYY_change
No	**Yes**	**Yes**	**Yes**	**No**	**No**	**No**	**No**	And what was the percentage change in revenue in YYYY compared YYYY? Your best estimate is fine.	f16c_perc_change
No	**Yes**	**Yes**	**Yes**	**No**	**No**	**No**	**No**	Was this an increase, a decrease, or no change in total expenses for [NAME BUSINESS] in YYYY compared to YYYY?	f17b_total_exp_YYYY_change
No	**Yes**	**Yes**	**Yes**	**No**	**No**	**No**	**No**	And what was the percentage change in total expenses in YYYY compared to YYYY? Your best estimate is fine.	f17c_perc_change
No	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	Please estimate [NAME BUSINESS]'s total research and development expenses for calendar year YYYY, including materials, equipment, space, salaries, wages, benefits, and consulting fees?	f19a_res_dev_amt_YYYY
No	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	Investments in intangible assets are expenditures expected to produce long-term benefits for businesses. I'm going to read you some types of intangible assets. When thinking about each category, please consider the cost of in-house activities in these areas including the time of the business owner(s), as well as services or license fees from outside providers. Did [NAME BUSINESS] have expenditures in [ITEM/The design of new and improved products and services] in calendar year YYYY?	f19b_a_design

Appendix A – Continued

The question appear in follow-up								Question	Variable Name
0	1	2	3	4	5	6	7		
No	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	Investments in intangible assets are expenditures expected to produce long-term benefits for businesses. I'm going to read you some types of intangible assets. When thinking about each category, please consider the cost of in-house activities in these areas including the time of the business owner(s), as well as services or license fees from outside providers. Did [NAME BUSINESS] have expenditures in [ITEM/Investments in software or databases] in calendar year YYYY?	f19b_b_investments
No	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	Investments in intangible assets are expenditures expected to produce long-term benefits for businesses. I'm going to read you some types of intangible assets. When thinking about each category, please consider the cost of in-house activities in these areas including the time of the business owner(s), as well as services or license fees from outside providers. Did [NAME BUSINESS] have expenditures in [ITEM/Brand development such as advertising or marketing] in calendar year YYYY?	f19b_c_brand_dev
No	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	Investments in intangible assets are expenditures expected to produce long-term benefits for businesses. I'm going to read you some types of intangible assets. When thinking about each category, please consider the cost of in-house activities in these areas including the time of the business owner(s), as well as services or license fees from outside providers. Did [NAME BUSINESS] have expenditures in [ITEM/Organizational development such as company formation expenses or management consulting] in calendar year YYYY?	f19b_d_org_dev

Appendix A – Continued

The question appear in follow-up								Question	Variable Name
0	1	2	3	4	5	6	7		
No	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	Investments in intangible assets are expenditures expected to produce long-term benefits for businesses. I'm going to read you some types of intangible assets. When thinking about each category, please consider the cost of in-house activities in these areas including the time of the business owner(s), as well as services or license fees from outside providers. Did [NAME BUSINESS] have expenditures in [ITEM/Worker training] in calendar year YYYY?	f19b_e_worker_training
No	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	Investments in intangible assets are expenditures expected to produce long-term benefits for businesses. I'm going to read you some types of intangible assets. When thinking about each category, please consider the cost of in-house activities in these areas including the time of the business owner(s), as well as services or license fees from outside providers. Did [NAME BUSINESS] have expenditures in [ITEM/Any other intangible asset investments] in calendar year YYYY?	f19b_f_other
No	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	Investments in intangible assets are expenditures expected to produce long-term benefits for businesses. I'm going to read you some types of intangible assets. When thinking about each category, please consider the cost of in-house activities in these areas including the time of the business owner(s), as well as services or license fees from outside providers. Did [NAME BUSINESS] have expenditures in [ITEM/Any other intangible asset investments - specify] in calendar year YYYY?	f19b_f_other_specify
No	**No**	**No**	**No**	**Yes**	**No**	**No**	**No**	Thinking about all the intangible asset expenditures [LIST IF NECESSARY] you just told me about, please estimate [NAME BUSINESS]'s total expenses on intangible assets for calendar year YYYY.	f19c_intangassets_amt

Appendix A – Continued

The question appear in follow-up								Question	Variable Name
0	1	2	3	4	5	6	7		
No	**No**	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	During calendar year YYYY, how much money did (BUSINESS NAME) spend on [INTANGIBLE ASSETS ITEM]/The design of new and improved products and services?	f19c_a_design_amt
No	**No**	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	During calendar year YYYY, how much money did (BUSINESS NAME) spend on [INTANGIBLE ASSETS ITEM]/Investments in software or databases?	f19c_b_investments_amt
No	**No**	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	During calendar year YYYY, how much money did (BUSINESS NAME) spend on [INTANGIBLE ASSETS ITEM]/ Brand development such as advertising or marketing?	f19c_c_brand_dev_amt
No	**No**	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	During calendar year YYYY, how much money did (BUSINESS NAME) spend on [INTANGIBLE ASSETS ITEM]/Organizational development such as company formation expenses or management consulting?	f19c_d_org_dev_amt
No	**No**	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	During calendar year YYYY, how much money did (BUSINESS NAME) spend on [INTANGIBLE ASSETS ITEM]/Worker training?	f19c_e_worker_training_amt
No	**No**	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	During calendar year YYYY, how much money did (BUSINESS NAME) spend on [INTANGIBLE ASSETS ITEM]/Any other intangible asset investments?	f19c_f_other_amt
No	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	Did [NAME BUSINESS] file for Chapter 11 bankruptcy protection at any time during calendar year YYYY?	f32_chap11_bankruptcy
No	**No**	**No**	**No**	**Yes**	**No**	**No**	**No**	Now I'd like you to think about how much you expected [NAME BUSINESS] to grow since the business was started. How much do you think [NAME BUSINESS] met your expectations for growth between when the business was started and December 31, YYYY? Would you say [NAME BUSINESS]'s growth . . .	f33_expected_growth

Appendix A – Continued

The question appear in follow-up								Question	Variable Name
0	1	2	3	4	5	6	7		
No	**No**	**No**	**No**	**Yes**	**No**	**No**	**No**	Compared to [NAME BUSINESS]'s revenues for calendar year YYYY, what do you expect [NAME BUSINESS]'s revenues will be in calendar year 2011? Do you think revenues will . . .	f34_future_revenue
No	**No**	**No**	**Yes**	**No**	**No**	**No**	**No**	What was the primary field of study for this degree?	g9_fieldofstudy_resp
No	**No**	**No**	**Yes**	**No**	**No**	**No**	**No**	What was the primary field of study for this degree?	g9_fieldofstudy_resp_2nd
No	**No**	**No**	**Yes**	**No**	**No**	**No**	**No**		g9_fieldofstudy_cip_desc
No	**No**	**No**	**Yes**	**No**	**No**	**No**	**No**		g9_fieldofstudy2nd_cip_desc
No	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	What is your marital status?	g10b_marital_status
No	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes**	Including the equity in your home and business, what is your approximate total net worth, which are all your assets minus all debts?	g10c_net_worth
No	**No**	**No**	**No**	**Yes**	**No**	**No**	**No**	How much do you agree with the following statement? In uncertain times, I usually expect the best. Would you say you . . .	g10d_personal_outlook

2.10. Appendix B

Variable	Definition
sampleinfo_samplestrata_0	The technology and gender ownership sampling strata Mathematica used to select the KFS sample. 101 = high tech, woman owned 102 = high tech, not woman owned 201 = medium tech, woman owned 202 = medium tech, not woman owned 301 = non tech, woman owned 302 = non tech, not woman owned
fstatus_0	The response status of Baseline Survey cases. 1= Complete 21 = Ineligible, out of business 22 = Ineligible, does not meet project requirements 3 = Refusal 4 = Unlocatable
resp_0	Response Indicator – All completes and ineligibles are considered respondents under this measure. 1 = Respondent , 0 = Non-respondent
samplingweight_0	Initial sampling weight when the sample was drawn.
wgt_ini_0	Weight after release adjustment.
wgt_ploct_0	Weight after location model.
wgt_final_0	Final Baseline weight after non-response model.
loct_0	Location indicator – Indicates whether business was located. 1 = Located , 0 = Unlocatable
fstatus_1	The response status of First Follow-Up cases. 1 = Complete 21 = Ineligible, out of business 22 = Ineligible, does not meet project expectations 3 = Refusal 4 = Unlocatable
resp_1	Response Indicator for First Follow-Up cases – All completes and ineligibles are considered respondents under this measure. 1 = Respondent , 0 = Non-respondent
wgt_final_1	Final weight after non-response model for First Follow-Up cases.
fstatus_f2_2	The response status of Second Follow-Up cases. 1 = Complete 21 = Ineligible, out of business 22 = Ineligible, does not meet project expectations 3 = Refusal 4 = Unlocatable

Appendix B – Continued

Variable	Definition
resp_f2_2	Response Indicator for Second Follow-Up cases – All completes and ineligibles are considered respondents under this measure. 1 = Respondent , 0 = Non-respondent
wgt_final_f2_2	Second Follow-Up final weight – Cross-Sectional.
wgt_final_f12_long_2	Second Follow-Up final weight – Longitudinal.
resp_f12_long_2	Indicates if there is a Second Follow-Up longitudinal weight. 1 = Yes , 0 = No
fstatus_f3_3	The response status of Third Follow-Up cases. 1 = Complete 21 = Ineligible, out of business 22 = Ineligible, does not meet project expectations 3 = Refusal 4 = Unlocatable
resp_f3_3	Response Indicator for Third Follow-Up cases – All completes and ineligibles are considered respondents under this measure. 1 = Respondent 0 = Non-respondent
wgt_final_f3_3	Third Follow-Up final weight – Cross-Sectional.
wgt_final_f123_long_3	Third Follow-Up final weight – Longitudinal.
resp_f123_long_3	Indicates if there is a Third Follow-Up longitudinal weight. 1 = Yes , 0 = No
fstatus_f4_4	The response status of Fourth Follow-Up cases. 1 = Complete 21 = Ineligible, out of business 22 = Ineligible, does not meet project expectations 3 = Refusal 4 = Unlocatable
resp_f4_4	Response Indicator for Fourth Follow-Up cases – All completes and ineligibles are considered respondents under this measure. 1 = Respondent , 0 = Non-respondent
wgt_final_f4_4	Fourth Follow-Up final weight – Cross-Sectional.
wgt_final_f1234_long_4	Fourth Follow-Up final weight – Longitudinal.
resp_long_f4_4	Indicates if there is a Fourth Follow-Up longitudinal weight. 1 = Yes , 0 = No
fstatus_f5_5	The response status of Fifth Follow-Up cases. 1 = complete 21 = ineligible, out of business 22 = Ineligible, does not meet project expectations 3 = Refusal 4 = Unlocatable

Appendix B – Continued

Variable	Definition
resp_f5_5	Response Indicator for Fifth Follow-Up cases – All completes and ineligibles are considered respondents under this measure. 1 = Respondent , 0 = Non-respondent
wgt_final_f5_5	Fifth Follow-Up final weight – Cross-Sectional.
wgt_final_f5_long_5	Fifth Follow-Up final weight – Longitudinal.
resp_long_f5_5	Indicates if there is a Fifth Follow-Up longitudinal weight. 1 = Yes , 0 = No
fstatus_f6_6	The response status of Sixth Follow-Up cases. 1 = complete 21 = ineligible, out of business 22 = Ineligible, does not meet project expectations 3 = Refusal 4 = Unlocatable
resp_f6_6	Response Indicator for Sixth Follow-Up cases – All completes and ineligibles are considered respondents under this measure. 1 = Respondent , 0 = Non-respondent
resp_long_f6_6	Indicates if there is a Sixth Follow-Up longitudinal weight. 1 = Yes , 0 = No
wgt_final_f6_6	Sixth Follow-Up final weight – Cross-Sectional.
wgt_final_f6_long_6	Sixth Follow-Up final weight – Longitudinal.
mprid	The identification number provided by Mathematica to each sampled business.
final_status_code_0	Final Disposition Code for sampled businesses in the Baseline Survey 10 = Telephone Complete , 30 = Web Complete
final_status_code_1	Final Disposition Code for sampled businesses in the First Follow-Up Survey. 10 = Telephone Complete 30 = Web Complete 200 = Refusal 210 = Refusal by gatekeeper 220 = Refusal by other 330 = Effort ended/Case retired 401 = Language Barrier (Spanish) 431 = Temporarily Stopped Operations During Field Period 450 = Business moved out of country 463 = No Longer in Business 465 = Started Previous to 2004 – The business was engaged in new business activity prior to calendar year 2004. New business activity is defined as applying for an EIN, submitting a Schedule C or C-EZ, paying state unemployment insurance taxes, or making FICA payments 590 = Unlocatable

Appendix B – Continued

Variable	Definition
final_status_code_2	Final Disposition Code for sampled businesses in the Second Follow-Up Survey. 10 = CATI complete 30 = Web Complete 200 = Refusal by known respondent 209 = Adamant refusal by known respondent 210 = Refusal by gatekeeper 220 = Refusal by unknown person 330 = Effort ended/Case retired 431 = Temporarily Stopped Operations During Field Period 463 = No Longer in Business 468 = Duplicate case 590 = Unlocatable
final_status_code_3	Final Disposition Code for sampled businesses in the Third Follow-Up Survey. 10 = CATI complete 30 = Web Complete 200 = Refusal by known respondent 209 = Adamant refusal by known respondent 210 = Refusal by gatekeeper 220 = Refusal by unknown person 330 = Effort ended/Case retired 431 = Temporarily Stopped Operations During Field Period 450 = Business moved out of country 463 = No Longer in Business 590 = Unlocatable
final_status_code_4	Final Disposition Code for sampled businesses in the Fourth Follow-Up Survey. 10 = CATI complete 30 = Web Complete 200 = Refusal by known respondent 209 = Adamant refusal by known respondent 210 = Refusal by gatekeeper 220 = Refusal by unknown person 330 = Effort ended/Case retired 430 = No owner/respondent available during field period 431 = Temporarily Stopped Operations During Field Period 450 = Business moved out of country 463 = No Longer in Business 590 = Unlocatable

Appendix B – Continued

Variable	Definition
final_status_code_5	Final Disposition Code for sampled businesses in the Fifth Follow-Up Survey. 10 = CATI complete 30 = Web Complete 200 = Refusal by known respondent 209 = Adamant refusal by known respondent 210 = Refusal by gatekeeper 219 = Adamant refusal by gatekeeper 220 = Refusal by unknown person 229 = Adamant refusal by unknown person 330 = Effort ended/Case retired 430 = No owner/respondent available during field period 431 = Temporarily Stopped Operations During Field Period 463 = No Longer in Business 590 = Unlocatable
final_status_code_6	Final Disposition Code for sampled businesses in the Sixth Follow-Up Survey. 10 = CATI complete 30 = Web Complete 200 = Refusal by known respondent 209 = Adamant refusal by known respondent 210 = Refusal by gatekeeper 219 = Adamant refusal by gatekeeper 220 = Refusal by unknown person 229 = Adamant refusal by unknown person 330 = Effort ended/Case retired 431 = Temporarily Stopped Operations During Field Period 463 = No Longer in Business 590 = Unlocatable
sampleinfo_wave_0	The sample wave number the sample business was included in during the Baseline Survey.
sampleinfo_release_0	The Baseline Survey sample was released in six batches: 1 = 1st release sent 07/29/05 2 = 2nd release sent 09/23/05 3 = 3rd release sent 11/4/05 4 = 4th release sent 12/1/05 5 = 5th release sent 01/11/06 6 = 6th release sent 02/28/06
sampleinfo_release_1	The First Follow-Up Survey sample was released in four batches: 1 = 1st release sent 06/14/06 2 = 2nd release sent 07/6/06 3 = 3rd release sent 08/07/06 4 = 4th release sent 08/31/06

Appendix B – Continued

Variable	Definition
sampleinfo_release_2	The Second Follow-Up Survey sample was released in three batches: 1 = 1st release sent 05/17/07 2 = 2nd release sent 05/31/07 3 = 3rd release sent 06/25/07
sampleinfo_release_3	The Third Follow-Up Survey sample was released in three batches: 1 = 1st release sent 06/24/08 2 = 2nd release sent 07/09/08 3 = 3rd release sent 07/23/08
interviewdate	The date the interview was completed
sampleinfo_strata_0	The high-tech strata of businesses in the D&B sample frame were defined by using the SIC codes of businesses listed below: 1 = High Tech 28 Chemicals and allied products 35 Industrial machinery and equipment 36 Electrical and electronic equipment 38 Instruments and related products 2 = Medium Tech 131 Crude Petroleum and natural gas operations 211 Cigarettes 229 Miscellaneous textile goods 261 Pulp mills 267 Miscellaneous converted paper products 291 Petroleum refining 299 Miscellaneous petroleum and coal products 335 Nonferrous rolling and drawing 348 Ordnance and accessories, not elsewhere classified 371 Motor vehicles and equipment 372 Aircraft and parts 376 Guided missiles, space vehicles, parts 379 Miscellaneous transportation equipment 737 Computer and data processing services 871 Engineering and architectural services 873 Research and testing services 874 Management and public relations 899 Services, not elsewhere classified 3 = Not High Tech Includes all other industries not listed above
sampleinfo_release_4	The Fourth Follow-Up Survey sample was released in three batches: 1 = 1st release sent 06/3/09 2 = 2nd release sent 06/17/09 3 = 3rd release sent 06/29/09

Appendix B – Continued

Variable	Definition
sampleinfo_release_5	The Fifth Follow-Up Survey sample was released in three batches: 1 = 1st release sent 05/12/2010 2 = 2nd release sent 05/25/2010 3 = 3rd release sent 06/08/2010
sampleinfo_release_6	The Sixth Follow-Up Survey sample was released in three batches: 1 = 1st release sent 05/10/2011 2 = 2nd release sent 05/24/2011 3 = 3rd release sent 06/06/2011
timezone	The time zone the business is located in: 2 = Hawaiian/Aleutian Time Zone 3 = Alaska Time Zone 4 = Pacific Time Zone 5 = Mountain Time Zone 6 = Central Time Zone 7 = Eastern Time Zone
sampleinfo_sex_0	The gender variable for the business principal provided with the D&B listing. 1 = Male 2 = Female
sampleinfo_sales_volume_0	Sales volume provided by D&B for the sampled business
sampleinfo_sales_volume_code_0	Code provided by D&B that helps define the SampleInfo_Sales_Volume variable. 0 = SampleInfo_Sales_Volume contains a real value 1 = SampleInfo_Sales_Volume contains the low end of a real range 2 (with an all zero SampleInfo_Sales_Volume) = SampleInfo_Sales_Volume is unknown 2 (with a non-zero SampleInfo_Sales_Volume) = SampleInfo_Sales_Volume is an estimate based on defined norms for this industry and size of business
sampleinfo_employees_total_0	Total employees provided by D&B for the sampled business.
sampleinfo_manufacturingindica_0	Code provided by D&B indicating whether manufacturing operations occur at this location. 0 = Manufacturing is done here 1 = No manufacturing done here
sampleinfo_legal_status_0	Code provided by D&B indicating the legal status of the establishment. 000 = not available 003 = corporation 012 = partnership of unknown type 013 = proprietorship type

Appendix B – Continued

Variable	Definition
sampleinfo_majorindustry_categ_0	Code provided by D&B that denotes the major industry category under which a sampled business falls. 1 = agriculture 2 = mining 3 = construction 4 = manufacturing 5 = transportation, communications, utilities 6 = wholesale trade 7 = retail trade 8 = finance, insurance, real estate 9 = services
sampleinfo_line_business_0	A brief description of the line of business provided by D&B for the sampled business based on the SIC code.
sampleinfo_sic_1_0	The first Standard Industrial Classification (SIC) code provided by D&B for the sampled business. The SIC code taxonomy assigns a code to businesses and other organizations, classifying and subdividing the activity performed by the establishment at that location.
sampleinfo_sic_2_0	The second SIC code provided by D&B for the sampled business.
sampleinfo_sic_3_0	The third SIC code provided by D&B for the sampled business.
sampleinfo_sic_4_0	The fourth SIC code provided by D&B for the sampled business.
sampleinfo_sic_5_0	The fifth SIC code provided by D&B for the sampled business.
sampleinfo_sic_6_0	The sixth SIC code provided by D&B for the sampled business.
subsidiary_indicator_0	Indicating if this business is a corporation that is more than 50 percent owned by another company according to D&B: 0 = Non subsidiary 3 = Subsidiary
sampleinfo_naics_code_0	The NAICS code provided by D&B for the sampled business. Refer to NAICS—North American Industry Classification System for additional documentation of this variable.

Appendix B – Continued

Variable	Definition
sampleinfo_naics_desp_0	Descriptions of the NAICS code provided by D&B for the sampled business. Refer to NAICS—North American Industry Classification System for additional documentation of this variable.
sampleinfo_woman_owned_0	Code provided by D&B indicating if the majority of the establishment is owned by a woman: Y = Woman owned N = Not woman owned
sampleinfo_fips_msa_0	The Federal Information Processing System (FIPS) Metropolitan Statistical Area (MSA) code provided by D&B for the sampled business. Refer to Current Lists of Metropolitan and Micropolitan Statistical Areas and Definitions for additional documentation of this variable
credrisk	Credit risk variables categorized as shown below: Credit Score Risk Class / Credit Score Percentile / Commercial Credit Score 1 / 91-100 / 536-670 2 / 71-90 / 493-535 3 / 31-70 / 423-492 4 / 11-30 / 376-422 5 / 1-10 / 101-375
emphr*	Employees Here – Number of employees at the DUNS location.
ephrind*	Employee Here Indicator: 0=Actual 1=Lower end of Range 2=Estimate of N/A
emptot*	Total Employees across all HQ/Branches and Subsidiary locations.
epttind*	Total Employees Indicator: 0=Actual 1=Lower end of Range 2=Estimate of N/A
salvol*	Total Sales
slvlind*	Total Sales Indicator: 0=Actual 1=Lower end of Range 2=Estimate of N/A

Appendix B – Continued

Variable	Definition
paysc	PAYDEX Score – a unique, dollar weighted indicator of payment performance based on payment experiences, as reported to D&B by trade references. A PAYDEX score will not be calculated for businesses with less than three experiences. There must also be two suppliers reporting trade on that business for a PAYDEX to be calculated. If there is insufficient trade to calculate a PAYDEX score, 999 (unavailable) is entered in the monthly PAYDEX. The following chart outlines the specific 1-100 score and what each means.

PAYDEX	Payment Practices
100	Anticipate
90-99	Discount
80-89	Prompt
70-79	15 Days Beyond Terms
60-69	22 Days Beyond Terms
50-59	30 Days Beyond Terms
40-49	60 Days Beyond Terms
30-39	90 Days Beyond Terms
20-29	120 Days Beyond Terms
1-19	Over 120 Days Beyond Terms

fssp	The Financial Stress Score predicts a business‘s likelihood of experiencing financial stress over the next 12-month period. D&B defines a financially stressed company as one that obtains legal relief from creditors, ceases operations with debts outstanding, goes into receivership or reorganization, or makes an arrangement for the benefit of creditors over the next 12-month period, based on the information in D&B‘s commercial database.

Financial Stress Score Percentile	Probability Of Failure
95–100	0.03%
69–94	0.09%
34–68	0.24%
2–33	0.84%
1	4.70%

Appendix B – Continued

Variable	Definition
totalowners	The total number of owners of the business, including owner- operators collected at question C4, and non-operating equity owners collected at question F3.
f6check	The total percentage of ownership accounted for prior to reconciliation. This includes the percentages collected among owner-operators at the question F2 series and among non- operating equity owners at question F5
naics_code	Primary industry of the business confirmed or updated during the surveys. Refer to NAICS—North American Industry Classification System for additional documentation of this variable.
state_final	The state abbreviation for the sampled business.
zip_final	The zip code for the sampled business.
msa_final_3 msa_final_4 msa_final_5 msa_final_6	Metropolitan Statistical Area, based on most recent zip code collected or confirmed from the panel businesses. Refer to Current Lists of Metropolitan and Micropolitan Statistical Areas and Definitions for additional documentation of this Variable.
respondent_1	Owner-operator who completed the First Follow-Up Survey.
respondent_2	Owner-operator who completed the Second Follow-Up Survey.
respondent_3	Owner-operator who completed the Third Follow-Up Survey.
respondent_4	Owner-operator who completed the Fourth Follow-Up Survey.
respondent_5	Owner-operator who completed the Fifth Follow-Up Survey.
respondent_6	Owner-operator who completed the Sixth Follow-Up Survey.
owner_active_oo_1	Active owner-operator(s) 1-10 at time of First Follow-Up Survey. oo= 01 to 10
owner_active_oo_2	Active owner-operator(s) 1-10 at time of Second Follow-Up Survey. oo = 01 to 10
owner_active_oo_3	Active owner-operator(s) 1-11 at time of Third Follow-Up Survey. oo = 01 to 11
owner_active_oo_4	Active owner-operator(s) 1-14 at time of Fourth Follow-Up Survey. oo = 01 to 14
owner_active_oo_5	Active owner-operator(s) 1-15 at time of Fifth Follow-Up Survey. oo = 01 to 15
owner_active_oo_6	Active owner-operator(s) 1-15 at time of Sixth Follow-Up Survey. oo = 01 to 15
techempl	Technology Employers - Following Chapple et al. (2004), industries (based on NAICS_Code) where employment of these occupations exceeds three times the national averages of 3.33%, or 9.98%. 1 = Technology Employer 0 = Non-Technology Employer

Appendix B – Continued

Variable	Definition
techgenr	Technology Generators - Following Chapple et al. (2004), industries (based on NAICS_Code) that are generators of technology, which are defined by the NSF's Survey of Industrial Research and Development as industries that exceed the U.S. average for both research and development expenditures for employee ($11,972) and the proportion of full-time-equivalent R&D scientists and engineers in the industry workforce (5.9%). 1 = Technology Generator 0 = Non-Technology Generator
hightech	High technology industry indicator based on whether the business is a Technology Employer or Technology Generator. 1 = High tech 0 = Non-high tech
email	Indicates whether respondent provided an email address for the business or owner. 1 = yes, email address provided 0 = no, email address not provided
website	Indicates whether respondent provided a website for the business. 1 = yes, website provided 0 = no, website not provided

3.1. KFS Data Structure

Data from a repeated-measures design can be set up in two different data formats: wide format and long format. In the wide format, all multi-wave variables from the same business and associated owners form just one record. For example, consider the following data: mprid is the businesses ID and var1_0, var1_1 and var1_2 are the sales for the baseline, first and second follow-up, and var2_0 is the gender of the owner, which was collected in the baseline survey only (constant over time).

mprid	var1_0	var1_1	var1_2	var2_0
1	485	2542	4095	1
2	2724	9292	.	1
3	9924	8049	2966	0

In the long (panel) format data, all variables from each wave and for the same business and associated owners form one record. For example:

mprid	suffix	var1	var2_0
1	_0	485	1
1	_1	2542	1
1	_2	4095	1
2	_0	2724	1
2	_1	9292	1
2	_2	.	1
3	_0	9924	0
3	_1	8049	0
3	_2	2966	0

Two identification variables are needed for the long (panel) format data; in addition to the business ID, we need to have time variable (suffix), but only one variable is needed for the measurements “Var1” and “Var2.”

The long (panel) format is very useful for any longitudinal data analysis of the KFS panel. Longitudinal businesses that closed in a particular follow-up, were sold, or merged (records with no information) can be dropped in the long format, but not in the wide format. For example, if the business with ID 2 was sold in time 2, then we can drop that record in time 2, and as a result, we will be dealing with unbalanced panel data.

mprid	suffix	var1	var2
1	_0	485	1
1	_1	2542	1
1	_2	4095	1
2	_0	2724	1
2	_1	9292	1
3	_0	9924	0
3	_1	8049	0
3	_2	2966	0

The advantage of the wide format data is that it is more convenient for the analysis of transitions and sequences, cross-tab (e.g. wave 1 vs. wave 2), lagged regression ($y_t = \alpha + \beta y_{t-1} + \gamma x_{t-1}$) , recoding data into soft and hard missing value, logical imputation, defining subpopulation based on time varying variables, survival analysis, cross-sectional analysis, SEM analysis, and some other data manipulation.

Both the confidential version of the KFS and the public version data come in a wide format. Thus, KFS panel data is stored in cross-sectional data format.

Most statistical packages allow restructuring data from the wide format to the long format and vice versa.

3.1.1. Data Reshaping: Wide Format ⇔ Long Format

In Stata®, the reshape function is very useful when dealing with restructuring data from one format to another. For example, to reshape our example above to a long format, we need the following Stata® command:

	mprid	var1_0	var1_1	var1_2	var2_0
1	1	485	2542	4095	1
2	2	2724	9292	.	1
3	3	9924	8049	2966	0

```
reshape long var1, i(mprid) j(suffix)   string

(note: j = _0 _1 _2)
Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                        3   ->      9
Number of variables                   5   ->      4
j variable (3 values)                     ->   suffix
xij variables:
                  var1_0 var1_1 var1_2    ->   var1
```

	mprid	suffix	var1	var2_0
1	1	_0	485	1
2	1	_1	2542	1
3	1	_2	4095	1
4	2	_0	2724	1
5	2	_1	9292	1
6	2	_2	.	1
7	3	_0	9924	0
8	3	_1	8049	0
9	3	_2	2966	0

Because we did not specify var2_0 in the command, Stata assumes that it is constant within the MPRID observations. To go back to wide after using reshape long:

```
reshape wide

(note: j = _0 _1 _2)

Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                        9   ->      3
Number of variables                   4   ->      5
j variable (3 values)            suffix   ->   (dropped)
xij variables:
                                   var1   ->   var1_0 var1_1 var1_2
```

	mprid	var1_0	var1_1	var1_2	var2_0
1	1	485	2542	4095	1
2	2	2724	9292	.	1
3	3	9924	8049	2966	0

3.1.2. Wide vs. Long Format for Multiply Imputed Data

For multiply imputed data, the wide versus long terminology is borrowed from reshape and the structures are similar but not equivalent. All the KFS multiply imputed (mi) data files are formatted using the long style, which means, in addition to the original KFS data (regardless of the original data format, m=0), we have another five imputed datasets of the KFS (m=1,2,3,4,5).

Example: Original KFS in wide format - mi in long format

m	mprid	var1_0	var1_1	var1_2	var2_0
0	1	485	2542	4095	1
0	2	2724	9292	.	1
0	3	9924	8049	2966	0
1	1	485	2542	4095	1
1	2	2724	9292	Imputed 1	1
1	3	9924	8049	2966	0
2	1	485	2542	4095	1
2	2	2724	9292	Imputed 2	1
2	3	9924	8049	2966	0
3	1	485	2542	4095	1
3	2	2724	9292	Imputed 3	1
3	3	9924	8049	2966	0
4	1	485	2542	4095	1
4	2	2724	9292	Imputed 4	1
4	3	9924	8049	2966	0
5	1	485	2542	4095	1
5	2	2724	9292	Imputed 5	1
5	3	9924	8049	2966	0

Example: Original KFS in long format - mi in long format

m	mprid	suffix	var1	var2_0
0	1	_0	485	1
0	1	_1	2542	1
0	1	_2	4095	1
0	2	_0	2724	1
0	2	_1	9292	1
0	2	_2	.	1
0	3	_0	9924	0
0	3	_1	8049	0
0	3	_2	2966	0
1	1	_0	485	1
1	1	_1	2542	1
1	1	_2	4095	1
1	2	_0	2724	1
1	2	_1	9292	1
1	2	_2	Imputed 1	1
1	3	_0	9924	0
1	3	_1	8049	0
1	3	_2	2966	0
2	1	_0	485	1
2	1	_1	2542	1
2	1	_2	4095	1
2	2	_0	2724	1
2	2	_1	9292	1
2	2	_2	Imputed 2	1
2	3	_0	9924	0
2	3	_1	8049	0
2	3	_2	2966	0
3	1	_0	485	1
3	1	_1	2542	1
3	1	_2	4095	1
.	.	.	.	.
5	3	_0	9924	0
5	3	_1	8049	0
5	3	_2	2966	0

3.2. KFS Data Files at NORC

This section describes the data files available at NORC and how to work with these files using Stata.

3.2.1. The Original KFS Data File

The original KFS data file that was produced by MPR is available at NORC [Kfs8_enclave_14oct13.dta]. The missing values in this file are not coded into soft and hard missing values; all missing values are reported as soft missing.

3.2.2. The KFS Data File after Data Editing (Logical Imputation)

A data set file that was subject to recoding of missing values into soft and hard missing values and implementing logical imputations for the baseline and follow-ups surveys is available at NORC under the name "KFS8_LI.dta." The "KFS8_LI" file is available in the wide format at NORC.

3.2.2.1. Reshape the Data from Wide to Long Format

The following Stata code will reshape the KFS8_LI file from wide to long format. The Stata code produces two files: a cross sectional in long format [KFS8_LI_CS_Long.dta] and longitudinal (3,140 penal) in long format [KFS8_LI_L_Long.dta].

```
/*-----------------------------------*/
  clear
  clear mata
  clear matrix
  capture log close
  set more off
  set maxvar 32767
  program drop _all
/*-----------------------------------*/
  cd " xxx:\KFS_Manual_and_Data"

  use KFS8_LI,clear
* Data Management Var
rename one_ower_00 one_owner_0
rename one_ower_01 one_owner_1
rename one_ower_02 one_owner_2
rename one_ower_03 one_owner_3
rename one_ower_04 one_owner_4
rename one_ower_05 one_owner_5
rename one_ower_06 one_owner_6
rename one_ower_07 one_owner_7
* Make sure that the race do not change over time due to entry errors
forvalues i = 1/7 {
foreach ow in $owners_1_15  {

replace      g6_race_amind_owner_`ow'_`i'      =     g6_race_amind_owner_`ow'_0 if
       g6_race_amind_owner_`ow'_`i'      <.     &     g6_race_amind_owner_`ow'_0<.
replace      g6_race_asian_owner_`ow'_`i'      =     g6_race_asian_owner_`ow'_0 if
       g6_race_asian_owner_`ow'_`i'      <.     &     g6_race_asian_owner_`ow'_0<.
replace      g6_race_black_owner_`ow'_`i'      =     g6_race_black_owner_`ow'_0 if
       g6_race_black_owner_`ow'_`i'      <.     &     g6_race_black_owner_`ow'_0<.
replace      g6_race_nathaw_owner_`ow'_`i'     =     g6_race_nathaw_owner_`ow'_0
       if     g6_race_nathaw_owner_`ow'_`i'     <.     &
       g6_race_nathaw_owner_`ow'_0<.
replace      g6_race_other_owner_`ow'_`i'      =     g6_race_other_owner_`ow'_0 if
       g6_race_other_owner_`ow'_`i'      <.     &     g6_race_other_owner_`ow'_0<.
replace      g6_race_white_owner_`ow'_`i'      =     g6_race_white_owner_`ow'_0 if
       g6_race_white_owner_`ow'_`i'      <.     &     g6_race_white_owner_`ow'_0<.
}
}


#delimit ;
```

```
   reshape long
one_owner   f29_assetval_equip   g6_race_nathaw_owner_01
age_owner_01_r   f29_assetval_inv   g6_race_nathaw_owner_02
age_owner_02_r   f29_assetval_landbuild   g6_race_nathaw_owner_03
age_owner_03_r   f29_assetval_othbusprop   g6_race_nathaw_owner_04
age_owner_04_r   f29_assetval_other   g6_race_nathaw_owner_05
age_owner_05_r   f29_assetval_veh   g6_race_nathaw_owner_06
age_owner_06_r   f3_eq_invest_angels   g6_race_nathaw_owner_07
age_owner_07_r   f3_eq_invest_companies   g6_race_nathaw_owner_08
age_owner_08_r   f3_eq_invest_govt   g6_race_nathaw_owner_09
age_owner_09_r   f3_eq_invest_other   g6_race_nathaw_owner_10
age_owner_10_r   f3_eq_invest_parents   g6_race_nathaw_owner_11
age_owner_11_r   f3_eq_invest_spouse   g6_race_nathaw_owner_12
age_owner_12_r   f3_eq_invest_vent_cap   g6_race_nathaw_owner_13
age_owner_13_r   f30a_liab_acctpay   g6_race_nathaw_owner_14
age_owner_14_r   f30b_liab_pension   g6_race_nathaw_owner_15
age_owner_15_r   f30c_liab_other   g6_race_other_owner_01
b1_bus_start   f31_value_acctpay   g6_race_other_owner_02
c10_morelocations   f31_value_other   g6_race_other_owner_03
c11_num_locations   f31_value_pension   g6_race_other_owner_04
c12a_sba   f32_chap11_bankruptcy   g6_race_other_owner_05
c12b_fed_gov   f33_expected_growth   g6_race_other_owner_06
c12c_statelocal_gov   f34_future_revenue   g6_race_other_owner_07
c12d_non_profit   f4_eq_amt_angels   g6_race_other_owner_08
c12e_college_univ   f4_eq_amt_angels_allyrs   g6_race_other_owner_09
c12f_chamber_of_comm   f4_eq_amt_companies   g6_race_other_owner_10
c12g_for_profit_org   f4_eq_amt_companies_allyrs   g6_race_other_owner_11
c12h_other   f4_eq_amt_govt   g6_race_other_owner_12
c1z2_legal_status   f4_eq_amt_govt_allyrs   g6_race_other_owner_13
c2_owners   f4_eq_amt_other   g6_race_other_owner_14
c3a_owner_operators   f4_eq_amt_other_allyrs   g6_race_other_owner_15
c4_numowners_confirm   f4_eq_amt_parents   g6_race_white_owner_01
c5_num_employees   f4_eq_amt_parents_allyrs   g6_race_white_owner_02
c6_num_ft_employees   f4_eq_amt_spouse   g6_race_white_owner_03
c7_num_pt_employees   f4_eq_amt_spouse_allyrs   g6_race_white_owner_04
c8_primary_loc   f4_eq_amt_vent_cap   g6_race_white_owner_05
c9_loc_change_reason   f4_eq_amt_vent_cap_allyrs   g6_race_white_owner_06
classf   f5_perc_owned_angels   g6_race_white_owner_07
credrisk   f5_perc_owned_companies   g6_race_white_owner_08
cswgt_final   f5_perc_owned_govt   g6_race_white_owner_09
d1_a_new_product   f5_perc_owned_other   g6_race_white_owner_10
d1_b_new_to_market   f5_perc_owned_parents   g6_race_white_owner_11
d1a_provide_service   f5_perc_owned_spouse   g6_race_white_owner_12
d1b_provide_product   f5_perc_owned_vent_cap   g6_race_white_owner_13
d1c_a_regional   f5a_seek_equity   g6_race_white_owner_14
d1c_b_national   f6_perc_owned_owner_01   g6_race_white_owner_15
d1c_c_international   f6_perc_owned_owner_02   g6b_race_group_01
d1d_new_processes   f6_perc_owned_owner_03   g6b_race_group_02
d2_comp_advantage   f6_perc_owned_owner_04   g6b_race_group_03
d2a_compadv_comp_reason   f6_perc_owned_owner_05   g6b_race_group_04
d2a_compadv_govlab_reason   f6_perc_owned_owner_06   g6b_race_group_05
d2a_compadv_patents_reason   f6_perc_owned_owner_07   g6b_race_group_06
d2a_compadv_univ_reason   f6_perc_owned_owner_08   g6b_race_group_07
d2b_compadv_comp_strength   f6_perc_owned_owner_09   g6b_race_group_08
d2b_compadv_govlab_strength   f6_perc_owned_owner_10   g6b_race_group_09
d2b_compadv_patents_strength   f6_perc_owned_owner_11   g6b_race_group_10
d2b_compadv_univ_strength   f6_perc_owned_owner_12   g6b_race_group_11
d2c_compadv_cost_reason   f6_perc_owned_owner_13   g6b_race_group_12
d2c_compadv_design_reason   f6_perc_owned_owner_14   g6b_race_group_13
d2c_compadv_expertise_reason   f6_perc_owned_owner_15   g6b_race_group_14
d2c_compadv_marketing_reason   f6a_personal_use   g6b_race_group_15
```

```
d2c_compadv_price_reason   f6b_personal_use_amt   g7_native_born_owner_01
d2c_compadv_reputation_reason   f6check   g7_native_born_owner_02
d2c_compadv_speed_reason   f6z_family_owned   g7_native_born_owner_03
d2d_compadv_cost_strength   f7a_bus_credcard   g7_native_born_owner_04
d2d_compadv_design_strength   f7a_pers_credcard   g7_native_born_owner_05
d2d_compadv_expertise_strength   f7a_pers_loan_bank   g7_native_born_owner_06
d2d_compadv_marketing_strength   f7a_pers_loan_fam   g7_native_born_owner_07
d2d_compadv_price_strength   f7a_pers_loan_other   g7_native_born_owner_08
d2d_compadv_reput_strength   f7a_pers_other   g7_native_born_owner_09
d2d_compadv_speed_strength   f7b_bus_credcard_numused   g7_native_born_owner_10
d3_a_have_patent   f7b_pers_credcard_numused   g7_native_born_owner_11
d3_a_num_patent   f7b_pers_loan_bank_numused   g7_native_born_owner_12
d3_b_have_copyright   f7b_pers_loan_fam_numused   g7_native_born_owner_13
d3_b_num_copyright   f7b_pers_loan_other_numused   g7_native_born_owner_14
d3_c_have_trademark   f7b_pers_other_numused   g7_native_born_owner_15
d3_c_num_trademark   f8a_bus_credcard_line   g8_us_cit_owner_01
d4_a_lic_out_patent   f8a_pers_credcard_line   g8_us_cit_owner_02
d4_b_lic_out_copyright   f8b_bus_credcard_bal   g8_us_cit_owner_03
d4_c_lic_out_trademark   f8b_pers_credcard_bal   g8_us_cit_owner_04
d5_a_lic_in_patent   f8c_pers_loan_bank_amt   g8_us_cit_owner_05
d5_b_lic_in_copyright   f8c_pers_loan_fam_amt   g8_us_cit_owner_06
d5_c_lic_in_trademark   f8c_pers_loan_other_amt   g8_us_cit_owner_07
d5a_founded_newprod   f8c_pers_other_amt   g8_us_cit_owner_08
d5b_a_personaluse   f8d_pers_loan_bank_owed   g8_us_cit_owner_09
d5b_b_previousjob   f8d_pers_loan_fam_owed   g8_us_cit_owner_10
d5b_c_startingbus   f8d_pers_loan_other_owed   g8_us_cit_owner_11
d6_have_sales   f8d_pers_other_owed   g8_us_cit_owner_12
d7_perc_sales_bus   f9a_bus_credcard   g8_us_cit_owner_13
d7_perc_sales_govt   f9a_pers_credcard   g8_us_cit_owner_14
d7_perc_sales_indiv   f9a_pers_loan_bank   g8_us_cit_owner_15
d8_customer_locations   f9a_pers_loan_fam   g9_education_owner_01
d8a_international_sales   f9a_pers_loan_other   g9_education_owner_02
d8b_perc_international_sales   f9a_pers_other   g9_education_owner_03
d9_internet_sales   f9b_bus_credcard_numused   g9_education_owner_04
d9a_perc_internet_sales   f9b_pers_credcard_numused   g9_education_owner_05
e1_a_num_human_res   f9b_pers_loan_bank_numused   g9_education_owner_06
e1_b_num_sales   f9b_pers_loan_fam_numused   g9_education_owner_07
e1_c_num_exec_admin   f9b_pers_loan_other_numused   g9_education_owner_08
e1_d_num_resdev   f9b_pers_other_numused   g9_education_owner_09
e1_e_num_prod_manu   final_status_code   g9_education_owner_10
e1_f_num_gen_admin   fssp   g9_education_owner_11
e1_g_num_fin_admin   fstatus   g9_education_owner_12
e1_h_num_other   g10_gender_owner_01   g9_education_owner_13
e2a_ft_emp_bonus_plan   g10_gender_owner_02   g9_education_owner_14
e2a_ft_emp_flex_time   g10_gender_owner_03   g9_education_owner_15
e2a_ft_emp_hlth_plan   g10_gender_owner_04   hightech
e2a_ft_emp_other   g10_gender_owner_05   msa_final
e2a_ft_emp_paid_sick   g10_gender_owner_06   naics_code
e2a_ft_emp_paid_vaca   g10_gender_owner_07   owner_active_01
e2a_ft_emp_retire_plan   g10_gender_owner_08   owner_active_02
e2a_ft_emp_stock_own   g10_gender_owner_09   owner_active_03
e2a_ft_emp_tuit_reim   g10_gender_owner_10   owner_active_04
e2b_pt_emp_bonus_plan   g10_gender_owner_11   owner_active_05
e2b_pt_emp_flex_time   g10_gender_owner_12   owner_active_06
e2b_pt_emp_hlth_plan   g10_gender_owner_13   owner_active_07
e2b_pt_emp_other   g10_gender_owner_14   owner_active_08
e2b_pt_emp_paid_sick   g10_gender_owner_15   owner_active_09
e2b_pt_emp_paid_vaca   g10b_marital_status_01   owner_active_10
e2b_pt_emp_retire_plan   g10b_marital_status_02   owner_active_11
e2b_pt_emp_stock_own   g10b_marital_status_03   owner_active_12
e2b_pt_emp_tuit_reim   g10b_marital_status_04   owner_active_13
```

```
email   g10b_marital_status_05   owner_active_14
f10a_bus_credcard_line   g10b_marital_status_06   owner_active_15
f10a_pers_credcard_line   g10b_marital_status_07   paysc
f10b_bus_credcard_bal   g10b_marital_status_08   respondent
f10b_pers_credcard_bal   g10b_marital_status_09   sampleinfo_samplestrata
f10c_pers_loan_bank_amt   g10b_marital_status_10   sampleinfo_strata
f10c_pers_loan_fam_amt   g10b_marital_status_11   state_final
f10c_pers_loan_other_amt   g10b_marital_status_12   status
f10c_pers_other_amt   g10b_marital_status_13   techempl
f10d_pers_loan_bank_owed   g10b_marital_status_14   techgenr
f10d_pers_loan_fam_owed   g10b_marital_status_15   timezone
f10d_pers_loan_other_owed   g10c_net_worth_01   tot_asset_acct_rec_r
f10d_pers_other_owed   g10c_net_worth_02   tot_asset_cash_r
f11a_bus_cred_line   g10c_net_worth_03   tot_asset_equip_r
f11a_bus_credcard   g10c_net_worth_04   tot_asset_inv_r
f11a_bus_loans_bank   g10c_net_worth_05   tot_asset_landbuild_r
f11a_bus_loans_emp   g10c_net_worth_06   tot_asset_other_bus_prop_r
f11a_bus_loans_fam   g10c_net_worth_07   tot_asset_other_r
f11a_bus_loans_govt   g10c_net_worth_08   tot_asset_veh_r
f11a_bus_loans_nonbank   g10c_net_worth_09   tot_assets
f11a_bus_loans_other_bus   g10c_net_worth_10   tot_bus_credcard_bal_others_r
f11a_bus_loans_other_ind   g10c_net_worth_11   tot_bus_credcard_bal_resp_r
f11a_bus_loans_owner   g10c_net_worth_12   tot_bus_credcard_line_others_r
f11a_bus_other   g10c_net_worth_13   tot_bus_credcard_line_resp_r
f11a_busloans_otherbus_numused   g10c_net_worth_14   tot_bus_debt_other_r
f11b_bus_cred_line_numused   g10c_net_worth_15   tot_bus_debt_owed
f11b_bus_credcard_numused   g10d_personal_outlook   tot_bus_loans_bank_owed_r
f11b_bus_loans_bank_numused   g1a_emp_owner_01   tot_bus_loans_emp_owed_r
f11b_bus_loans_emp_numused   g1a_emp_owner_02   tot_bus_loans_fam_owed_r
f11b_bus_loans_fam_numused   g1a_emp_owner_03   tot_bus_loans_govt_owed_r
f11b_bus_loans_govt_numused   g1a_emp_owner_04   tot_bus_loans_nonbank_owed_r
f11b_bus_loans_nonbank_numused   g1a_emp_owner_05   tot_bus_loans_other_owed_r
f11b_bus_loans_owner_numused   g1a_emp_owner_06   tot_bus_loans_otherbus_owed_r
f11b_bus_other_numused   g1a_emp_owner_07   tot_bus_loans_otherind_owed_r
f11b_busloans_otherind_numused   g1a_emp_owner_08   tot_bus_loans_owner_owed_r
f12a_bus_cred_line   g1a_emp_owner_09   tot_cred_line_bus_bal_r
f12a_bus_credcard_line   g1a_emp_owner_10   tot_cred_line_bus_line_r
f12b_bus_cred_line_bal   g1a_emp_owner_11   tot_credcard_bal_bus_r
f12b_bus_credcard_bal   g1a_emp_owner_12   tot_credcard_line_bus_r
f12c_bus_loans_bank_amt   g1a_emp_owner_13   tot_debt
f12c_bus_loans_bus_amt   g1a_emp_owner_14   tot_debt_bus
f12c_bus_loans_emp_amt   g1a_emp_owner_15   tot_debt_liab_equity
f12c_bus_loans_fam_amt   g1b1_hours_owner_01   tot_debt_owed
f12c_bus_loans_govt_amt   g1b1_hours_owner_02   tot_debt_owed_owner_operators
f12c_bus_loans_nonbank_amt   g1b1_hours_owner_03   tot_debt_owner_operators
f12c_bus_loans_other_ind_amt   g1b1_hours_owner_04   tot_equity
f12c_bus_loans_owner_amt   g1b1_hours_owner_05   tot_equity_allyrs
f12c_bus_other_amt   g1b1_hours_owner_06   tot_equity_allyrs_owner_01_r
f12d_bus_loans_bank_owed   g1b1_hours_owner_07   tot_equity_allyrs_owner_02_r
f12d_bus_loans_bus_owed   g1b1_hours_owner_08   tot_equity_allyrs_owner_03_r
f12d_bus_loans_emp_owed   g1b1_hours_owner_09   tot_equity_allyrs_owner_04_r
f12d_bus_loans_fam_owed   g1b1_hours_owner_10   tot_equity_allyrs_owner_05_r
f12d_bus_loans_govt_owed   g1b1_hours_owner_11   tot_equity_allyrs_owner_06_r
f12d_bus_loans_nonbank_owed   g1b1_hours_owner_12   tot_equity_allyrs_owner_07_r
f12d_bus_loans_other_ind_owed   g1b1_hours_owner_13   tot_equity_allyrs_owner_08_r
f12d_bus_loans_owner_owed   g1b1_hours_owner_14   tot_equity_allyrs_owner_09_r
f12d_bus_other_owed   g1b1_hours_owner_15   tot_equity_allyrs_owner_10_r
f12e_collateral   g1b2_reasonfor_business   tot_equity_allyrs_owner_11_r
f12f_bus_real_estate   g2_work_exp_owner_01   tot_equity_allyrs_owner_12_r
f12f_business_equip_veh   g2_work_exp_owner_02   tot_equity_allyrs_owner_13_r
f12f_business_sec_dep   g2_work_exp_owner_03   tot_equity_allyrs_owner_14_r
```

```
f12f_intellectual_prop   g2_work_exp_owner_04   tot_equity_allyrs_owner_15_r
f12f_inventory_acctrec   g2_work_exp_owner_05   tot_equity_angels_allyrs_r
f12f_other   g2_work_exp_owner_06   tot_equity_angels_r
f12f_other_pers_assets   g2_work_exp_owner_07   tot_equity_companies_allyrs_r
f12f_pers_real_estate   g2_work_exp_owner_08   tot_equity_companies_r
f13_trade_fin   g2_work_exp_owner_09   tot_equity_govt_allyrs_r
f14a_trade_fin_amt   g2_work_exp_owner_10   tot_equity_govt_r
f14d_new_loans   g2_work_exp_owner_11   tot_equity_nonownerop_allyrs
f14e_approved_denied   g2_work_exp_owner_12   tot_equity_nonowneroperators
f14f_bus_credit_hist   g2_work_exp_owner_13   tot_equity_other_allyrs_r
f14f_inadeq_doc   g2_work_exp_owner_14   tot_equity_other_r
f14f_insuff_coll   g2_work_exp_owner_15   tot_equity_owner_01_r
f14f_loan_toolarge   g3a_oth_bus_owner_01   tot_equity_owner_02_r
f14f_new_bus   g3a_oth_bus_owner_02   tot_equity_owner_03_r
f14f_other   g3a_oth_bus_owner_03   tot_equity_owner_04_r
f14f_pers_credit_hist   g3a_oth_bus_owner_04   tot_equity_owner_05_r
f14f_restr_on_lending   g3a_oth_bus_owner_05   tot_equity_owner_06_r
f14g_didnotapply   g3a_oth_bus_owner_06   tot_equity_owner_07_r
f14h_loan_guarantees   g3a_oth_bus_owner_07   tot_equity_owner_08_r
f14i_economy_effect   g3a_oth_bus_owner_08   tot_equity_owner_09_r
f14j_most_challenging   g3a_oth_bus_owner_09   tot_equity_owner_10_r
f15_revenue   g3a_oth_bus_owner_10   tot_equity_owner_11_r
f16a_rev_amt   g3a_oth_bus_owner_11   tot_equity_owner_12_r
f17a_total_exp_amt   g3a_oth_bus_owner_12   tot_equity_owner_13_r
f18a_wage_exp_amt   g3a_oth_bus_owner_13   tot_equity_owner_14_r
f19_res_dev   g3a_oth_bus_owner_14   tot_equity_owner_15_r
f19a_res_dev_amt   g3a_oth_bus_owner_15   tot_equity_owner_operators
f19b_a_design   g3b_bus_same_ind_owner_01   tot_equity_owneroper_allyrs
f19b_b_investments   g3b_bus_same_ind_owner_02   tot_equity_parents_allyrs_r
f19b_c_brand_dev   g3b_bus_same_ind_owner_03   tot_equity_parents_r
f19b_d_org_dev   g3b_bus_same_ind_owner_04

f19b_e_worker_training   g3b_bus_same_ind_owner_05   tot_equity_spouse_r
f19b_f_other   g3b_bus_same_ind_owner_06   tot_equity_vent_cap_allyrs_r
f19c_a_design_amt   g3b_bus_same_ind_owner_07   tot_equity_vent_cap_r
f19c_b_investments_amt   g3b_bus_same_ind_owner_08   tot_expenses_r
f19c_c_brand_dev_amt   g3b_bus_same_ind_owner_09   tot_intang_assets_r
f19c_d_org_dev_amt   g3b_bus_same_ind_owner_10   tot_intangassets_branddev_r
f19c_e_worker_training_amt   g3b_bus_same_ind_owner_11   tot_intangassets_design_r
f19c_f_other_amt   g3b_bus_same_ind_owner_12   tot_intangassets_invest_r
f19c_intangassets_amt   g3b_bus_same_ind_owner_13   tot_intangassets_orgdev_r
f2_owner_amt_eq_invest_01   g3b_bus_same_ind_owner_14   tot_intangassets_other_r
f2_owner_amt_eq_invest_02   g3b_bus_same_ind_owner_15   tot_intangassets_wkrtrng_r
f2_owner_amt_eq_invest_03   g4_age_owner_01   tot_liab
f2_owner_amt_eq_invest_04   g4_age_owner_02   tot_liab_acct_pay_r
f2_owner_amt_eq_invest_05   g4_age_owner_03   tot_liab_other_r
f2_owner_amt_eq_invest_06   g4_age_owner_04   tot_liab_pension_r
f2_owner_amt_eq_invest_07   g4_age_owner_05   tot_loan_bank_bus_r
f2_owner_amt_eq_invest_08   g4_age_owner_06   tot_loan_emp_bus_r
f2_owner_amt_eq_invest_09   g4_age_owner_07   tot_loan_fam_bus_r
f2_owner_amt_eq_invest_10   g4_age_owner_08   tot_loan_govt_bus_r
f2_owner_amt_eq_invest_11   g4_age_owner_09   tot_loan_nonbank_bus_r
f2_owner_amt_eq_invest_12   g4_age_owner_10   tot_loan_other_bus_r
f2_owner_amt_eq_invest_13   g4_age_owner_11   tot_loan_other_ind_r
f2_owner_amt_eq_invest_14   g4_age_owner_12   tot_loan_owner_bus_r
f2_owner_amt_eq_invest_15   g4_age_owner_13   tot_loss_r
f2_owner_eq_invest_01   g4_age_owner_14   tot_pers_credcard_bal_others_r
f2_owner_eq_invest_02   g4_age_owner_15   tot_pers_credcard_bal_resp_r
f2_owner_eq_invest_03   g5_hisp_origin_owner_01   tot_pers_credcard_line_othrs_r
f2_owner_eq_invest_04   g5_hisp_origin_owner_02   tot_pers_credcard_line_resp_r
f2_owner_eq_invest_05   g5_hisp_origin_owner_03   tot_pers_debt_other_owners
```

```
f2_owner_eq_invest_06   g5_hisp_origin_owner_04   tot_pers_debt_owed_othrownrs
f2_owner_eq_invest_07   g5_hisp_origin_owner_05   tot_pers_debt_owed_resp
f2_owner_eq_invest_08   g5_hisp_origin_owner_06   tot_pers_debt_resp
f2_owner_eq_invest_09   g5_hisp_origin_owner_07   tot_pers_loan_bank_others_r
f2_owner_eq_invest_10   g5_hisp_origin_owner_08   tot_pers_loan_bank_owed_resp_r
f2_owner_eq_invest_11   g5_hisp_origin_owner_09   tot_pers_loan_bank_resp_r
f2_owner_eq_invest_12   g5_hisp_origin_owner_10   tot_pers_loan_fam_owed_resp_r
f2_owner_eq_invest_13   g5_hisp_origin_owner_11   tot_pers_loan_fam_resp_r
f2_owner_eq_invest_14   g5_hisp_origin_owner_12   tot_pers_loan_other_owners_r
f2_owner_eq_invest_15   g5_hisp_origin_owner_13   tot_pers_loan_other_resp_r
f2_owner_perc_own_01   g5_hisp_origin_owner_14   tot_pers_other_other_owners_r
f2_owner_perc_own_02   g5_hisp_origin_owner_15   tot_pers_other_owed_others_r
f2_owner_perc_own_03   g6_race_amind_owner_01   tot_pers_other_owed_resp_r
f2_owner_perc_own_04   g6_race_amind_owner_02   tot_pers_other_resp_r
f2_owner_perc_own_05   g6_race_amind_owner_03   tot_persloan_bank_owed_othrs_r
f2_owner_perc_own_06   g6_race_amind_owner_04   tot_persloan_fam_othrowners_r
f2_owner_perc_own_07   g6_race_amind_owner_05   tot_persloan_fam_owed_othrs_r
f2_owner_perc_own_08   g6_race_amind_owner_06   tot_persloan_other_owed_resp_r
f2_owner_perc_own_09   g6_race_amind_owner_07   tot_persloan_othr_owed_othrs_r
f2_owner_perc_own_10   g6_race_amind_owner_08   tot_personal_use_r
f2_owner_perc_own_11   g6_race_amind_owner_09   tot_profit_r
f2_owner_perc_own_12   g6_race_amind_owner_10   tot_res_dev_r
f2_owner_perc_own_13   g6_race_amind_owner_11   tot_revenue_r
f2_owner_perc_own_14   g6_race_amind_owner_12   tot_trade_finan_r
f2_owner_perc_own_15   g6_race_amind_owner_13   tot_wages_r
f2_ownr_amt_eqinvest_allyrs_01   g6_race_amind_owner_14   total_hours_owner_01_r
f2_ownr_amt_eqinvest_allyrs_02   g6_race_amind_owner_15   total_hours_owner_02_r
f2_ownr_amt_eqinvest_allyrs_03   g6_race_asian_owner_01   total_hours_owner_03_r
f2_ownr_amt_eqinvest_allyrs_04   g6_race_asian_owner_02   total_hours_owner_04_r
f2_ownr_amt_eqinvest_allyrs_05   g6_race_asian_owner_03   total_hours_owner_05_r
f2_ownr_amt_eqinvest_allyrs_06   g6_race_asian_owner_04   total_hours_owner_06_r
f2_ownr_amt_eqinvest_allyrs_07   g6_race_asian_owner_05   total_hours_owner_07_r
f2_ownr_amt_eqinvest_allyrs_08   g6_race_asian_owner_06   total_hours_owner_08_r
f2_ownr_amt_eqinvest_allyrs_09   g6_race_asian_owner_07   total_hours_owner_09_r
f2_ownr_amt_eqinvest_allyrs_10   g6_race_asian_owner_08   total_hours_owner_10_r
f2_ownr_amt_eqinvest_allyrs_11   g6_race_asian_owner_09   total_hours_owner_11_r
f2_ownr_amt_eqinvest_allyrs_12   g6_race_asian_owner_10   total_hours_owner_12_r
f2_ownr_amt_eqinvest_allyrs_13   g6_race_asian_owner_11   total_hours_owner_13_r
f2_ownr_amt_eqinvest_allyrs_14   g6_race_asian_owner_12   total_hours_owner_14_r
f2_ownr_amt_eqinvest_allyrs_15   g6_race_asian_owner_13   total_hours_owner_15_r
f20_mach   g6_race_asian_owner_14   website
f21_land_rent   g6_race_asian_owner_15
f22_mach_rent   g6_race_black_owner_01
f23_profit_or_loss   g6_race_black_owner_02
f24_profit_amt   g6_race_black_owner_03
f24_profitloss_amt   g6_race_black_owner_04
f26_loss_amt   g6_race_black_owner_05
f28a_asset_cash   g6_race_black_owner_06
f28b_asset_acct_rec   g6_race_black_owner_07
f28c_asset_inv   g6_race_black_owner_08
f28d_asset_equip   g6_race_black_owner_09
f28e_asset_landbuild   g6_race_black_owner_10
f28f_asset_veh   g6_race_black_owner_11
f28g_other_bus_prop   g6_race_black_owner_12
f28h_other_assets   g6_race_black_owner_13
f29_assetval_acctrec   g6_race_black_owner_14
f29_assetval_cash   g6_race_black_owner_15
zip_final      , i(mprid) j(y)   string ;
 #delimit cr
gen year=substr(y,2,1)
destring year, replace force
```

```
replace year=year+2004
drop y

replace wgt_1_long =. if year>2005
replace wgt_2_long =. if year>2006
replace wgt_3_long =. if year>2007
replace wgt_4_long =. if year>2008
replace wgt_5_long =. if year>2009
replace wgt_6_long =. if year>2010

* Save the Cross Sectional in long Format

cd " xxx:\KFS_Manual_and_Data "

save KFS8_LI_CS_Long,replace

gen  tempvar=year-2004

keep if tempvar<Duration
drop tempvar

keep if wgt_7_long>0

* Save the Seventh year panel in long Format

save KFS8_LI_L_Long,replace
```

3.2.2.2. Creating New Variables

This section illustrates the basics of creating new variables in KFS using Stata

3.2.2.2.1. Total Amount – Financial Variables

For all financial variables in the KFS, if the respondent did not provide the exact amount of the variable in dollars, the respondent was asked to provide a range of the amount instead. The range interval classes were standard across all the financial variables in the KFS. For businesses that reported the range rather than the exact value, replacing missing continuous values by the midpoints of the class interval is a common approach by KFS researchers.

The following table shows the names of the new financial variables and the variables used to create them.

Variable Name	Description
Tot_Equity_Owner_Operators	Sum of f2_owner_amt_eq_invest_*
Tot_Equity_NonOwnerOperators	Sum of f4_eq_amt_*
Tot_Equity	Tot_Equity_Owner_Operators + Tot_Equity_NonOwnerOperators
Tot_Equity_OwnerOper_AllYrs	Sum of f2_ownr_amt_eqinvest_allyrs_*
Tot_Equity_NonOwnerOp_AllYrs	Sum of f4_eq_amt_*_allyrs
Tot_Equity_AllYrs	Tot_Equity_OwnerOper_AllYrs + Tot_Equity_NonOwnerOp_AllYrs
Tot_Assets	Sum of f29_assetval_*
Tot_Liab	Sum of f31_value_*
Tot_Pers_Debt_Resp	Sum of f8b_pers_credcard_bal, f8b_bus_credcard_bal, f8c_pers_loan_bank_amt, f8c_pers_loan_fam_amt, f8c_pers_loan_other_amt, f8c_pers_other_amt
Tot_Pers_Debt_Other_Owners	Sum of f10b_pers_credcard_bal, f10c_pers_loan_bank_amt, f10b_bus_credcard_bal, f10c_pers_loan_fam_amt, f10c_pers_loan_other_amt, f10c_pers_other_amt
Tot_Debt_Owner_Operators	Tot_Pers_Debt_Resp + Tot_Pers_Debt_Other_Owners

Tot_Debt_Bus	Sum of f12b_bus_credcard_bal, f12c_bus_loans_bank_amt, f12b_bus_cred_line_bal, f12c_bus_loans_nonbank_amt, f12c_bus_loans_fam_amt, f12c_bus_loans_govt_amt, f12c_bus_loans_emp_amt, f12c_bus_loans_other_ind_amt, f12c_bus_loans_owner_amt, f12c_bus_loans_bus_amt, f12c_bus_other_amt
Tot_Debt	Tot_Debt_Owner_Operators + Tot_Debt_Bus
Tot_Pers_Debt_Owed_Resp	Sum of f8b_pers_credcard_bal, f8b_bus_credcard_bal, f8d_pers_loan_bank_owed, f8d_pers_loan_fam_owed, f8d_pers_loan_other_owed, f8d_pers_other_owed
Tot_Pers_Debt_Owed_OthrOwnrs	f10b_pers_credcard_bal, f10b_bus_credcard_bal, f10d_pers_loan_bank_owed, f10d_pers_loan_fam_owed, f10d_pers_loan_other_owed, f10d_pers_other_owed
Tot_Debt_Owed_Owner_Operators	Tot_Pers_Debt_Owed_Resp + Tot_Pers_Debt_Owed_OthrOwnrs
Tot_Bus_Debt_Owed	Sum of f12b_bus_cred_line_bal, f12b_bus_credcard_bal, f12d_bus_loans_bank_owed, f12d_bus_loans_nonbank_owed, f12d_bus_loans_emp_owed, f12d_bus_loans_fam_owed, f12d_bus_loans_govt_owed, f12d_bus_loans_other_ind_owed, f12d_bus_loans_owner_owed, f12d_bus_loans_bus_owed, f12d_bus_other_owed
Tot_Debt_Owed	Tot_Debt_Owed_Owner_Operators + Tot_Bus_Debt_Owed
Net_Profit	-1*f26_loss_amt Or f24_profit_amt

3.2.2.2.2. Primary Owner and Active-Owner-Operators Characteristics

To determine the owner(s) demographics at the firm level, researchers can use the primary owner characteristic or use the average of the characteristics across all active owner-operators.

For the primary owner approach, Robb's Stata code was used to create the primary owner file. The file is available at NORC under the name "primary_owner.dta"; the file has two variables MPRID and primary_owner.[1] The primary_owner variable indicates the number of the primary owner of the business at the baseline survey. The following table shows the names of the new variables for primary owner characteristics.

Variable Name	Description
PO_emp	Primary owner is paid employee (1=yes,0=no)
PO_hours	Average weekly working hours by Primary Owner
PO_work_exp	Primary Owner work experience (years)
PO_oth_bus_owner	Number of other new businesses started by primary owner
PO_bus_same_ind	New business(es) in the same industry as the current business 1=yes,0=no)
PO_age_owner	Age of primary owner
PO_hisp_origin	Primary owner is Hispanic or Latino origin (1=yes,0=no)
PO_race_group	Primary owner race (1=American Indian or Alaska Native, 2=Native Hawaiian or Other Pacific Islander, 3= Asian, 4=Black or African American, 5=White, 6=Other Races or Mixed Race)
PO_race_amind_owner	Primary owner is American Indian or Alaska Native(1=yes, 0=no)
PO_race_asian_owner	Primary owner is Asian(1=yes, 0=no)
Variable Name	
PO_race_black_owner	Primary owner is Black(1=yes, 0=no)
PO_race_nathaw_owner	Primary owner is Native Hawaiian or Other Pacific Islander(1=yes, 0=no)
PO_race_other_owner	Primary owner is Other Races or Mixed Race(1=yes, 0=no)
PO_race_white_owner	Primary owner is White(1=yes, 0=no)
PO_native_born	Primary owner was born in the United States(1=yes, 0=no)
PO_us_cit	Primary owner is a U.S. citizen(1=yes, 0=no)

[1] For firms with multiple owners, the primary owner is identified as the one with the largest equity share. In cases where two or more owners owned equal shares, hours worked and a series of other variables were used to create a rank of owners to define the primary owner.

PO_education	Primary owner highest level of education (1= Less than 9th grade; 2= Some high school, but no diploma; 3= High school graduate; 4= Technical, trade or vocational degree; 5= Some college, but no degree; 6= Associate's degree; 7= Bachelor's degree: 8= Some graduate school but no degree; 9= Master's degree; 10= Professional school or doctorate)
PO_gender	Primary owner gender (1=male, 0=female)

For the average of the characteristics across all active owner-operators approach, the average characteristics of all active owner-operators are calculated at the firm level.

In addition, we show how to calculate the diversity/similarity index among the managing team. We use Blau's index to compute diversity/similarity for a particular characteristic among active owner-operators.

$$\text{Diversity Index } A = 1 - \sum_{i=1}^{n} P_i^2$$

$$\text{Similarity Index } A = \sum_{i=1}^{n} P_i^2$$

where P is the proportion of owners who have the same characteristics and n is the number of different characteristics.

The following table shows the names of the new variables for all active owner-operators.

Variable Name	Description
OO_emp_owner	Percentage of active owner-operators who are paid employees
OO_hours_owner	Average weekly working hours by active owner-operators
OO_work_exp_owner	Average work experience of active owner-operators
OO_oth_bus_owner	Average number of other new businesses started by active owner-operators
OO_bus_same_ind_owner	Percentage of new business(es) in the same industry as the current business
OO_age_owner	Average age of active owner-operators
OO_hisp_origin_owner	Percentage of Hispanic or Latino origin active owner-operators
OO_race_amind_owner	Percentage of American Indian or Alaska Native active owner-operators
OO_race_asian_owner	Percentage of Asian active owner-operators
OO_race_black_owner	Percentage of Black active owner-operators
OO_race_nathaw_owner	Percentage of Native Hawaiian or Other Pacific Islander active owner-operators

Variable Name	Description
OO_race_other_owner	Percentage of Other Races or Mixed Race active owner-operators
OO_race_white_owner	Percentage of White active owner-operators
OO_native_born_owner	Percentage of active owner-operators born in the United States
OO_us_cit_owner	Percentage of active owner-operators who are U.S. citizens
OO_education_owner	Average highest level of education of active owner-operators
md_education_owner	Median highest level of education of active owner-operators
OO_D_education_owner	Percentage of active owner-operators who hold a college degree or above
OO_ gender_owner	Percentage of males active owner-operators
Race_similarity	Blau's index
Race_diversity	Blau's index
Gender_similarity	Blau's index
Gender_diversity	Blau's index

3.2.2.2.3. Business level Characteristics

In addition to the owner level characteristics, we show how to create business level variables. The following table shows the names of the new variables for the business characteristics.

Variable Name	Description
Home_Based	Home based Business(1=yes, 0=no)
Sole_Proprietorship	Sole Proprietorship Business(1=yes, 0=no)
Comp_advantage	Having competitive advantage(1=yes, 0=no)
Have_IP	The business have Patent or Copyright or Trademark (1=yes, 0=no)
Full_Part_Time_Employees	=c5_num_employees
Full_Time_Employees	=c6_num_ft_employees
Part_Time_Employees	=c7_num_pt_employees
Employee_Owner	Total number of active owner-operators who are paid employees
Total_Employees	Full_Part_Time_Employees + Employee_Owner

3.2.2.2.4. Stata Code: Cross Sectional in Wide Format

The following Stata code will create the new variables using the KFS wide format file "KFS8_LI". The code will save the new cross sectional (n=4928) file under the name "KFS8_CS_w1.dta".

```
clear
clear mata
clear matrix
capture log close
set more off
set maxvar 32767
program drop _all
adopath + " xxx:\KFS_Manual_and_Data\Farhat_Robb_Commands\"
cd " xxx:\KFS_Manual_and_Data "
use KFS8_LI,clear

set more off
* Data Management Var
rename one_ower_00 one_owner_0
rename one_ower_01 one_owner_1
rename one_ower_02 one_owner_2
rename one_ower_03 one_owner_3
rename one_ower_04 one_owner_4
rename one_ower_05 one_owner_5
rename one_ower_06 one_owner_6
rename one_ower_07 one_owner_7


*Replacing missing continuous values by the midpoints of the class intervals

foreach totfin in tot_*_r_* {
recode `totfin'     (0=0)(1=250)(2=750)(3=2000)(4=4000)(5=7500) (6=17500) (7=62500) (8=550000) (9=1000000)
    }

global owners_1_15 "01 02 03 04 05 06 07 08 09 10 11 12 13 14 15"

forvalues i = 0/7 {
foreach ow in $owners_1_15  {
replace  f2_owner_amt_eq_invest_`ow'_`i'=tot_equity_owner_`ow'_r_`i'   if  f2_owner_amt_eq_invest_`ow'_`i'==.
replace               f2_ownr_amt_eqinvest_allyrs_`ow'_`i'=tot_equity_allyrs_owner_`ow'_r_`i'                        if
```

```
f2_ownr_amt_eqinvest_allyrs_`ow'_`i'==.
}
}

forvalues i = 0/7 {
replace  f4_eq_amt_angels_`i'=tot_equity_angels_r_`i'   if  f4_eq_amt_angels_`i'==.
replace  f4_eq_amt_angels_allyrs_`i'=tot_equity_angels_allyrs_r_`i'   if  f4_eq_amt_angels_allyrs_`i'==.
replace  f4_eq_amt_companies_`i'=tot_equity_companies_r_`i'   if  f4_eq_amt_companies_`i'==.
replace  f4_eq_amt_companies_allyrs_`i'=tot_equity_companies_allyrs_r_`i'   if  f4_eq_amt_companies_allyrs_`i'==.
replace  f4_eq_amt_govt_`i'=tot_equity_govt_r_`i'   if  f4_eq_amt_govt_`i'==.
replace  f4_eq_amt_govt_allyrs_`i'=tot_equity_govt_allyrs_r_`i'   if  f4_eq_amt_govt_allyrs_`i'==.
replace  f4_eq_amt_other_`i'=tot_equity_other_r_`i'   if  f4_eq_amt_other_`i'==.
replace  f4_eq_amt_other_allyrs_`i'=tot_equity_other_allyrs_r_`i'   if  f4_eq_amt_other_allyrs_`i'==.
replace  f4_eq_amt_parents_`i'=tot_equity_parents_r_`i'   if  f4_eq_amt_parents_`i'==.
replace  f4_eq_amt_parents_allyrs_`i'=tot_equity_parents_allyrs_r_`i'   if  f4_eq_amt_parents_allyrs_`i'==.
replace  f4_eq_amt_spouse_`i'=tot_equity_spouse_r_`i'   if  f4_eq_amt_spouse_`i'==.
replace  f4_eq_amt_spouse_allyrs_`i'=tot_equity_spouse_allyrs_r_`i'   if  f4_eq_amt_spouse_allyrs_`i'==.
replace  f4_eq_amt_vent_cap_`i'=tot_equity_vent_cap_r_`i'   if  f4_eq_amt_vent_cap_`i'==.
replace  f4_eq_amt_vent_cap_allyrs_`i'=tot_equity_vent_cap_allyrs_r_`i'   if  f4_eq_amt_vent_cap_allyrs_`i'==.
replace  f6b_personal_use_amt_`i'=tot_personal_use_r_`i'   if  f6b_personal_use_amt_`i'==.
replace  f8a_bus_credcard_line_`i'=tot_bus_credcard_line_resp_r_`i'   if  f8a_bus_credcard_line_`i'==.
replace  f8b_bus_credcard_bal_`i'=tot_bus_credcard_bal_resp_r_`i'   if  f8b_bus_credcard_bal_`i'==.
replace  f8a_pers_credcard_line_`i'=tot_pers_credcard_line_resp_r_`i'   if  f8a_pers_credcard_line_`i'==.
replace  f8b_pers_credcard_bal_`i'=tot_pers_credcard_bal_resp_r_`i'   if  f8b_pers_credcard_bal_`i'==.
replace  f8c_pers_loan_bank_amt_`i'=tot_pers_loan_bank_resp_r_`i'   if  f8c_pers_loan_bank_amt_`i'==.
replace  f8d_pers_loan_bank_owed_`i'=tot_pers_loan_bank_owed_resp_r_`i'   if  f8d_pers_loan_bank_owed_`i'==.
replace  f8c_pers_loan_fam_amt_`i'=tot_pers_loan_fam_resp_r_`i'   if  f8c_pers_loan_fam_amt_`i'==.
replace  f8d_pers_loan_fam_owed_`i'=tot_pers_loan_fam_owed_resp_r_`i'   if  f8d_pers_loan_fam_owed_`i'==.
replace  f8c_pers_loan_other_amt_`i'=tot_pers_loan_other_resp_r_`i'   if  f8c_pers_loan_other_amt_`i'==.
replace  f8d_pers_loan_other_owed_`i'=tot_persloan_other_owed_resp_r_`i'   if  f8d_pers_loan_other_owed_`i'==.
replace  f8c_pers_other_amt_`i'=tot_pers_other_resp_r_`i'   if  f8c_pers_other_amt_`i'==.
replace  f8d_pers_other_owed_`i'=tot_pers_other_owed_resp_r_`i'   if  f8d_pers_other_owed_`i'==.
replace  f10a_bus_credcard_line_`i'=tot_bus_credcard_line_others_r_`i'   if  f10a_bus_credcard_line_`i'==.
replace  f10b_bus_credcard_bal_`i'=tot_bus_credcard_bal_others_r_`i'   if  f10b_bus_credcard_bal_`i'==.
replace  f10a_pers_credcard_line_`i'=tot_pers_credcard_line_othrs_r_`i'   if  f10a_pers_credcard_line_`i'==.
replace  f10b_pers_credcard_bal_`i'=tot_pers_credcard_bal_others_r_`i'   if  f10b_pers_credcard_bal_`i'==.
replace  f10c_pers_loan_bank_amt_`i'=tot_pers_loan_bank_others_r_`i'   if  f10c_pers_loan_bank_amt_`i'==.
replace  f10d_pers_loan_bank_owed_`i'=tot_persloan_bank_owed_othrs_r_`i'   if  f10d_pers_loan_bank_owed_`i'==.
replace  f10c_pers_loan_fam_amt_`i'=tot_persloan_fam_othrowners_r_`i'   if  f10c_pers_loan_fam_amt_`i'==.
replace  f10d_pers_loan_fam_owed_`i'=tot_persloan_fam_owed_othrs_r_`i'   if  f10d_pers_loan_fam_owed_`i'==.
replace  f10c_pers_loan_other_amt_`i'=tot_pers_loan_other_owners_r_`i'   if  f10c_pers_loan_other_amt_`i'==.
replace  f10d_pers_loan_other_owed_`i'=tot_persloan_othr_owed_othrs_r_`i'   if  f10d_pers_loan_other_owed_`i'==.
```

```
replace  f10c_pers_other_amt_`i'=tot_pers_other_other_owners_r_`i'   if  f10c_pers_other_amt_`i'==.
replace  f10d_pers_other_owed_`i'=tot_pers_other_owed_others_r_`i'   if  f10d_pers_other_owed_`i'==.
replace  f12a_bus_cred_line_`i'=tot_cred_line_bus_line_r_`i'   if  f12a_bus_cred_line_`i'==.
replace  f12b_bus_cred_line_bal_`i'=tot_cred_line_bus_bal_r_`i'   if  f12b_bus_cred_line_bal_`i'==.
replace  f12a_bus_credcard_line_`i'=tot_credcard_line_bus_r_`i'   if  f12a_bus_credcard_line_`i'==.
replace  f12b_bus_credcard_bal_`i'=tot_credcard_bal_bus_r_`i'   if  f12b_bus_credcard_bal_`i'==.
replace  f12c_bus_loans_bank_amt_`i'=tot_loan_bank_bus_r_`i'   if  f12c_bus_loans_bank_amt_`i'==.
replace  f12d_bus_loans_bank_owed_`i'=tot_bus_loans_bank_owed_r_`i'   if  f12d_bus_loans_bank_owed_`i'==.
replace  f12c_bus_loans_nonbank_amt_`i'=tot_loan_nonbank_bus_r_`i'   if  f12c_bus_loans_nonbank_amt_`i'==.
replace  f12d_bus_loans_nonbank_owed_`i'=tot_bus_loans_nonbank_owed_r_`i'   if  f12d_bus_loans_nonbank_owed_`i'==.
replace  f12c_bus_loans_emp_amt_`i'=tot_loan_emp_bus_r_`i'   if  f12c_bus_loans_emp_amt_`i'==.
replace  f12d_bus_loans_emp_owed_`i'=tot_bus_loans_emp_owed_r_`i'   if  f12d_bus_loans_emp_owed_`i'==.
replace  f12c_bus_loans_fam_amt_`i'=tot_loan_fam_bus_r_`i'   if  f12c_bus_loans_fam_amt_`i'==.
replace  f12d_bus_loans_fam_owed_`i'=tot_bus_loans_fam_owed_r_`i'   if  f12d_bus_loans_fam_owed_`i'==.
replace  f12c_bus_loans_govt_amt_`i'=tot_loan_govt_bus_r_`i'   if  f12c_bus_loans_govt_amt_`i'==.
replace  f12d_bus_loans_govt_owed_`i'=tot_bus_loans_govt_owed_r_`i'   if  f12d_bus_loans_govt_owed_`i'==.
replace  f12c_bus_loans_other_ind_amt_`i'=tot_loan_other_ind_r_`i'   if  f12c_bus_loans_other_ind_amt_`i'==.
replace  f12d_bus_loans_other_ind_owed_`i'=tot_bus_loans_otherind_owed_r_`i'   if  f12d_bus_loans_other_ind_owed_`i'==.
replace  f12c_bus_loans_owner_amt_`i'=tot_loan_owner_bus_r_`i'   if  f12c_bus_loans_owner_amt_`i'==.
replace  f12d_bus_loans_owner_owed_`i'=tot_bus_loans_owner_owed_r_`i'   if  f12d_bus_loans_owner_owed_`i'==.
replace  f12c_bus_loans_bus_amt_`i'=tot_loan_other_bus_r_`i'   if  f12c_bus_loans_bus_amt_`i'==.
replace  f12d_bus_loans_bus_owed_`i'=tot_bus_loans_otherbus_owed_r_`i'   if  f12d_bus_loans_bus_owed_`i'==.
replace  f12c_bus_other_amt_`i'=tot_bus_debt_other_r_`i'   if  f12c_bus_other_amt_`i'==.
replace  f12d_bus_other_owed_`i'=tot_bus_loans_other_owed_r_`i'   if  f12d_bus_other_owed_`i'==.
replace  f14a_trade_fin_amt_`i'=tot_trade_finan_r_`i'   if  f14a_trade_fin_amt_`i'==.
replace  f16a_rev_amt_`i'=tot_revenue_r_`i'   if  f16a_rev_amt_`i'==.
replace  f17a_total_exp_amt_`i'=tot_expenses_r_`i'   if  f17a_total_exp_amt_`i'==.
replace  f18a_wage_exp_amt_`i'=tot_wages_r_`i'   if  f18a_wage_exp_amt_`i'==.
replace  f19a_res_dev_amt_`i'=tot_res_dev_r_`i'   if  f19a_res_dev_amt_`i'==.
replace  f19c_a_design_amt_`i'=tot_intangassets_design_r_`i'   if  f19c_a_design_amt_`i'==.
replace  f19c_b_investments_amt_`i'=tot_intangassets_invest_r_`i'   if  f19c_b_investments_amt_`i'==.
replace  f19c_c_brand_dev_amt_`i'=tot_intangassets_branddev_r_`i'   if  f19c_c_brand_dev_amt_`i'==.
replace  f19c_d_org_dev_amt_`i'=tot_intangassets_orgdev_r_`i'   if  f19c_d_org_dev_amt_`i'==.
replace  f19c_e_worker_training_amt_`i'=tot_intangassets_wkrtrng_r_`i'   if  f19c_e_worker_training_amt_`i'==.
replace  f19c_f_other_amt_`i'=tot_intangassets_other_r_`i'   if  f19c_f_other_amt_`i'==.
replace  f19c_intangassets_amt_`i'=tot_intang_assets_r_`i'   if  f19c_intangassets_amt_`i'==.
replace  f24_profit_amt_`i'=tot_profit_r_`i'   if  f24_profit_amt_`i'==.
replace  f26_loss_amt_`i'=tot_loss_r_`i'   if  f26_loss_amt_`i'==.
replace  f29_assetval_cash_`i'=tot_asset_cash_r_`i'   if  f29_assetval_cash_`i'==.
replace  f29_assetval_acctrec_`i'=tot_asset_acct_rec_r_`i'   if  f29_assetval_acctrec_`i'==.
replace  f29_assetval_inv_`i'=tot_asset_inv_r_`i'   if  f29_assetval_inv_`i'==.
replace  f29_assetval_equip_`i'=tot_asset_equip_r_`i'   if  f29_assetval_equip_`i'==.
```

```
replace  f29_assetval_landbuild_`i'=tot_asset_landbuild_r_`i'   if  f29_assetval_landbuild_`i'==.
replace  f29_assetval_veh_`i'=tot_asset_veh_r_`i'   if  f29_assetval_veh_`i'==.
replace  f29_assetval_othbusprop_`i'=tot_asset_other_bus_prop_r_`i'   if  f29_assetval_othbusprop_`i'==.
replace  f29_assetval_other_`i'=tot_asset_other_r_`i'   if  f29_assetval_other_`i'==.
replace  f31_value_acctpay_`i'=tot_liab_acct_pay_r_`i'   if  f31_value_acctpay_`i'==.
replace  f31_value_pension_`i'=tot_liab_pension_r_`i'   if  f31_value_pension_`i'==.
replace  f31_value_other_`i'=tot_liab_other_r_`i'   if  f31_value_other_`i'==.
}

forvalues i = 0/7 {
egen   Tot_Equity_Owner_Operators_`i'=  rowtotal(    f2_owner_amt_eq_invest_*_`i'      )      , missing
egen   Tot_Equity_OwnerOper_AllYrs_`i'=rowtotal(     f2_ownr_amt_eqinvest_allyrs_*_`i' ), missing
}


forvalues i = 0/7 {
foreach ow in $owners_1_15  {
/* Recode legitimate (hard) missing values */
replace Tot_Equity_Owner_Operators_`i' =.a if classf_`i'<6
replace Tot_Equity_OwnerOper_AllYrs_`i'=.a if classf_`i'<6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Equity_Owner_Operators_`i' =.   if  f2_owner_amt_eq_invest_`ow'_`i'==.
replace Tot_Equity_OwnerOper_AllYrs_`i'=.   if  f2_ownr_amt_eqinvest_allyrs_`ow'_`i'==.
}
}

global List1 "spouse parents angels companies govt vent_cap other"

forvalues i = 0/7 {
egen  Tot_Equity_NonOwnerOperators_`i'=rowtotal(f4_eq_amt_angels_`i' f4_eq_amt_companies_`i' f4_eq_amt_govt_`i' ///
f4_eq_amt_other_`i' f4_eq_amt_parents_`i' f4_eq_amt_spouse_`i'  f4_eq_amt_vent_cap_`i')      , missing
egen  Tot_Equity_NonOwnerOp_AllYrs_`i'=rowtotal(f4_eq_amt_*_allyrs_`i')   , missing
/* Recode legitimate (hard) missing values */
replace Tot_Equity_NonOwnerOperators_`i' =.a if classf_`i'<6
replace Tot_Equity_NonOwnerOperators_`i' =.a if  c1z2_legal_status_`i'==1
replace Tot_Equity_NonOwnerOperators_`i' =.a if  one_owner_`i'==1
replace Tot_Equity_NonOwnerOp_AllYrs_`i' =.a if classf_`i'<6
replace Tot_Equity_NonOwnerOp_AllYrs_`i' =.a if  c1z2_legal_status_`i'==1
replace Tot_Equity_NonOwnerOp_AllYrs_`i' =.a if  one_owner_`i'==1
}

forvalues i = 0/7 {
```

```
foreach name in $List1  {
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Equity_NonOwnerOperators_`i' =. if f4_eq_amt_`name'_`i'==.
replace Tot_Equity_NonOwnerOp_AllYrs_`i' =. if f4_eq_amt_`name'_allyrs_`i'==.
}
}

forvalues i = 0/7 {
egen  Tot_Equity_`i'       =rowtotal(Tot_Equity_Owner_Operators_`i' Tot_Equity_NonOwnerOperators_`i'), missing
egen  Tot_Equity_AllYrs_`i'=rowtotal(Tot_Equity_OwnerOper_AllYrs_`i' Tot_Equity_NonOwnerOp_AllYrs_`i'), missing
/* Recode legitimate (hard) missing values */
replace Tot_Equity_`i'       =.a if classf_`i'<6
replace Tot_Equity_AllYrs_`i'=.a if classf_`i'<6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Equity_`i'       =. if Tot_Equity_Owner_Operators_`i'==.
replace Tot_Equity_`i'       =. if Tot_Equity_NonOwnerOperators_`i'==.
replace Tot_Equity_AllYrs_`i'=. if Tot_Equity_NonOwnerOperators_`i'==.
replace Tot_Equity_AllYrs_`i'=. if Tot_Equity_NonOwnerOp_AllYrs_`i'==.
}

forvalues i = 0/7 {
egen  Tot_Assets_`i'=rowtotal(f29_assetval_*_`i')    , missing
egen  Tot_Liab_`i'=rowtotal(f31_value_*_`i'), missing
/* Recode legitimate (hard) missing values */
replace Tot_Assets_`i' =.a if classf_`i'<6
replace Tot_Liab_`i'   =.a if classf_`i'<6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Assets_`i' =. if f29_assetval_acctrec_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_cash_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_equip_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_inv_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_landbuild_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_othbusprop_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_other_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_veh_`i'==.
replace Tot_Liab_`i' =. if  f31_value_acctpay_`i'==.
replace Tot_Liab_`i' =. if  f31_value_other_`i'==.
replace Tot_Liab_`i' =. if  f31_value_pension_`i'==.
}

forvalues i = 0/7 {
egen    Tot_Pers_Debt_Resp_`i'=rowtotal(f8b_pers_credcard_bal_`i'  f8b_bus_credcard_bal_`i'  f8c_pers_loan_bank_amt_`i'
```

```
f8c_pers_loan_fam_amt_`i' f8c_pers_loan_other_amt_`i' f8c_pers_other_amt_`i'),missing
egen                    Tot_Pers_Debt_Owed_Resp_`i'=rowtotal(f8b_pers_credcard_bal_`i'          f8b_bus_credcard_bal_`i'
f8d_pers_loan_bank_owed_`i' f8d_pers_loan_fam_owed_`i' f8d_pers_loan_other_owed_`i' f8d_pers_other_owed_`i'),missing
/* Recode legitimate (hard) missing values */
replace Tot_Pers_Debt_Resp_`i'      =.a if classf_`i'<6
replace Tot_Pers_Debt_Owed_Resp_`i' =.a if classf_`i'<6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Pers_Debt_Resp_`i' =. if f8b_pers_credcard_bal_`i'==.
replace Tot_Pers_Debt_Resp_`i' =. if f8b_bus_credcard_bal_`i'==.
replace Tot_Pers_Debt_Resp_`i' =. if f8c_pers_loan_bank_amt_`i'==.
replace Tot_Pers_Debt_Resp_`i' =. if f8c_pers_loan_fam_amt_`i'==.
replace Tot_Pers_Debt_Resp_`i' =. if f8c_pers_loan_other_amt_`i'==.
replace Tot_Pers_Debt_Resp_`i' =. if f8c_pers_other_amt_`i'==.
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8b_pers_credcard_bal_`i'==.
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8b_bus_credcard_bal_`i'==.
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8d_pers_loan_bank_owed_`i'==.
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8d_pers_loan_fam_owed_`i'==.
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8d_pers_loan_other_owed_`i'==.
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8d_pers_other_owed_`i'==.
}

forvalues i = 0/7 {
egen               Tot_Pers_Debt_Other_Owners_`i'=rowtotal(f10b_pers_credcard_bal_`i'        f10c_pers_loan_bank_amt_`i'
f10b_bus_credcard_bal_`i' f10c_pers_loan_fam_amt_`i' f10c_pers_loan_other_amt_`i'  f10c_pers_other_amt_`i'),missing
egen               Tot_Pers_Debt_Owed_OthrOwnrs_`i'=rowtotal(f10b_pers_credcard_bal_`i'        f10b_bus_credcard_bal_`i'
f10d_pers_loan_bank_owed_`i'                  f10d_pers_loan_fam_owed_`i'                  f10d_pers_loan_other_owed_`i'
f10d_pers_other_owed_`i'),missing
/* Recode legitimate (hard) missing values */
replace Tot_Pers_Debt_Other_Owners_`i'  =.a if classf_`i'<6
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i'=.a if classf_`i'<6
replace Tot_Pers_Debt_Other_Owners_`i'  =.a if c4_numowners_confirm_`i'<2
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i'=.a if c4_numowners_confirm_`i'<2
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10b_pers_credcard_bal_`i'==.
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10b_bus_credcard_bal_`i'==.
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10c_pers_loan_bank_amt_`i'==.
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10c_pers_loan_fam_amt_`i'==.
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10c_pers_loan_other_amt_`i'==.
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10c_pers_other_amt_`i'==.
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10b_pers_credcard_bal_`i'==.
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10b_bus_credcard_bal_`i'==.
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10d_pers_loan_bank_owed_`i'==.
```

```
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10d_pers_loan_fam_owed_`i'==.
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10d_pers_loan_other_owed_`i'==.
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10d_pers_other_owed_`i'==.
}

forvalues i = 0/7 {
egen  Tot_Debt_Owner_Operators_`i'=rowtotal(Tot_Pers_Debt_Resp_`i' Tot_Pers_Debt_Other_Owners_`i'),missing
egen  Tot_Debt_Owed_Owner_Operators_`i'=rowtotal(Tot_Pers_Debt_Owed_Resp_`i' Tot_Pers_Debt_Owed_OthrOwnrs_`i'),missing
/* Recode legitimate (hard) missing values */
replace Tot_Debt_Owner_Operators_`i'   =.a if classf_`i'<6
replace Tot_Debt_Owed_Owner_Operators_`i'=.a if classf_`i'<6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Debt_Owner_Operators_`i'   =. if Tot_Pers_Debt_Resp_`i'==.
replace Tot_Debt_Owner_Operators_`i'   =. if Tot_Pers_Debt_Other_Owners_`i'==.
replace Tot_Debt_Owed_Owner_Operators_`i'=. if Tot_Pers_Debt_Owed_Resp_`i'==.
replace Tot_Debt_Owed_Owner_Operators_`i'=. if Tot_Pers_Debt_Owed_OthrOwnrs_`i'==.
}

forvalues i = 0/7 {
egen    Tot_Debt_Bus_`i'=rowtotal(f12b_bus_credcard_bal_`i'    f12c_bus_loans_bank_amt_`i'    f12b_bus_cred_line_bal_`i'
f12c_bus_loans_nonbank_amt_`i'    f12c_bus_loans_fam_amt_`i'    f12c_bus_loans_govt_amt_`i'    f12c_bus_loans_emp_amt_`i'
f12c_bus_loans_other_ind_amt_`i'                f12c_bus_loans_owner_amt_`i'                f12c_bus_loans_bus_amt_`i'
      f12c_bus_other_amt_`i'),missing
egen  Tot_Bus_Debt_Owed_`i'=rowtotal(f12b_bus_cred_line_bal_`i'   f12b_bus_credcard_bal_`i'  f12d_bus_loans_bank_owed_`i'
f12d_bus_loans_nonbank_owed_`i'  f12d_bus_loans_emp_owed_`i'    f12d_bus_loans_fam_owed_`i'  f12d_bus_loans_govt_owed_`i'
f12d_bus_loans_other_ind_owed_`i'          f12d_bus_loans_owner_owed_`i'                f12d_bus_loans_bus_owed_`i'
f12d_bus_other_owed_`i'),missing
/* Recode legitimate (hard) missing values */
replace Tot_Debt_Bus_`i'     =.a if classf_`i'<6
replace Tot_Bus_Debt_Owed_`i'=.a if classf_`i'<6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Debt_Bus_`i'   =. if f12b_bus_credcard_bal_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12b_bus_cred_line_bal_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_bank_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_nonbank_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_fam_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_govt_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_emp_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_other_ind_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_owner_amt_`i' ==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_bus_amt_`i' ==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_other_amt_`i'==.
```

```
replace Tot_Bus_Debt_Owed_`i'   =. if f12b_bus_credcard_bal_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12b_bus_cred_line_bal_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_bank_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_nonbank_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_fam_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_govt_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_emp_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_other_ind_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_owner_owed_`i' ==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_bus_owed_`i'  ==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_other_owed_`i'==.
}

forvalues i = 0/7 {
egen  Tot_Debt_`i'=rowtotal(Tot_Debt_Owner_Operators_`i' Tot_Debt_Bus_`i'),missing
egen  Tot_Debt_Owed_`i'=rowtotal(Tot_Debt_Owed_Owner_Operators_`i' Tot_Bus_Debt_Owed_`i'),missing
/* Recode legitimate (hard) missing values */
replace Tot_Debt_`i'      =.a if classf_`i'<6
replace Tot_Debt_Owed_`i' =.a if classf_`i'<6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Debt_`i'   =. if Tot_Debt_Owner_Operators_`i'==.
replace Tot_Debt_`i'   =. if Tot_Debt_Bus_`i'==.
replace Tot_Debt_Owed_`i'=. if Tot_Debt_Owed_Owner_Operators_`i'==.
replace Tot_Debt_Owed_`i'=. if Tot_Bus_Debt_Owed_`i'==.
}

forvalues i = 0/7 {
gen  Net_Profit_`i'=f24_profitloss_amt_`i'
}

/*****************************************************************/
/* Merge the file with the primary owner file "primary_owner.dta" */

merge 1:1 mprid using "xxx:\KFS_Manual_and_Data\primary_owner.dta"
drop _merge

*Replacing missing continuous values by the midpoints of the class intervals

forvalues i = 0/7 {
foreach ow in $owners_1_15  {
recode       total_hours_owner_`ow'_r_`i'      (0=0) (1=9.5)(2=27.5)(3=40.5)(4=50.5)(5=60.5) (6=70.5)
recode       age_owner_`ow'_r_`i'       (1=21)(2=29.5)(3=39.5)(4=49.5)(5=59.5) (6=69.5) (7=79.5)
```

```
    }
      }

forvalues i = 0/7 {
foreach ow in $owners_1_15  {
replace  g1b1_hours_owner_`ow'_`i'=total_hours_owner_`ow'_r_`i'   if  g1b1_hours_owner_`ow'_`i'==.
replace  g4_age_owner_`ow'_`i'=age_owner_`ow'_r_`i'   if  g4_age_owner_`ow'_`i'==.
}
}

drop tot_assets_* tot_liab_* tot_equity_nonowneroperators_* tot_equity_nonownerop_allyrs_* ///
tot_pers_debt_owed_resp_*  tot_pers_debt_resp_* tot_pers_debt_owed_othrownrs_* ///
tot_pers_debt_other_owners_* tot_bus_debt_owed_* tot_debt_bus_* tot_debt_owner_operators_* ///
tot_debt_owed_owner_operators_* tot_debt_* tot_debt_owed_* tot_equity_owner_operators_*  ///
tot_equity_owneroper_allyrs_* tot_debt_liab_equity_* tot_equity_* tot_equity_allyrs_*

drop *_r_*

* Make sure that the race do not change over time due to entry errors
forvalues i = 1/7 {
foreach ow in $owners_1_15  {

replace      g6_race_amind_owner_`ow'_`i'      =     g6_race_amind_owner_`ow'_0 if     g6_race_amind_owner_`ow'_`i'
      <.     &      g6_race_amind_owner_`ow'_0<.
replace      g6_race_asian_owner_`ow'_`i'      =     g6_race_asian_owner_`ow'_0 if     g6_race_asian_owner_`ow'_`i'
      <.     &      g6_race_asian_owner_`ow'_0<.
replace      g6_race_black_owner_`ow'_`i'      =     g6_race_black_owner_`ow'_0 if     g6_race_black_owner_`ow'_`i'
      <.     &      g6_race_black_owner_`ow'_0<.
replace      g6_race_nathaw_owner_`ow'_`i'     =     g6_race_nathaw_owner_`ow'_0       if
      g6_race_nathaw_owner_`ow'_`i'     <.     &     g6_race_nathaw_owner_`ow'_0<.
replace      g6_race_other_owner_`ow'_`i'      =     g6_race_other_owner_`ow'_0 if     g6_race_other_owner_`ow'_`i'
      <.     &      g6_race_other_owner_`ow'_0<.
replace      g6_race_white_owner_`ow'_`i'      =     g6_race_white_owner_`ow'_0 if     g6_race_white_owner_`ow'_`i'
      <.     &      g6_race_white_owner_`ow'_0<.
}
}

* Primary Owner(PO) Characteristics at the Baseline: PO by Robb's Stata code

forvalues i = 0/0 {
gen  PO_emp_`i'               =.
gen  PO_hours_`i'             =.
```

```
gen  PO_work_exp_`i'              =.
gen  PO_oth_bus_owner_`i'      =.
gen  PO_bus_same_ind_`i'       =.
gen  PO_age_owner_`i'             =.
gen  PO_hisp_origin_`i'       =.
gen  PO_race_group_`i'        =.
gen  PO_race_amind_owner_`i'   =.
gen  PO_race_asian_owner_`i'   =.
gen  PO_race_black_owner_`i'   =.
gen  PO_race_nathaw_owner_`i'  =.
gen  PO_race_other_owner_`i'   =.
gen  PO_race_white_owner_`i'   =.
gen  PO_native_born_`i'       =.
gen  PO_us_cit_`i'             =.
gen  PO_education_`i'             =.
gen  PO_gender_`i'             =.
}

forvalues i = 0/0 {
forvalues po = 1/6 {
replace PO_emp_`i'          =   g1a_emp_owner_0`po'_`i'             if  primary_owner==`po'
replace PO_hours_`i'           =  g1b1_hours_owner_0`po'_`i'                 if  primary_owner==`po'
replace PO_work_exp_`i'    =      g2_work_exp_owner_0`po'_`i'              if  primary_owner==`po'
replace PO_oth_bus_owner_`i'=     g3a_oth_bus_owner_0`po'_`i'              if  primary_owner==`po'
replace PO_bus_same_ind_`i' =     g3b_bus_same_ind_owner_0`po'_`i' if  primary_owner==`po'
replace PO_age_owner_`i'   =      g4_age_owner_0`po'_`i'                     if  primary_owner==`po'
replace PO_hisp_origin_`i'        =      g5_hisp_origin_owner_0`po'_`i'        if  primary_owner==`po'
replace PO_race_group_`i'   =     g6b_race_group_0`po'_`i'              if  primary_owner==`po'
replace PO_race_amind_owner_`i' = g6_race_amind_owner_0`po'_`i'     if  primary_owner==`po'
replace PO_race_asian_owner_`i' = g6_race_asian_owner_0`po'_`i'     if  primary_owner==`po'
replace PO_race_black_owner_`i' = g6_race_black_owner_0`po'_`i'     if  primary_owner==`po'
replace PO_race_nathaw_owner_`i'= g6_race_nathaw_owner_0`po'_`i'    if  primary_owner==`po'
replace PO_race_other_owner_`i' = g6_race_other_owner_0`po'_`i'     if  primary_owner==`po'
replace PO_race_white_owner_`i' = g6_race_white_owner_0`po'_`i'     if  primary_owner==`po'
replace PO_native_born_`i'        =      g7_native_born_owner_0`po'_`i'        if  primary_owner==`po'
replace PO_us_cit_`i'          =  g8_us_cit_owner_0`po'_`i'             if  primary_owner==`po'
replace PO_education_`i'   =      g9_education_owner_0`po'_`i'              if  primary_owner==`po'
replace PO_gender_`i'          =  g10_gender_owner_0`po'_`i'            if  primary_owner==`po'
}
}

forvalues i = 1/7 {
```

```
gen  PO_emp_`i'          = PO_emp_0
gen  PO_hours_`i'        = PO_hours_0
gen  PO_work_exp_`i'        = PO_work_exp_0
gen  PO_oth_bus_owner_`i'= PO_oth_bus_owner_0
gen  PO_bus_same_ind_`i' = PO_bus_same_ind_0
replace PO_bus_same_ind_`i' =.a if PO_oth_bus_owner_0==0
gen  PO_age_owner_`i'       = PO_age_owner_0
gen  PO_hisp_origin_`i'  = PO_hisp_origin_0
gen  PO_race_group_`i'   = PO_race_group_0
gen  PO_race_amind_owner_`i' =PO_race_amind_owner_0
gen  PO_race_asian_owner_`i' =PO_race_asian_owner_0
gen  PO_race_black_owner_`i' =PO_race_black_owner_0
gen  PO_race_nathaw_owner_`i'=PO_race_nathaw_owner_0
gen  PO_race_other_owner_`i' =PO_race_other_owner_0
gen  PO_race_white_owner_`i' =PO_race_white_owner_0
gen  PO_native_born_`i'  = PO_native_born_0
gen  PO_us_cit_`i'          = PO_us_cit_0
gen  PO_education_`i'       = PO_education_0
gen  PO_gender_`i'       = PO_gender_0
/* Recode legitimate (hard) missing values */
replace  PO_emp_`i'                = .a if    classf_`i'  <6
replace  PO_hours_`i'              = .a if    classf_`i'  <6
replace  PO_work_exp_`i'        = .a if classf_`i'  <6
replace  PO_oth_bus_owner_`i'     = .a if     classf_`i'  <6
replace  PO_bus_same_ind_`i'      = .a if     classf_`i'  <6
replace  PO_age_owner_`i'       = .a if classf_`i'  <6
replace  PO_hisp_origin_`i'       = .a if     classf_`i'  <6
replace  PO_race_group_`i'        = .a if     classf_`i'  <6
replace  PO_race_amind_owner_`i' = .a if      classf_`i'  <6
replace  PO_race_asian_owner_`i' = .a if      classf_`i'  <6
replace  PO_race_black_owner_`i' = .a if      classf_`i'  <6
replace  PO_race_nathaw_owner_`i'= .a if      classf_`i'  <6
replace  PO_race_other_owner_`i' = .a if      classf_`i'  <6
replace  PO_race_white_owner_`i' = .a if      classf_`i'  <6
replace  PO_native_born_`i'       = .a if     classf_`i'  <6
replace  PO_us_cit_`i'             = .a if    classf_`i'  <6
replace  PO_education_`i'       = .a if classf_`i'  <6
replace  PO_gender_`i'             = .a if    classf_`i'  <6

}

* Active-Owner-Operators Characteristics (OO)
```

```
forvalues i = 0/7 {
egen   OO_emp_owner_`i'          =      rowmean(g1a_emp_owner_*_`i')
egen   OO_hours_owner_`i'        = rowmean(g1b1_hours_owner_*_`i')
egen   OO_work_exp_owner_`i'     =      rowmean(g2_work_exp_owner_*_`i'   )
egen   OO_oth_bus_owner_`i'      =      rowmean(g3a_oth_bus_owner_*_`i'   )
egen   OO_bus_same_ind_owner_`i' =      rowmean(g3b_bus_same_ind_owner_*_`i')
egen   OO_age_owner_`i'          =      rowmean(g4_age_owner_*_`i')
egen   OO_hisp_origin_owner_`i'  =      rowmean(g5_hisp_origin_owner_*_`i')
egen   OO_race_amind_owner_`i'   =      rowmean(g6_race_amind_owner_*_`i')
egen   OO_race_asian_owner_`i'   =      rowmean(g6_race_asian_owner_*_`i')
egen   OO_race_black_owner_`i'   =      rowmean(g6_race_black_owner_*_`i')
egen   OO_race_nathaw_owner_`i'  =      rowmean(g6_race_nathaw_owner_*_`i')
egen   OO_race_other_owner_`i'   =      rowmean(g6_race_other_owner_*_`i')
egen   OO_race_white_owner_`i'   =      rowmean(g6_race_white_owner_*_`i')
egen   OO_native_born_owner_`i'  =      rowmean(g7_native_born_owner_*_`i'      )
egen   OO_us_cit_owner_`i'       =      rowmean(g8_us_cit_owner_*_`i'    )
egen   OO_education_owner_`i'    =      rowmean(g9_education_owner_*_`i'  )
egen   md_education_owner_`i'    =      rowmedian(g9_education_owner_*_`i')
gen    OO_D_education_owner_`i'  =(md_education_owner_`i'    >6.99)        if     md_education_owner_`i'    <11

egen   OO_gender_owner_`i'       =      rowmean(g10_gender_owner_*_`i'   )
}

/* Recode to soft missing value if any of the total's component is soft missing*/
forvalues i = 0/7 {
foreach ow in $owners_1_15  {
replace OO_emp_owner_`i'              =. if g1a_emp_owner_`ow'_`i'==.
replace OO_hours_owner_`i'          =. if g1b1_hours_owner_`ow'_`i'==.
replace OO_work_exp_owner_`i'     =. if g2_work_exp_owner_`ow'_`i'==.
replace OO_oth_bus_owner_`i'      =. if g3a_oth_bus_owner_`ow'_`i'==.
replace OO_bus_same_ind_owner_`i' =. if g3b_bus_same_ind_owner_`ow'_`i'==.
replace OO_age_owner_`i'              =. if g4_age_owner_`ow'_`i'==.
replace OO_hisp_origin_owner_`i'  =. if g5_hisp_origin_owner_`ow'_`i'==.
replace OO_race_amind_owner_`i'   =. if g6_race_amind_owner_`ow'_`i'==.
replace OO_race_asian_owner_`i'   =. if g6_race_asian_owner_`ow'_`i'==.
replace OO_race_black_owner_`i'   =. if g6_race_black_owner_`ow'_`i'==.
replace OO_race_nathaw_owner_`i'  =. if g6_race_nathaw_owner_`ow'_`i'==.
replace OO_race_other_owner_`i'   =. if g6_race_other_owner_`ow'_`i'==.
replace OO_race_white_owner_`i'   =. if g6_race_white_owner_`ow'_`i'==.
replace OO_native_born_owner_`i'  =. if g7_native_born_owner_`ow'_`i'==.
replace OO_us_cit_owner_`i'       =. if g8_us_cit_owner_`ow'_`i'==.
```

```
replace OO_education_owner_`i'      =. if g9_education_owner_`ow'_`i'==.
replace md_education_owner_`i'      =. if g9_education_owner_`ow'_`i'==.
replace OO_D_education_owner_`i'  =. if g9_education_owner_`ow'_`i'==.
replace OO_gender_owner_`i'       =. if g10_gender_owner_`ow'_`i'==.
}
}

/* Recode legitimate (hard) missing values */
forvalues i = 0/7 {
replace     OO_bus_same_ind_owner_`i' = .a if   OO_oth_bus_owner_`i'==0
replace      OO_emp_owner_`i'          = .a if classf_`i'  <6
replace      OO_hours_owner_`i'        = .a if classf_`i'  <6
replace      OO_work_exp_owner_`i'     = .a if classf_`i'  <6
replace      OO_oth_bus_owner_`i'      = .a if classf_`i'  <6
replace      OO_bus_same_ind_owner_`i' = .a if classf_`i'  <6
replace      OO_age_owner_`i'          = .a if classf_`i'  <6
replace      OO_hisp_origin_owner_`i'  = .a if classf_`i'  <6
replace      OO_race_amind_owner_`i'   = .a if classf_`i'  <6
replace      OO_race_asian_owner_`i'   = .a if classf_`i'  <6
replace      OO_race_black_owner_`i'   = .a if classf_`i'  <6
replace      OO_race_nathaw_owner_`i'  = .a if classf_`i'  <6
replace      OO_race_other_owner_`i'   = .a if classf_`i'  <6
replace      OO_race_white_owner_`i'   = .a if classf_`i'  <6
replace      OO_native_born_owner_`i'  = .a if classf_`i'  <6
replace      OO_us_cit_owner_`i'       = .a if classf_`i'  <6
replace      OO_education_owner_`i'    = .a if classf_`i'  <6
replace      md_education_owner_`i'    = .a if classf_`i'  <6
replace      OO_D_education_owner_`i'  = .a if classf_`i'  <6

replace      OO_gender_owner_`i'       = .a if classf_`i'  <6
}

/*****************************************************************************/
*Diversity / Similarity index
forvalues i = 0/7 {
gen xr1_`i'= OO_race_amind_owner_`i'    *      OO_race_amind_owner_`i'
gen xr2_`i'= OO_race_asian_owner_`i'    *      OO_race_asian_owner_`i'
gen xr3_`i'= OO_race_black_owner_`i'    *      OO_race_black_owner_`i'
gen xr4_`i'= OO_race_nathaw_owner_`i'*  OO_race_nathaw_owner_`i'
gen xr5_`i'= OO_race_other_owner_`i' *  OO_race_other_owner_`i'
gen xr6_`i'= OO_race_white_owner_`i'    *      OO_race_white_owner_`i'
* Race_similarity
```

```
egen Race_similarity_`i'=rowtotal(xr1_`i' xr2_`i' xr3_`i' xr4_`i' xr5_`i' xr6_`i'),missing
drop xr*_`i'
* Race_ diversity
gen Race_diversity_`i'=1-Race_similarity_`i'
}

forvalues i = 0/7 {
foreach ow in $owners_1_15  {
/* Recode to soft missing value if any of the var's component is soft missing*/
replace Race_similarity_`i'   =. if g6_race_amind_owner_`ow'_`i'==.
replace Race_similarity_`i'   =. if g6_race_asian_owner_`ow'_`i'==.
replace Race_similarity_`i'   =. if g6_race_black_owner_`ow'_`i'==.
replace Race_similarity_`i'   =. if g6_race_nathaw_owner_`ow'_`i'==.
replace Race_similarity_`i'   =. if g6_race_other_owner_`ow'_`i'==.
replace Race_similarity_`i'   =. if g6_race_white_owner_`ow'_`i'==.

}
}
forvalues i = 0/7 {
/* Recode legitimate (hard) missing values */
replace Race_similarity_`i'   =.a if classf_`i'<6
replace Race_diversity_`i'   =.a if Race_similarity_`i'==.a
replace Race_diversity_`i'   =.  if Race_similarity_`i'==.
}

forvalues i = 0/7 {
gen fmal_`i'=1-OO_gender_owner_`i'
gen xr1_`i'= OO_gender_owner_`i' *      OO_gender_owner_`i'
gen xr2_`i'= fmal_`i'      *      fmal_`i'

* Gender_similarity
egen Gender_similarity_`i'=rowtotal(xr1_`i' xr2_`i'  ),missing
drop xr*_`i' fmal_`i'
* Gender_diversity
gen Gender_diversity_`i'=1-Gender_similarity_`i'
}

forvalues i = 0/7 {
foreach ow in $owners_1_15  {
/* Recode to soft missing value if any of the var's component is soft missing*/
replace  Gender_similarity_`i'      =. if g10_gender_owner_`ow'_`i'==.
}
```

```
}

forvalues i = 0/7 {
/* Recode legitimate (hard) missing values */
replace Gender_similarity_`i'   =.a if classf_`i'<6
replace Gender_diversity_`i'   =.a if Gender_similarity_`i'==.a
replace Gender_diversity_`i'   =.  if Gender_similarity_`i'==.
}

forvalues i = 0/7 {
mean Race_similarity_`i' Gender_similarity_`i' if c4_numowners_confirm_`i'>1 & c4_numowners_confirm_`i'<.
}

*Business level Characteristics
forvalues i = 0/7 {
*Home Based Dummy
recode c8_primary_loc_`i' (1=1 "Home Based") (nonmiss=0 "Non Home Based" ) ,into (Home_Based_`i')
*Sole_Proprietorship Dummy
recode    c1z2_legal_status_`i'    (1=1    "Sole_Proprietorship")    (nonmiss=0    "Limited    Liability"    )    ,into
(Sole_Proprietorship_`i')
rename d2_comp_advantage_`i' Comp_advantage_`i'
egen Have_IP_`i'=anymatch(d3_a_have_patent_`i' d3_b_have_copyright_`i' d3_c_have_trademark_`i'), values(1)
rename c5_num_employees_`i'     Full_Part_Time_Employees_`i'
rename c6_num_ft_employees_`i'  Full_Time_Employees_`i'
rename c7_num_pt_employees_`i'  Part_Time_Employees_`i'
egen   Employee_Owner_`i'          =    rowtotal(g1a_emp_owner_*_`i')
egen   Total_Employees_`i'=rowtotal(Employee_Owner_`i' Full_Part_Time_Employees_`i')
}

forvalues i = 0/7 {
foreach ow in $owners_1_15  {
/* Recode to soft missing value if any of the var's component is soft missing*/
replace Have_IP_`i'=. if d3_a_have_patent_`i'==. | d3_b_have_copyright_`i'==. | d3_c_have_trademark_`i'==.
replace Employee_Owner_`i'             =. if g1a_emp_owner_`ow'_`i'==.
replace Total_Employees_`i'            =. if g1a_emp_owner_`ow'_`i'==.
replace Total_Employees_`i'            =. if  Full_Part_Time_Employees_`i'==.
/* Recode legitimate (hard) missing values */
replace Have_IP_`i'        =.a if classf_`i'<6
replace Employee_Owner_`i' =.a if classf_`i'<6
replace Total_Employees_`i'=.a if classf_`i'<6
}
}
```

```
saveold KFS8_LI_CS_w1,replace

forvalues i = 0/7 {

  set logtype text , permanently

  log using KFS8_LI_CS_`i'.csv, replace
    FR_Sum_CS *_`i' [pw=cswgt_final_`i']
  log close
}
```

3.2.2.2.5. Stata Code: Longitudinal in Wide Format

The following Stata code will create the new variables using the KFS wide format file "KFS8_LI." The code will save the new longitudinal (n=3140) file under the name "KFS8_L7_w1.dta."

```
clear
clear mata
clear matrix
capture log close
set more off
set maxvar 32767
program drop _all
adopath + " xxx:\KFS_Manual_and_Data\Farhat_Robb_Commands\"
cd " xxx:\KFS_Manual_and_Data "
use KFS8_LI,clear

keep if wgt_7_long>0

set more off
* Data Management Var
rename one_ower_00 one_owner_0
rename one_ower_01 one_owner_1
rename one_ower_02 one_owner_2
rename one_ower_03 one_owner_3
rename one_ower_04 one_owner_4
rename one_ower_05 one_owner_5
rename one_ower_06 one_owner_6
rename one_ower_07 one_owner_7


*Replacing missing continuous values by the midpoints of the class intervals

foreach totfin in tot_*_r_* {
recode `totfin'     (0=0)(1=250)(2=750)(3=2000)(4=4000)(5=7500) (6=17500) (7=62500) (8=550000) (9=1000000)
    }

global owners_1_15 "01 02 03 04 05 06 07 08 09 10 11 12 13 14 15"

forvalues i = 0/7 {
foreach ow in $owners_1_15  {
```

```
replace  f2_owner_amt_eq_invest_`ow'_`i'=tot_equity_owner_`ow'_r_`i'   if  f2_owner_amt_eq_invest_`ow'_`i'==.
replace               f2_ownr_amt_eqinvest_allyrs_`ow'_`i'=tot_equity_allyrs_owner_`ow'_r_`i'                        if
f2_ownr_amt_eqinvest_allyrs_`ow'_`i'==.
}
}

forvalues i = 0/7 {
replace  f4_eq_amt_angels_`i'=tot_equity_angels_r_`i'   if  f4_eq_amt_angels_`i'==.
replace  f4_eq_amt_angels_allyrs_`i'=tot_equity_angels_allyrs_r_`i'   if  f4_eq_amt_angels_allyrs_`i'==.
replace  f4_eq_amt_companies_`i'=tot_equity_companies_r_`i'   if  f4_eq_amt_companies_`i'==.
replace  f4_eq_amt_companies_allyrs_`i'=tot_equity_companies_allyrs_r_`i'   if  f4_eq_amt_companies_allyrs_`i'==.
replace  f4_eq_amt_govt_`i'=tot_equity_govt_r_`i'   if  f4_eq_amt_govt_`i'==.
replace  f4_eq_amt_govt_allyrs_`i'=tot_equity_govt_allyrs_r_`i'   if  f4_eq_amt_govt_allyrs_`i'==.
replace  f4_eq_amt_other_`i'=tot_equity_other_r_`i'   if  f4_eq_amt_other_`i'==.
replace  f4_eq_amt_other_allyrs_`i'=tot_equity_other_allyrs_r_`i'   if  f4_eq_amt_other_allyrs_`i'==.
replace  f4_eq_amt_parents_`i'=tot_equity_parents_r_`i'   if  f4_eq_amt_parents_`i'==.
replace  f4_eq_amt_parents_allyrs_`i'=tot_equity_parents_allyrs_r_`i'   if  f4_eq_amt_parents_allyrs_`i'==.
replace  f4_eq_amt_spouse_`i'=tot_equity_spouse_r_`i'   if  f4_eq_amt_spouse_`i'==.
replace  f4_eq_amt_spouse_allyrs_`i'=tot_equity_spouse_allyrs_r_`i'   if  f4_eq_amt_spouse_allyrs_`i'==.
replace  f4_eq_amt_vent_cap_`i'=tot_equity_vent_cap_r_`i'   if  f4_eq_amt_vent_cap_`i'==.
replace  f4_eq_amt_vent_cap_allyrs_`i'=tot_equity_vent_cap_allyrs_r_`i'   if  f4_eq_amt_vent_cap_allyrs_`i'==.
replace  f6b_personal_use_amt_`i'=tot_personal_use_r_`i'   if  f6b_personal_use_amt_`i'==.
replace  f8a_bus_credcard_line_`i'=tot_bus_credcard_line_resp_r_`i'   if  f8a_bus_credcard_line_`i'==.
replace  f8b_bus_credcard_bal_`i'=tot_bus_credcard_bal_resp_r_`i'   if  f8b_bus_credcard_bal_`i'==.
replace  f8a_pers_credcard_line_`i'=tot_pers_credcard_line_resp_r_`i'   if  f8a_pers_credcard_line_`i'==.
replace  f8b_pers_credcard_bal_`i'=tot_pers_credcard_bal_resp_r_`i'   if  f8b_pers_credcard_bal_`i'==.
replace  f8c_pers_loan_bank_amt_`i'=tot_pers_loan_bank_resp_r_`i'   if  f8c_pers_loan_bank_amt_`i'==.
replace  f8d_pers_loan_bank_owed_`i'=tot_pers_loan_bank_owed_resp_r_`i'   if  f8d_pers_loan_bank_owed_`i'==.
replace  f8c_pers_loan_fam_amt_`i'=tot_pers_loan_fam_resp_r_`i'   if  f8c_pers_loan_fam_amt_`i'==.
replace  f8d_pers_loan_fam_owed_`i'=tot_pers_loan_fam_owed_resp_r_`i'   if  f8d_pers_loan_fam_owed_`i'==.
replace  f8c_pers_loan_other_amt_`i'=tot_pers_loan_other_resp_r_`i'   if  f8c_pers_loan_other_amt_`i'==.
replace  f8d_pers_loan_other_owed_`i'=tot_persloan_other_owed_resp_r_`i'   if  f8d_pers_loan_other_owed_`i'==.
replace  f8c_pers_other_amt_`i'=tot_pers_other_resp_r_`i'   if  f8c_pers_other_amt_`i'==.
replace  f8d_pers_other_owed_`i'=tot_pers_other_owed_resp_r_`i'   if  f8d_pers_other_owed_`i'==.
replace  f10a_bus_credcard_line_`i'=tot_bus_credcard_line_others_r_`i'   if  f10a_bus_credcard_line_`i'==.
replace  f10b_bus_credcard_bal_`i'=tot_bus_credcard_bal_others_r_`i'   if  f10b_bus_credcard_bal_`i'==.
replace  f10a_pers_credcard_line_`i'=tot_pers_credcard_line_othrs_r_`i'   if  f10a_pers_credcard_line_`i'==.
replace  f10b_pers_credcard_bal_`i'=tot_pers_credcard_bal_others_r_`i'   if  f10b_pers_credcard_bal_`i'==.
replace  f10c_pers_loan_bank_amt_`i'=tot_pers_loan_bank_others_r_`i'   if  f10c_pers_loan_bank_amt_`i'==.
replace  f10d_pers_loan_bank_owed_`i'=tot_persloan_bank_owed_othrs_r_`i'   if  f10d_pers_loan_bank_owed_`i'==.
replace  f10c_pers_loan_fam_amt_`i'=tot_persloan_fam_othrowners_r_`i'   if  f10c_pers_loan_fam_amt_`i'==.
replace  f10d_pers_loan_fam_owed_`i'=tot_persloan_fam_owed_othrs_r_`i'   if  f10d_pers_loan_fam_owed_`i'==.
```

```
replace  f10c_pers_loan_other_amt_`i'=tot_pers_loan_other_owners_r_`i'   if  f10c_pers_loan_other_amt_`i'==.
replace  f10d_pers_loan_other_owed_`i'=tot_persloan_othr_owed_othrs_r_`i'   if  f10d_pers_loan_other_owed_`i'==.
replace  f10c_pers_other_amt_`i'=tot_pers_other_other_owners_r_`i'   if  f10c_pers_other_amt_`i'==.
replace  f10d_pers_other_owed_`i'=tot_pers_other_owed_others_r_`i'   if  f10d_pers_other_owed_`i'==.
replace  f12a_bus_cred_line_`i'=tot_cred_line_bus_line_r_`i'   if  f12a_bus_cred_line_`i'==.
replace  f12b_bus_cred_line_bal_`i'=tot_cred_line_bus_bal_r_`i'   if  f12b_bus_cred_line_bal_`i'==.
replace  f12a_bus_credcard_line_`i'=tot_credcard_line_bus_r_`i'   if  f12a_bus_credcard_line_`i'==.
replace  f12b_bus_credcard_bal_`i'=tot_credcard_bal_bus_r_`i'   if  f12b_bus_credcard_bal_`i'==.
replace  f12c_bus_loans_bank_amt_`i'=tot_loan_bank_bus_r_`i'   if  f12c_bus_loans_bank_amt_`i'==.
replace  f12d_bus_loans_bank_owed_`i'=tot_bus_loans_bank_owed_r_`i'   if  f12d_bus_loans_bank_owed_`i'==.
replace  f12c_bus_loans_nonbank_amt_`i'=tot_loan_nonbank_bus_r_`i'   if  f12c_bus_loans_nonbank_amt_`i'==.
replace  f12d_bus_loans_nonbank_owed_`i'=tot_bus_loans_nonbank_owed_r_`i'   if  f12d_bus_loans_nonbank_owed_`i'==.
replace  f12c_bus_loans_emp_amt_`i'=tot_loan_emp_bus_r_`i'   if  f12c_bus_loans_emp_amt_`i'==.
replace  f12d_bus_loans_emp_owed_`i'=tot_bus_loans_emp_owed_r_`i'   if  f12d_bus_loans_emp_owed_`i'==.
replace  f12c_bus_loans_fam_amt_`i'=tot_loan_fam_bus_r_`i'   if  f12c_bus_loans_fam_amt_`i'==.
replace  f12d_bus_loans_fam_owed_`i'=tot_bus_loans_fam_owed_r_`i'   if  f12d_bus_loans_fam_owed_`i'==.
replace  f12c_bus_loans_govt_amt_`i'=tot_loan_govt_bus_r_`i'   if  f12c_bus_loans_govt_amt_`i'==.
replace  f12d_bus_loans_govt_owed_`i'=tot_bus_loans_govt_owed_r_`i'   if  f12d_bus_loans_govt_owed_`i'==.
replace  f12c_bus_loans_other_ind_amt_`i'=tot_loan_other_ind_r_`i'   if  f12c_bus_loans_other_ind_amt_`i'==.
replace  f12d_bus_loans_other_ind_owed_`i'=tot_bus_loans_otherind_owed_r_`i'   if  f12d_bus_loans_other_ind_owed_`i'==.
replace  f12c_bus_loans_owner_amt_`i'=tot_loan_owner_bus_r_`i'   if  f12c_bus_loans_owner_amt_`i'==.
replace  f12d_bus_loans_owner_owed_`i'=tot_bus_loans_owner_owed_r_`i'   if  f12d_bus_loans_owner_owed_`i'==.
replace  f12c_bus_loans_bus_amt_`i'=tot_loan_other_bus_r_`i'   if  f12c_bus_loans_bus_amt_`i'==.
replace  f12d_bus_loans_bus_owed_`i'=tot_bus_loans_otherbus_owed_r_`i'   if  f12d_bus_loans_bus_owed_`i'==.
replace  f12c_bus_other_amt_`i'=tot_bus_debt_other_r_`i'   if  f12c_bus_other_amt_`i'==.
replace  f12d_bus_other_owed_`i'=tot_bus_loans_other_owed_r_`i'   if  f12d_bus_other_owed_`i'==.
replace  f14a_trade_fin_amt_`i'=tot_trade_finan_r_`i'   if  f14a_trade_fin_amt_`i'==.
replace  f16a_rev_amt_`i'=tot_revenue_r_`i'   if  f16a_rev_amt_`i'==.
replace  f17a_total_exp_amt_`i'=tot_expenses_r_`i'   if  f17a_total_exp_amt_`i'==.
replace  f18a_wage_exp_amt_`i'=tot_wages_r_`i'   if  f18a_wage_exp_amt_`i'==.
replace  f19a_res_dev_amt_`i'=tot_res_dev_r_`i'   if  f19a_res_dev_amt_`i'==.
replace  f19c_a_design_amt_`i'=tot_intangassets_design_r_`i'   if  f19c_a_design_amt_`i'==.
replace  f19c_b_investments_amt_`i'=tot_intangassets_invest_r_`i'   if  f19c_b_investments_amt_`i'==.
replace  f19c_c_brand_dev_amt_`i'=tot_intangassets_branddev_r_`i'   if  f19c_c_brand_dev_amt_`i'==.
replace  f19c_d_org_dev_amt_`i'=tot_intangassets_orgdev_r_`i'   if  f19c_d_org_dev_amt_`i'==.
replace  f19c_e_worker_training_amt_`i'=tot_intangassets_wkrtrng_r_`i'   if  f19c_e_worker_training_amt_`i'==.
replace  f19c_f_other_amt_`i'=tot_intangassets_other_r_`i'   if  f19c_f_other_amt_`i'==.
replace  f19c_intangassets_amt_`i'=tot_intang_assets_r_`i'   if  f19c_intangassets_amt_`i'==.
replace  f24_profit_amt_`i'=tot_profit_r_`i'   if  f24_profit_amt_`i'==.
replace  f26_loss_amt_`i'=tot_loss_r_`i'   if  f26_loss_amt_`i'==.
replace  f29_assetval_cash_`i'=tot_asset_cash_r_`i'   if  f29_assetval_cash_`i'==.
replace  f29_assetval_acctrec_`i'=tot_asset_acct_rec_r_`i'   if  f29_assetval_acctrec_`i'==.
```

```
replace  f29_assetval_inv_`i'=tot_asset_inv_r_`i'   if  f29_assetval_inv_`i'==.
replace  f29_assetval_equip_`i'=tot_asset_equip_r_`i'   if  f29_assetval_equip_`i'==.
replace  f29_assetval_landbuild_`i'=tot_asset_landbuild_r_`i'   if  f29_assetval_landbuild_`i'==.
replace  f29_assetval_veh_`i'=tot_asset_veh_r_`i'   if  f29_assetval_veh_`i'==.
replace  f29_assetval_othbusprop_`i'=tot_asset_other_bus_prop_r_`i'   if  f29_assetval_othbusprop_`i'==.
replace  f29_assetval_other_`i'=tot_asset_other_r_`i'   if  f29_assetval_other_`i'==.
replace  f31_value_acctpay_`i'=tot_liab_acct_pay_r_`i'   if  f31_value_acctpay_`i'==.
replace  f31_value_pension_`i'=tot_liab_pension_r_`i'   if  f31_value_pension_`i'==.
replace  f31_value_other_`i'=tot_liab_other_r_`i'   if  f31_value_other_`i'==.
}

forvalues i = 0/7 {
egen   Tot_Equity_Owner_Operators_`i'=  rowtotal(     f2_owner_amt_eq_invest_*_`i'      )      , missing
egen   Tot_Equity_OwnerOper_AllYrs_`i'=rowtotal(      f2_ownr_amt_eqinvest_allyrs_*_`i' ), missing
}


forvalues i = 0/7 {
foreach ow in $owners_1_15  {
/* Recode legitimate (hard) missing values */
replace Tot_Equity_Owner_Operators_`i' =.a if classf_`i'<6
replace Tot_Equity_OwnerOper_AllYrs_`i'=.a if classf_`i'<6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Equity_Owner_Operators_`i' =.   if  f2_owner_amt_eq_invest_`ow'_`i'==.
replace Tot_Equity_OwnerOper_AllYrs_`i'=.   if  f2_ownr_amt_eqinvest_allyrs_`ow'_`i'==.
}
}

global List1 "spouse parents angels companies govt vent_cap other"

forvalues i = 0/7 {
egen  Tot_Equity_NonOwnerOperators_`i'=rowtotal(f4_eq_amt_angels_`i' f4_eq_amt_companies_`i' f4_eq_amt_govt_`i' ///
f4_eq_amt_other_`i' f4_eq_amt_parents_`i' f4_eq_amt_spouse_`i'  f4_eq_amt_vent_cap_`i')      , missing
egen  Tot_Equity_NonOwnerOp_AllYrs_`i'=rowtotal(f4_eq_amt_*_allyrs_`i')   , missing
/* Recode legitimate (hard) missing values */
replace Tot_Equity_NonOwnerOperators_`i' =.a if classf_`i'<6
replace Tot_Equity_NonOwnerOperators_`i' =.a if  c1z2_legal_status_`i'==1
replace Tot_Equity_NonOwnerOperators_`i' =.a if  one_owner_`i'==1
replace Tot_Equity_NonOwnerOp_AllYrs_`i' =.a if classf_`i'<6
replace Tot_Equity_NonOwnerOp_AllYrs_`i' =.a if  c1z2_legal_status_`i'==1
replace Tot_Equity_NonOwnerOp_AllYrs_`i' =.a if  one_owner_`i'==1
}
```

```
forvalues i = 0/7 {
foreach name in $List1  {
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Equity_NonOwnerOperators_`i' =. if f4_eq_amt_`name'_`i'==.
replace Tot_Equity_NonOwnerOp_AllYrs_`i' =. if f4_eq_amt_`name'_allyrs_`i'==.
}
}

forvalues i = 0/7 {
egen  Tot_Equity_`i'       =rowtotal(Tot_Equity_Owner_Operators_`i' Tot_Equity_NonOwnerOperators_`i'), missing
egen  Tot_Equity_AllYrs_`i'=rowtotal(Tot_Equity_OwnerOper_AllYrs_`i' Tot_Equity_NonOwnerOp_AllYrs_`i'), missing
/* Recode legitimate (hard) missing values */
replace Tot_Equity_`i'       =.a if classf_`i'<6
replace Tot_Equity_AllYrs_`i'=.a if classf_`i'<6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Equity_`i'       =. if Tot_Equity_Owner_Operators_`i'==.
replace Tot_Equity_`i'       =. if Tot_Equity_NonOwnerOperators_`i'==.
replace Tot_Equity_AllYrs_`i'=. if Tot_Equity_NonOwnerOperators_`i'==.
replace Tot_Equity_AllYrs_`i'=. if Tot_Equity_NonOwnerOp_AllYrs_`i'==.
}

forvalues i = 0/7 {
egen  Tot_Assets_`i'=rowtotal(f29_assetval_*_`i')    , missing
egen  Tot_Liab_`i'=rowtotal(f31_value_*_`i'), missing
/* Recode legitimate (hard) missing values */
replace Tot_Assets_`i' =.a if classf_`i'<6
replace Tot_Liab_`i'   =.a if classf_`i'<6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Assets_`i' =. if f29_assetval_acctrec_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_cash_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_equip_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_inv_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_landbuild_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_othbusprop_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_other_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_veh_`i'==.
replace Tot_Liab_`i' =. if  f31_value_acctpay_`i'==.
replace Tot_Liab_`i' =. if  f31_value_other_`i'==.
replace Tot_Liab_`i' =. if  f31_value_pension_`i'==.
}
```

```
forvalues i = 0/7 {
egen    Tot_Pers_Debt_Resp_`i'=rowtotal(f8b_pers_credcard_bal_`i'  f8b_bus_credcard_bal_`i'  f8c_pers_loan_bank_amt_`i'
f8c_pers_loan_fam_amt_`i' f8c_pers_loan_other_amt_`i' f8c_pers_other_amt_`i'),missing
egen                  Tot_Pers_Debt_Owed_Resp_`i'=rowtotal(f8b_pers_credcard_bal_`i'          f8b_bus_credcard_bal_`i'
f8d_pers_loan_bank_owed_`i' f8d_pers_loan_fam_owed_`i' f8d_pers_loan_other_owed_`i' f8d_pers_other_owed_`i'),missing
/* Recode legitimate (hard) missing values */
replace Tot_Pers_Debt_Resp_`i'      =.a if classf_`i'<6
replace Tot_Pers_Debt_Owed_Resp_`i' =.a if classf_`i'<6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Pers_Debt_Resp_`i' =. if f8b_pers_credcard_bal_`i'==.
replace Tot_Pers_Debt_Resp_`i' =. if f8b_bus_credcard_bal_`i'==.
replace Tot_Pers_Debt_Resp_`i' =. if f8c_pers_loan_bank_amt_`i'==.
replace Tot_Pers_Debt_Resp_`i' =. if f8c_pers_loan_fam_amt_`i'==.
replace Tot_Pers_Debt_Resp_`i' =. if f8c_pers_loan_other_amt_`i'==.
replace Tot_Pers_Debt_Resp_`i' =. if f8c_pers_other_amt_`i'==.
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8b_pers_credcard_bal_`i'==.
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8b_bus_credcard_bal_`i'==.
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8d_pers_loan_bank_owed_`i'==.
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8d_pers_loan_fam_owed_`i'==.
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8d_pers_loan_other_owed_`i'==.
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8d_pers_other_owed_`i'==.
}

forvalues i = 0/7 {
egen               Tot_Pers_Debt_Other_Owners_`i'=rowtotal(f10b_pers_credcard_bal_`i'        f10c_pers_loan_bank_amt_`i'
f10b_bus_credcard_bal_`i' f10c_pers_loan_fam_amt_`i' f10c_pers_loan_other_amt_`i'  f10c_pers_other_amt_`i'),missing
egen               Tot_Pers_Debt_Owed_OthrOwnrs_`i'=rowtotal(f10b_pers_credcard_bal_`i'        f10b_bus_credcard_bal_`i'
f10d_pers_loan_bank_owed_`i'                  f10d_pers_loan_fam_owed_`i'                  f10d_pers_loan_other_owed_`i'
f10d_pers_other_owed_`i'),missing
/* Recode legitimate (hard) missing values */
replace Tot_Pers_Debt_Other_Owners_`i'  =.a if classf_`i'<6
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i'=.a if classf_`i'<6
replace Tot_Pers_Debt_Other_Owners_`i'  =.a if c4_numowners_confirm_`i'<2
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i'=.a if c4_numowners_confirm_`i'<2
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10b_pers_credcard_bal_`i'==.
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10b_bus_credcard_bal_`i'==.
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10c_pers_loan_bank_amt_`i'==.
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10c_pers_loan_fam_amt_`i'==.
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10c_pers_loan_other_amt_`i'==.
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10c_pers_other_amt_`i'==.
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10b_pers_credcard_bal_`i'==.
```

```
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10b_bus_credcard_bal_`i'==.
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10d_pers_loan_bank_owed_`i'==.
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10d_pers_loan_fam_owed_`i'==.
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10d_pers_loan_other_owed_`i'==.
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10d_pers_other_owed_`i'==.
}

forvalues i = 0/7 {
egen  Tot_Debt_Owner_Operators_`i'=rowtotal(Tot_Pers_Debt_Resp_`i' Tot_Pers_Debt_Other_Owners_`i'),missing
egen  Tot_Debt_Owed_Owner_Operators_`i'=rowtotal(Tot_Pers_Debt_Owed_Resp_`i' Tot_Pers_Debt_Owed_OthrOwnrs_`i'),missing
/* Recode legitimate (hard) missing values */
replace Tot_Debt_Owner_Operators_`i'   =.a if classf_`i'<6
replace Tot_Debt_Owed_Owner_Operators_`i'=.a if classf_`i'<6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Debt_Owner_Operators_`i'   =. if Tot_Pers_Debt_Resp_`i'==.
replace Tot_Debt_Owner_Operators_`i'   =. if Tot_Pers_Debt_Other_Owners_`i'==.
replace Tot_Debt_Owed_Owner_Operators_`i'=. if Tot_Pers_Debt_Owed_Resp_`i'==.
replace Tot_Debt_Owed_Owner_Operators_`i'=. if Tot_Pers_Debt_Owed_OthrOwnrs_`i'==.
}

forvalues i = 0/7 {
egen    Tot_Debt_Bus_`i'=rowtotal(f12b_bus_credcard_bal_`i'    f12c_bus_loans_bank_amt_`i'    f12b_bus_cred_line_bal_`i'
f12c_bus_loans_nonbank_amt_`i'    f12c_bus_loans_fam_amt_`i'    f12c_bus_loans_govt_amt_`i'    f12c_bus_loans_emp_amt_`i'
f12c_bus_loans_other_ind_amt_`i'                f12c_bus_loans_owner_amt_`i'                f12c_bus_loans_bus_amt_`i'
      f12c_bus_other_amt_`i'),missing
egen  Tot_Bus_Debt_Owed_`i'=rowtotal(f12b_bus_cred_line_bal_`i'   f12b_bus_credcard_bal_`i'  f12d_bus_loans_bank_owed_`i'
f12d_bus_loans_nonbank_owed_`i'  f12d_bus_loans_emp_owed_`i'    f12d_bus_loans_fam_owed_`i'  f12d_bus_loans_govt_owed_`i'
f12d_bus_loans_other_ind_owed_`i'          f12d_bus_loans_owner_owed_`i'                    f12d_bus_loans_bus_owed_`i'
f12d_bus_other_owed_`i'),missing
/* Recode legitimate (hard) missing values */
replace Tot_Debt_Bus_`i'     =.a if classf_`i'<6
replace Tot_Bus_Debt_Owed_`i'=.a if classf_`i'<6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Debt_Bus_`i'   =. if f12b_bus_credcard_bal_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12b_bus_cred_line_bal_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_bank_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_nonbank_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_fam_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_govt_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_emp_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_other_ind_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_owner_amt_`i' ==.
```

```
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_bus_amt_`i' ==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_other_amt_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12b_bus_credcard_bal_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12b_bus_cred_line_bal_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_bank_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_nonbank_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_fam_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_govt_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_emp_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_other_ind_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_owner_owed_`i' ==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_bus_owed_`i'  ==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_other_owed_`i'==.
}

forvalues i = 0/7 {
egen  Tot_Debt_`i'=rowtotal(Tot_Debt_Owner_Operators_`i' Tot_Debt_Bus_`i'),missing
egen  Tot_Debt_Owed_`i'=rowtotal(Tot_Debt_Owed_Owner_Operators_`i' Tot_Bus_Debt_Owed_`i'),missing
/* Recode legitimate (hard) missing values */
replace Tot_Debt_`i'      =.a if classf_`i'<6
replace Tot_Debt_Owed_`i' =.a if classf_`i'<6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Debt_`i'   =. if Tot_Debt_Owner_Operators_`i'==.
replace Tot_Debt_`i'   =. if Tot_Debt_Bus_`i'==.
replace Tot_Debt_Owed_`i'=. if Tot_Debt_Owed_Owner_Operators_`i'==.
replace Tot_Debt_Owed_`i'=. if Tot_Bus_Debt_Owed_`i'==.
}

forvalues i = 0/7 {
gen  Net_Profit_`i'=f24_profitloss_amt_`i'
}

/*****************************************************************/
/* Merge the file with the primary owner file "primary_owner.dta" */

merge 1:1 mprid using "xxx:\KFS_Manual_and_Data\primary_owner.dta"
keep if _merge==3
drop _merge

*Replacing missing continuous values by the midpoints of the class intervals

forvalues i = 0/7 {
```

```
foreach ow in $owners_1_15  {
recode       total_hours_owner_`ow'_r_`i'     (0=0) (1=9.5)(2=27.5)(3=40.5)(4=50.5)(5=60.5) (6=70.5)
recode       age_owner_`ow'_r_`i'       (1=21)(2=29.5)(3=39.5)(4=49.5)(5=59.5) (6=69.5) (7=79.5)
    }
      }

forvalues i = 0/7 {
foreach ow in $owners_1_15  {
replace  g1b1_hours_owner_`ow'_`i'=total_hours_owner_`ow'_r_`i'   if  g1b1_hours_owner_`ow'_`i'==.
replace  g4_age_owner_`ow'_`i'=age_owner_`ow'_r_`i'   if  g4_age_owner_`ow'_`i'==.
}
}

drop tot_assets_* tot_liab_* tot_equity_nonowneroperators_* tot_equity_nonownerop_allyrs_* ///
tot_pers_debt_owed_resp_*  tot_pers_debt_resp_* tot_pers_debt_owed_othrownrs_* ///
tot_pers_debt_other_owners_* tot_bus_debt_owed_* tot_debt_bus_* tot_debt_owner_operators_* ///
tot_debt_owed_owner_operators_* tot_debt_* tot_debt_owed_* tot_equity_owner_operators_*  ///
tot_equity_owneroper_allyrs_* tot_debt_liab_equity_* tot_equity_* tot_equity_allyrs_*

drop *_r_*

* Make sure that the race do not change over time due to entry errors
forvalues i = 1/7 {
foreach ow in $owners_1_15  {

replace      g6_race_amind_owner_`ow'_`i'      =     g6_race_amind_owner_`ow'_0 if     g6_race_amind_owner_`ow'_`i'
      <.     &      g6_race_amind_owner_`ow'_0<.
replace      g6_race_asian_owner_`ow'_`i'      =     g6_race_asian_owner_`ow'_0 if     g6_race_asian_owner_`ow'_`i'
      <.     &      g6_race_asian_owner_`ow'_0<.
replace      g6_race_black_owner_`ow'_`i'      =     g6_race_black_owner_`ow'_0 if     g6_race_black_owner_`ow'_`i'
      <.     &      g6_race_black_owner_`ow'_0<.
replace      g6_race_nathaw_owner_`ow'_`i'     =     g6_race_nathaw_owner_`ow'_0       if
      g6_race_nathaw_owner_`ow'_`i'     <.     &     g6_race_nathaw_owner_`ow'_0<.
replace      g6_race_other_owner_`ow'_`i'      =     g6_race_other_owner_`ow'_0 if     g6_race_other_owner_`ow'_`i'
      <.     &      g6_race_other_owner_`ow'_0<.
replace      g6_race_white_owner_`ow'_`i'      =     g6_race_white_owner_`ow'_0 if     g6_race_white_owner_`ow'_`i'
      <.     &      g6_race_white_owner_`ow'_0<.
}
}

* Primary Owner(PO) Characteristics at the Baseline: PO by Robb's Stata code
```

```
forvalues i = 0/0 {
gen  PO_emp_`i'               =.
gen  PO_hours_`i'             =.
gen  PO_work_exp_`i'             =.
gen  PO_oth_bus_owner_`i'     =.
gen  PO_bus_same_ind_`i'      =.
gen  PO_age_owner_`i'            =.
gen  PO_hisp_origin_`i'      =.
gen  PO_race_group_`i'       =.
gen  PO_race_amind_owner_`i'   =.
gen  PO_race_asian_owner_`i'   =.
gen  PO_race_black_owner_`i'   =.
gen  PO_race_nathaw_owner_`i'  =.
gen  PO_race_other_owner_`i'   =.
gen  PO_race_white_owner_`i'   =.
gen  PO_native_born_`i'      =.
gen  PO_us_cit_`i'            =.
gen  PO_education_`i'            =.
gen  PO_gender_`i'            =.
}

forvalues i = 0/0 {
forvalues po = 1/6 {
replace PO_emp_`i'          =   g1a_emp_owner_0`po'_`i'            if  primary_owner==`po'
replace PO_hours_`i'          =  g1b1_hours_owner_0`po'_`i'               if  primary_owner==`po'
replace PO_work_exp_`i'    =     g2_work_exp_owner_0`po'_`i'            if  primary_owner==`po'
replace PO_oth_bus_owner_`i'=    g3a_oth_bus_owner_0`po'_`i'            if  primary_owner==`po'
replace PO_bus_same_ind_`i' =    g3b_bus_same_ind_owner_0`po'_`i' if  primary_owner==`po'
replace PO_age_owner_`i'   =     g4_age_owner_0`po'_`i'                 if  primary_owner==`po'
replace PO_hisp_origin_`i'       =     g5_hisp_origin_owner_0`po'_`i'       if  primary_owner==`po'
replace PO_race_group_`i'   =    g6b_race_group_0`po'_`i'             if  primary_owner==`po'
replace PO_race_amind_owner_`i' = g6_race_amind_owner_0`po'_`i'     if  primary_owner==`po'
replace PO_race_asian_owner_`i' = g6_race_asian_owner_0`po'_`i'     if  primary_owner==`po'
replace PO_race_black_owner_`i' = g6_race_black_owner_0`po'_`i'     if  primary_owner==`po'
replace PO_race_nathaw_owner_`i'= g6_race_nathaw_owner_0`po'_`i'    if  primary_owner==`po'
replace PO_race_other_owner_`i' = g6_race_other_owner_0`po'_`i'     if  primary_owner==`po'
replace PO_race_white_owner_`i' = g6_race_white_owner_0`po'_`i'     if  primary_owner==`po'
replace PO_native_born_`i'       =     g7_native_born_owner_0`po'_`i'       if  primary_owner==`po'
replace PO_us_cit_`i'         =  g8_us_cit_owner_0`po'_`i'            if  primary_owner==`po'
replace PO_education_`i'   =     g9_education_owner_0`po'_`i'           if  primary_owner==`po'
replace PO_gender_`i'         =  g10_gender_owner_0`po'_`i'           if  primary_owner==`po'
}
```

```
}

forvalues i = 1/7 {
gen  PO_emp_`i'          = PO_emp_0
gen  PO_hours_`i'        = PO_hours_0
gen  PO_work_exp_`i'        = PO_work_exp_0
gen  PO_oth_bus_owner_`i'= PO_oth_bus_owner_0
gen  PO_bus_same_ind_`i' = PO_bus_same_ind_0
replace PO_bus_same_ind_`i' =.a if PO_oth_bus_owner_0==0
gen  PO_age_owner_`i'       = PO_age_owner_0
gen  PO_hisp_origin_`i'  = PO_hisp_origin_0
gen  PO_race_group_`i'   = PO_race_group_0
gen  PO_race_amind_owner_`i' =PO_race_amind_owner_0
gen  PO_race_asian_owner_`i' =PO_race_asian_owner_0
gen  PO_race_black_owner_`i' =PO_race_black_owner_0
gen  PO_race_nathaw_owner_`i'=PO_race_nathaw_owner_0
gen  PO_race_other_owner_`i' =PO_race_other_owner_0
gen  PO_race_white_owner_`i' =PO_race_white_owner_0
gen  PO_native_born_`i'  = PO_native_born_0
gen  PO_us_cit_`i'          = PO_us_cit_0
gen  PO_education_`i'       = PO_education_0
gen  PO_gender_`i'       = PO_gender_0
/* Recode legitimate (hard) missing values */
replace  PO_emp_`i'                = .a if    classf_`i'  <6
replace  PO_hours_`i'              = .a if    classf_`i'  <6
replace  PO_work_exp_`i'       = .a if classf_`i'  <6
replace  PO_oth_bus_owner_`i'     = .a if     classf_`i'  <6
replace  PO_bus_same_ind_`i'      = .a if     classf_`i'  <6
replace  PO_age_owner_`i'      = .a if classf_`i'  <6
replace  PO_hisp_origin_`i'       = .a if     classf_`i'  <6
replace  PO_race_group_`i'        = .a if     classf_`i'  <6
replace  PO_race_amind_owner_`i' = .a if      classf_`i'  <6
replace  PO_race_asian_owner_`i' = .a if      classf_`i'  <6
replace  PO_race_black_owner_`i' = .a if      classf_`i'  <6
replace  PO_race_nathaw_owner_`i'= .a if      classf_`i'  <6
replace  PO_race_other_owner_`i' = .a if      classf_`i'  <6
replace  PO_race_white_owner_`i' = .a if      classf_`i'  <6
replace  PO_native_born_`i'       = .a if     classf_`i'  <6
replace  PO_us_cit_`i'             = .a if    classf_`i'  <6
replace  PO_education_`i'      = .a if classf_`i'  <6
replace  PO_gender_`i'             = .a if    classf_`i'  <6
```

```
}

* Active-Owner-Operators Characteristics (OO)

forvalues i = 0/7 {
egen   OO_emp_owner_`i'          =      rowmean(g1a_emp_owner_*_`i')
egen   OO_hours_owner_`i'        = rowmean(g1b1_hours_owner_*_`i')
egen   OO_work_exp_owner_`i'     =      rowmean(g2_work_exp_owner_*_`i'   )
egen   OO_oth_bus_owner_`i'      =      rowmean(g3a_oth_bus_owner_*_`i'   )
egen   OO_bus_same_ind_owner_`i' =      rowmean(g3b_bus_same_ind_owner_*_`i')
egen   OO_age_owner_`i'          =      rowmean(g4_age_owner_*_`i')
egen   OO_hisp_origin_owner_`i'  =      rowmean(g5_hisp_origin_owner_*_`i')
egen   OO_race_amind_owner_`i'   =      rowmean(g6_race_amind_owner_*_`i')
egen   OO_race_asian_owner_`i'   =      rowmean(g6_race_asian_owner_*_`i')
egen   OO_race_black_owner_`i'   =      rowmean(g6_race_black_owner_*_`i')
egen   OO_race_nathaw_owner_`i'  =      rowmean(g6_race_nathaw_owner_*_`i')
egen   OO_race_other_owner_`i'   =      rowmean(g6_race_other_owner_*_`i')
egen   OO_race_white_owner_`i'   =      rowmean(g6_race_white_owner_*_`i')
egen   OO_native_born_owner_`i'  =      rowmean(g7_native_born_owner_*_`i'      )
egen   OO_us_cit_owner_`i'       =      rowmean(g8_us_cit_owner_*_`i'     )
egen   OO_education_owner_`i'    =      rowmean(g9_education_owner_*_`i'  )
egen   md_education_owner_`i'    =      rowmedian(g9_education_owner_*_`i')
gen    OO_D_education_owner_`i'  =(md_education_owner_`i'    >6.99)        if      md_education_owner_`i'    <11

egen   OO_gender_owner_`i'       =      rowmean(g10_gender_owner_*_`i'    )
}

/* Recode to soft missing value if any of the total's component is soft missing*/
forvalues i = 0/7 {
foreach ow in $owners_1_15  {
replace OO_emp_owner_`i'              =. if g1a_emp_owner_`ow'_`i'==.
replace OO_hours_owner_`i'          =. if g1b1_hours_owner_`ow'_`i'==.
replace OO_work_exp_owner_`i'     =. if g2_work_exp_owner_`ow'_`i'==.
replace OO_oth_bus_owner_`i'      =. if g3a_oth_bus_owner_`ow'_`i'==.
replace OO_bus_same_ind_owner_`i' =. if g3b_bus_same_ind_owner_`ow'_`i'==.
replace OO_age_owner_`i'              =. if g4_age_owner_`ow'_`i'==.
replace OO_hisp_origin_owner_`i'  =. if g5_hisp_origin_owner_`ow'_`i'==.
replace OO_race_amind_owner_`i'   =. if g6_race_amind_owner_`ow'_`i'==.
replace OO_race_asian_owner_`i'   =. if g6_race_asian_owner_`ow'_`i'==.
replace OO_race_black_owner_`i'   =. if g6_race_black_owner_`ow'_`i'==.
replace OO_race_nathaw_owner_`i'  =. if g6_race_nathaw_owner_`ow'_`i'==.
replace OO_race_other_owner_`i'   =. if g6_race_other_owner_`ow'_`i'==.
```

```
replace OO_race_white_owner_`i'   =. if g6_race_white_owner_`ow'_`i'==.
replace OO_native_born_owner_`i'  =. if g7_native_born_owner_`ow'_`i'==.
replace OO_us_cit_owner_`i'       =. if g8_us_cit_owner_`ow'_`i'==.
replace OO_education_owner_`i'     =. if g9_education_owner_`ow'_`i'==.
replace md_education_owner_`i'     =. if g9_education_owner_`ow'_`i'==.
replace OO_D_education_owner_`i'  =. if g9_education_owner_`ow'_`i'==.
replace OO_gender_owner_`i'       =. if g10_gender_owner_`ow'_`i'==.
}
}
/* Recode legitimate (hard) missing values */
forvalues i = 0/7 {
replace     OO_bus_same_ind_owner_`i' = .a if   OO_oth_bus_owner_`i'==0
replace      OO_emp_owner_`i'          = .a if classf_`i'  <6
replace      OO_hours_owner_`i'        = .a if classf_`i'  <6
replace      OO_work_exp_owner_`i'     = .a if classf_`i'  <6
replace      OO_oth_bus_owner_`i'      = .a if classf_`i'  <6
replace      OO_bus_same_ind_owner_`i' = .a if classf_`i'  <6
replace      OO_age_owner_`i'          = .a if classf_`i'  <6
replace      OO_hisp_origin_owner_`i'  = .a if classf_`i'  <6
replace      OO_race_amind_owner_`i'   = .a if classf_`i'  <6
replace      OO_race_asian_owner_`i'   = .a if classf_`i'  <6
replace      OO_race_black_owner_`i'   = .a if classf_`i'  <6
replace      OO_race_nathaw_owner_`i'  = .a if classf_`i'  <6
replace      OO_race_other_owner_`i'   = .a if classf_`i'  <6
replace      OO_race_white_owner_`i'   = .a if classf_`i'  <6
replace      OO_native_born_owner_`i'  = .a if classf_`i'  <6
replace      OO_us_cit_owner_`i'       = .a if classf_`i'  <6
replace      OO_education_owner_`i'    = .a if classf_`i'  <6
replace      md_education_owner_`i'    = .a if classf_`i'  <6
replace      OO_D_education_owner_`i'  = .a if classf_`i'  <6
replace      OO_gender_owner_`i'       = .a if classf_`i'  <6
}
/*****************************************************************************/
*Diversity / Similarity index
forvalues i = 0/7 {
gen xr1_`i'= OO_race_amind_owner_`i'    *      OO_race_amind_owner_`i'
gen xr2_`i'= OO_race_asian_owner_`i'    *      OO_race_asian_owner_`i'
gen xr3_`i'= OO_race_black_owner_`i'    *      OO_race_black_owner_`i'
gen xr4_`i'= OO_race_nathaw_owner_`i'*  OO_race_nathaw_owner_`i'
gen xr5_`i'= OO_race_other_owner_`i' *  OO_race_other_owner_`i'
gen xr6_`i'= OO_race_white_owner_`i'    *      OO_race_white_owner_`i'
* Race_similarity
```

```
egen Race_similarity_`i'=rowtotal(xr1_`i' xr2_`i' xr3_`i' xr4_`i' xr5_`i' xr6_`i'),missing
drop xr*_`i'
* Race_ diversity
gen Race_diversity_`i'=1-Race_similarity_`i'
}
forvalues i = 0/7 {
foreach ow in $owners_1_15  {
/* Recode to soft missing value if any of the var's component is soft missing*/
replace Race_similarity_`i'   =. if g6_race_amind_owner_`ow'_`i'==.
replace Race_similarity_`i'   =. if g6_race_asian_owner_`ow'_`i'==.
replace Race_similarity_`i'   =. if g6_race_black_owner_`ow'_`i'==.
replace Race_similarity_`i'   =. if g6_race_nathaw_owner_`ow'_`i'==.
replace Race_similarity_`i'   =. if g6_race_other_owner_`ow'_`i'==.
replace Race_similarity_`i'   =. if g6_race_white_owner_`ow'_`i'==.

}
}
forvalues i = 0/7 {
/* Recode legitimate (hard) missing values */
replace Race_similarity_`i'   =.a if classf_`i'<6
replace Race_diversity_`i'   =.a if Race_similarity_`i'==.a
replace Race_diversity_`i'   =.  if Race_similarity_`i'==.
}

forvalues i = 0/7 {
gen fmal_`i'=1-OO_gender_owner_`i'
gen xr1_`i'= OO_gender_owner_`i' *     OO_gender_owner_`i'
gen xr2_`i'= fmal_`i'     *     fmal_`i'

* Gender_similarity
egen Gender_similarity_`i'=rowtotal(xr1_`i' xr2_`i'  ),missing
drop xr*_`i' fmal_`i'
* Gender_diversity
gen Gender_diversity_`i'=1-Gender_similarity_`i'
}

forvalues i = 0/7 {
foreach ow in $owners_1_15  {
/* Recode to soft missing value if any of the var's component is soft missing*/
replace  Gender_similarity_`i'      =. if g10_gender_owner_`ow'_`i'==.
}
}
```

```
forvalues i = 0/7 {
/* Recode legitimate (hard) missing values */
replace Gender_similarity_`i'   =.a if classf_`i'<6
replace Gender_diversity_`i'   =.a if Gender_similarity_`i'==.a
replace Gender_diversity_`i'   =.  if Gender_similarity_`i'==.
}

forvalues i = 0/7 {
mean Race_similarity_`i' Gender_similarity_`i' if c4_numowners_confirm_`i'>1 & c4_numowners_confirm_`i'<.
}

*Business level Characteristics
forvalues i = 0/7 {
*Home Based Dummy
recode c8_primary_loc_`i' (1=1 "Home Based") (nonmiss=0 "Non Home Based" ) ,into (Home_Based_`i')
*Sole_Proprietorship Dummy
recode    c1z2_legal_status_`i'    (1=1    "Sole_Proprietorship")    (nonmiss=0    "Limited    Liability"    )    ,into
(Sole_Proprietorship_`i')
rename d2_comp_advantage_`i' Comp_advantage_`i'
egen Have_IP_`i'=anymatch(d3_a_have_patent_`i' d3_b_have_copyright_`i' d3_c_have_trademark_`i'), values(1)
rename c5_num_employees_`i'     Full_Part_Time_Employees_`i'
rename c6_num_ft_employees_`i'  Full_Time_Employees_`i'
rename c7_num_pt_employees_`i'  Part_Time_Employees_`i'
egen   Employee_Owner_`i'          =    rowtotal(g1a_emp_owner_*_`i')
egen   Total_Employees_`i'=rowtotal(Employee_Owner_`i' Full_Part_Time_Employees_`i')
}

forvalues i = 0/7 {
foreach ow in $owners_1_15  {
/* Recode to soft missing value if any of the var's component is soft missing*/
replace Have_IP_`i'=. if d3_a_have_patent_`i'==. | d3_b_have_copyright_`i'==. | d3_c_have_trademark_`i'==.
replace Employee_Owner_`i'             =. if g1a_emp_owner_`ow'_`i'==.
replace Total_Employees_`i'            =. if g1a_emp_owner_`ow'_`i'==.
replace Total_Employees_`i'            =. if  Full_Part_Time_Employees_`i'==.
/* Recode legitimate (hard) missing values */
replace Have_IP_`i'        =.a if classf_`i'<6
replace Employee_Owner_`i' =.a if classf_`i'<6
replace Total_Employees_`i'=.a if classf_`i'<6
}
}
```

```
drop cswgt_final_0 wgt_1_long wgt_2_long wgt_3_long wgt_4_long wgt_5_long wgt_6_long

saveold KFS8_LI_L_w1,replace

forvalues i = 0/7 {

  set logtype text , permanently

  log using KFS8_LI_L_`i'.csv, replace
    FR_Sum_CS *_`i' [pw=wgt_7_long]
  log close
}
```

3.2.2.2.6. Stata Code: Cross Sectional in Long Format

The following Stata code will create the new variables using the KFS long format file "KFS8_LI_CS_Long." The code will save the new cross-sectional (n= 39424) file under the name "KFS8_CS_L1.dta."

```
clear
clear mata
clear matrix
capture log close
set more off
set maxvar 32767
program drop _all
adopath + " xxx:\KFS_Manual_and_Data\Farhat_Robb_Commands\"
cd "Xxx:\KFS_Manual_and_Data"
  use KFS8_LI_CS_Long,clear

drop tot_assets tot_liab tot_equity_nonowneroperators tot_equity_nonownerop_allyrs ///
tot_pers_debt_owed_resp  tot_pers_debt_resp tot_pers_debt_owed_othrownrs ///
tot_pers_debt_other_owners tot_bus_debt_owed tot_debt_bus tot_debt_owner_operators ///
tot_debt_owed_owner_operators tot_debt tot_debt_owed tot_equity_owner_operators  ///
tot_equity_owneroper_allyrs tot_debt_liab_equity tot_equity tot_equity_allyrs



*Replacing missing continuous values by the midpoints of the class intervals

foreach totfin in tot_*_r {
recode `totfin'     (0=0)(1=250)(2=750)(3=2000)(4=4000)(5=7500) (6=17500) (7=62500) (8=550000) (9=1000000)
    }

global owners_1_15 "01 02 03 04 05 06 07 08 09 10 11 12 13 14 15"

foreach ow in $owners_1_15  {
replace  f2_owner_amt_eq_invest_`ow' =tot_equity_owner_`ow'_r    if  f2_owner_amt_eq_invest_`ow' ==.
replace  f2_ownr_amt_eqinvest_allyrs_`ow' =tot_equity_allyrs_owner_`ow'_r    if  f2_ownr_amt_eqinvest_allyrs_`ow' ==.
}

replace  f4_eq_amt_angels =tot_equity_angels_r    if  f4_eq_amt_angels ==.
replace  f4_eq_amt_angels_allyrs =tot_equity_angels_allyrs_r    if  f4_eq_amt_angels_allyrs ==.
replace  f4_eq_amt_companies =tot_equity_companies_r    if  f4_eq_amt_companies ==.
replace  f4_eq_amt_companies_allyrs =tot_equity_companies_allyrs_r    if  f4_eq_amt_companies_allyrs ==.
```

```
replace  f4_eq_amt_govt =tot_equity_govt_r    if  f4_eq_amt_govt ==.
replace  f4_eq_amt_govt_allyrs =tot_equity_govt_allyrs_r    if  f4_eq_amt_govt_allyrs ==.
replace  f4_eq_amt_other =tot_equity_other_r    if  f4_eq_amt_other ==.
replace  f4_eq_amt_other_allyrs =tot_equity_other_allyrs_r    if  f4_eq_amt_other_allyrs ==.
replace  f4_eq_amt_parents =tot_equity_parents_r    if  f4_eq_amt_parents ==.
replace  f4_eq_amt_parents_allyrs =tot_equity_parents_allyrs_r    if  f4_eq_amt_parents_allyrs ==.
replace  f4_eq_amt_spouse =tot_equity_spouse_r    if  f4_eq_amt_spouse ==.
replace  f4_eq_amt_spouse_allyrs =tot_equity_spouse_allyrs_r    if  f4_eq_amt_spouse_allyrs ==.
replace  f4_eq_amt_vent_cap =tot_equity_vent_cap_r    if  f4_eq_amt_vent_cap ==.
replace  f4_eq_amt_vent_cap_allyrs =tot_equity_vent_cap_allyrs_r    if  f4_eq_amt_vent_cap_allyrs ==.
replace  f6b_personal_use_amt =tot_personal_use_r    if  f6b_personal_use_amt ==.
replace  f8a_bus_credcard_line =tot_bus_credcard_line_resp_r    if  f8a_bus_credcard_line ==.
replace  f8b_bus_credcard_bal =tot_bus_credcard_bal_resp_r    if  f8b_bus_credcard_bal ==.
replace  f8a_pers_credcard_line =tot_pers_credcard_line_resp_r    if  f8a_pers_credcard_line ==.
replace  f8b_pers_credcard_bal =tot_pers_credcard_bal_resp_r    if  f8b_pers_credcard_bal ==.
replace  f8c_pers_loan_bank_amt =tot_pers_loan_bank_resp_r    if  f8c_pers_loan_bank_amt ==.
replace  f8d_pers_loan_bank_owed =tot_pers_loan_bank_owed_resp_r    if  f8d_pers_loan_bank_owed ==.
replace  f8c_pers_loan_fam_amt =tot_pers_loan_fam_resp_r    if  f8c_pers_loan_fam_amt ==.
replace  f8d_pers_loan_fam_owed =tot_pers_loan_fam_owed_resp_r    if  f8d_pers_loan_fam_owed ==.
replace  f8c_pers_loan_other_amt =tot_pers_loan_other_resp_r    if  f8c_pers_loan_other_amt ==.
replace  f8d_pers_loan_other_owed =tot_persloan_other_owed_resp_r    if  f8d_pers_loan_other_owed ==.
replace  f8c_pers_other_amt =tot_pers_other_resp_r    if  f8c_pers_other_amt ==.
replace  f8d_pers_other_owed =tot_pers_other_owed_resp_r    if  f8d_pers_other_owed ==.
replace  f10a_bus_credcard_line =tot_bus_credcard_line_others_r    if  f10a_bus_credcard_line ==.
replace  f10b_bus_credcard_bal =tot_bus_credcard_bal_others_r    if  f10b_bus_credcard_bal ==.
replace  f10a_pers_credcard_line =tot_pers_credcard_line_othrs_r    if  f10a_pers_credcard_line ==.
replace  f10b_pers_credcard_bal =tot_pers_credcard_bal_others_r    if  f10b_pers_credcard_bal ==.
replace  f10c_pers_loan_bank_amt =tot_pers_loan_bank_others_r    if  f10c_pers_loan_bank_amt ==.
replace  f10d_pers_loan_bank_owed =tot_persloan_bank_owed_othrs_r    if  f10d_pers_loan_bank_owed ==.
replace  f10c_pers_loan_fam_amt =tot_persloan_fam_othrowners_r    if  f10c_pers_loan_fam_amt ==.
replace  f10d_pers_loan_fam_owed =tot_persloan_fam_owed_othrs_r    if  f10d_pers_loan_fam_owed ==.
replace  f10c_pers_loan_other_amt =tot_pers_loan_other_owners_r    if  f10c_pers_loan_other_amt ==.
replace  f10d_pers_loan_other_owed =tot_persloan_othr_owed_othrs_r    if  f10d_pers_loan_other_owed ==.
replace  f10c_pers_other_amt =tot_pers_other_other_owners_r    if  f10c_pers_other_amt ==.
replace  f10d_pers_other_owed =tot_pers_other_owed_others_r    if  f10d_pers_other_owed ==.
replace  f12a_bus_cred_line =tot_cred_line_bus_line_r    if  f12a_bus_cred_line ==.
replace  f12b_bus_cred_line_bal =tot_cred_line_bus_bal_r    if  f12b_bus_cred_line_bal ==.
replace  f12a_bus_credcard_line =tot_credcard_line_bus_r    if  f12a_bus_credcard_line ==.
replace  f12b_bus_credcard_bal =tot_credcard_bal_bus_r    if  f12b_bus_credcard_bal ==.
replace  f12c_bus_loans_bank_amt =tot_loan_bank_bus_r    if  f12c_bus_loans_bank_amt ==.
replace  f12d_bus_loans_bank_owed =tot_bus_loans_bank_owed_r    if  f12d_bus_loans_bank_owed ==.
replace  f12c_bus_loans_nonbank_amt =tot_loan_nonbank_bus_r    if  f12c_bus_loans_nonbank_amt ==.
```

```
replace  f12d_bus_loans_nonbank_owed =tot_bus_loans_nonbank_owed_r    if  f12d_bus_loans_nonbank_owed ==.
replace  f12c_bus_loans_emp_amt =tot_loan_emp_bus_r    if  f12c_bus_loans_emp_amt ==.
replace  f12d_bus_loans_emp_owed =tot_bus_loans_emp_owed_r    if  f12d_bus_loans_emp_owed ==.
replace  f12c_bus_loans_fam_amt =tot_loan_fam_bus_r    if  f12c_bus_loans_fam_amt ==.
replace  f12d_bus_loans_fam_owed =tot_bus_loans_fam_owed_r    if  f12d_bus_loans_fam_owed ==.
replace  f12c_bus_loans_govt_amt =tot_loan_govt_bus_r    if  f12c_bus_loans_govt_amt ==.
replace  f12d_bus_loans_govt_owed =tot_bus_loans_govt_owed_r    if  f12d_bus_loans_govt_owed ==.
replace  f12c_bus_loans_other_ind_amt =tot_loan_other_ind_r    if  f12c_bus_loans_other_ind_amt ==.
replace  f12d_bus_loans_other_ind_owed =tot_bus_loans_otherind_owed_r    if  f12d_bus_loans_other_ind_owed ==.
replace  f12c_bus_loans_owner_amt =tot_loan_owner_bus_r    if  f12c_bus_loans_owner_amt ==.
replace  f12d_bus_loans_owner_owed =tot_bus_loans_owner_owed_r    if  f12d_bus_loans_owner_owed ==.
replace  f12c_bus_loans_bus_amt =tot_loan_other_bus_r    if  f12c_bus_loans_bus_amt ==.
replace  f12d_bus_loans_bus_owed =tot_bus_loans_otherbus_owed_r    if  f12d_bus_loans_bus_owed ==.
replace  f12c_bus_other_amt =tot_bus_debt_other_r    if  f12c_bus_other_amt ==.
replace  f12d_bus_other_owed =tot_bus_loans_other_owed_r    if  f12d_bus_other_owed ==.
replace  f14a_trade_fin_amt =tot_trade_finan_r    if  f14a_trade_fin_amt ==.
replace  f16a_rev_amt =tot_revenue_r    if  f16a_rev_amt ==.
replace  f17a_total_exp_amt =tot_expenses_r    if  f17a_total_exp_amt ==.
replace  f18a_wage_exp_amt =tot_wages_r    if  f18a_wage_exp_amt ==.
replace  f19a_res_dev_amt =tot_res_dev_r    if  f19a_res_dev_amt ==.
replace  f19c_a_design_amt =tot_intangassets_design_r    if  f19c_a_design_amt ==.
replace  f19c_b_investments_amt =tot_intangassets_invest_r    if  f19c_b_investments_amt ==.
replace  f19c_c_brand_dev_amt =tot_intangassets_branddev_r    if  f19c_c_brand_dev_amt ==.
replace  f19c_d_org_dev_amt =tot_intangassets_orgdev_r    if  f19c_d_org_dev_amt ==.
replace  f19c_e_worker_training_amt =tot_intangassets_wkrtrng_r    if  f19c_e_worker_training_amt ==.
replace  f19c_f_other_amt =tot_intangassets_other_r    if  f19c_f_other_amt ==.
replace  f19c_intangassets_amt =tot_intang_assets_r    if  f19c_intangassets_amt ==.
replace  f24_profit_amt =tot_profit_r    if  f24_profit_amt ==.
replace  f26_loss_amt =tot_loss_r    if  f26_loss_amt ==.
replace  f29_assetval_cash =tot_asset_cash_r    if  f29_assetval_cash ==.
replace  f29_assetval_acctrec =tot_asset_acct_rec_r    if  f29_assetval_acctrec ==.
replace  f29_assetval_inv =tot_asset_inv_r    if  f29_assetval_inv ==.
replace  f29_assetval_equip =tot_asset_equip_r    if  f29_assetval_equip ==.
replace  f29_assetval_landbuild =tot_asset_landbuild_r    if  f29_assetval_landbuild ==.
replace  f29_assetval_veh =tot_asset_veh_r    if  f29_assetval_veh ==.
replace  f29_assetval_othbusprop =tot_asset_other_bus_prop_r    if  f29_assetval_othbusprop ==.
replace  f29_assetval_other =tot_asset_other_r    if  f29_assetval_other ==.
replace  f31_value_acctpay =tot_liab_acct_pay_r    if  f31_value_acctpay ==.
replace  f31_value_pension =tot_liab_pension_r    if  f31_value_pension ==.
replace  f31_value_other =tot_liab_other_r    if  f31_value_other ==.

egen   Tot_Equity_Owner_Operators =     rowtotal(    f2_owner_amt_eq_invest_*   )      , missing
```

```
egen   Tot_Equity_OwnerOper_AllYrs =rowtotal(  f2_ownr_amt_eqinvest_allyrs_*     ), missing

/* Recode legitimate (hard) missing values */
replace Tot_Equity_Owner_Operators  =.a if classf <6
replace Tot_Equity_OwnerOper_AllYrs =.a if classf <6

foreach ow in $owners_1_15  {
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Equity_Owner_Operators  =.   if  f2_owner_amt_eq_invest_`ow' ==.
replace Tot_Equity_OwnerOper_AllYrs =.   if  f2_ownr_amt_eqinvest_allyrs_`ow' ==.
}


global List1 "spouse parents angels companies govt vent_cap other"

egen  Tot_Equity_NonOwnerOperators =rowtotal(f4_eq_amt_angels  f4_eq_amt_companies  f4_eq_amt_govt  ///
f4_eq_amt_other  f4_eq_amt_parents  f4_eq_amt_spouse   f4_eq_amt_vent_cap )     , missing
egen  Tot_Equity_NonOwnerOp_AllYrs =rowtotal(f4_eq_amt_*_allyrs )  , missing
/* Recode legitimate (hard) missing values */
replace Tot_Equity_NonOwnerOperators  =.a if classf <6
replace Tot_Equity_NonOwnerOperators  =.a if  c1z2_legal_status ==1
replace Tot_Equity_NonOwnerOperators  =.a if  one_owner ==1
replace Tot_Equity_NonOwnerOp_AllYrs  =.a if classf <6
replace Tot_Equity_NonOwnerOp_AllYrs  =.a if  c1z2_legal_status ==1
replace Tot_Equity_NonOwnerOp_AllYrs  =.a if  one_owner ==1

foreach name in $List1  {
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Equity_NonOwnerOperators  =. if f4_eq_amt_`name' ==.
replace Tot_Equity_NonOwnerOp_AllYrs  =. if f4_eq_amt_`name'_allyrs ==.
}

egen  Tot_Equity        =rowtotal(Tot_Equity_Owner_Operators  Tot_Equity_NonOwnerOperators ), missing
egen  Tot_Equity_AllYrs =rowtotal(Tot_Equity_OwnerOper_AllYrs  Tot_Equity_NonOwnerOp_AllYrs ), missing
/* Recode legitimate (hard) missing values */
replace Tot_Equity        =.a if classf <6
replace Tot_Equity_AllYrs =.a if classf <6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Equity        =. if Tot_Equity_Owner_Operators ==.
replace Tot_Equity        =. if Tot_Equity_NonOwnerOperators ==.
replace Tot_Equity_AllYrs =. if Tot_Equity_NonOwnerOperators ==.
replace Tot_Equity_AllYrs =. if Tot_Equity_NonOwnerOp_AllYrs ==.
```

```
egen  Tot_Assets =rowtotal(f29_assetval_* )    , missing
egen  Tot_Liab =rowtotal(f31_value_* ), missing
/* Recode legitimate (hard) missing values */
replace Tot_Assets  =.a if classf <6
replace Tot_Liab    =.a if classf <6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Assets  =. if f29_assetval_acctrec ==.
replace Tot_Assets  =. if f29_assetval_cash ==.
replace Tot_Assets  =. if f29_assetval_equip ==.
replace Tot_Assets  =. if f29_assetval_inv ==.
replace Tot_Assets  =. if f29_assetval_landbuild ==.
replace Tot_Assets  =. if f29_assetval_othbusprop ==.
replace Tot_Assets  =. if f29_assetval_other ==.
replace Tot_Assets  =. if f29_assetval_veh ==.
replace Tot_Liab  =. if  f31_value_acctpay ==.
replace Tot_Liab  =. if  f31_value_other ==.
replace Tot_Liab  =. if  f31_value_pension ==.

egen       Tot_Pers_Debt_Resp    =rowtotal(f8b_pers_credcard_bal       f8b_bus_credcard_bal       f8c_pers_loan_bank_amt
f8c_pers_loan_fam_amt  f8c_pers_loan_other_amt  f8c_pers_other_amt ),missing
egen     Tot_Pers_Debt_Owed_Resp   =rowtotal(f8b_pers_credcard_bal     f8b_bus_credcard_bal     f8d_pers_loan_bank_owed
f8d_pers_loan_fam_owed  f8d_pers_loan_other_owed  f8d_pers_other_owed ),missing
/* Recode legitimate (hard) missing values */
replace Tot_Pers_Debt_Resp       =.a if classf <6
replace Tot_Pers_Debt_Owed_Resp  =.a if classf <6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Pers_Debt_Resp  =. if f8b_pers_credcard_bal ==.
replace Tot_Pers_Debt_Resp  =. if f8b_bus_credcard_bal ==.
replace Tot_Pers_Debt_Resp  =. if f8c_pers_loan_bank_amt ==.
replace Tot_Pers_Debt_Resp  =. if f8c_pers_loan_fam_amt ==.
replace Tot_Pers_Debt_Resp  =. if f8c_pers_loan_other_amt ==.
replace Tot_Pers_Debt_Resp  =. if f8c_pers_other_amt ==.
replace Tot_Pers_Debt_Owed_Resp  =. if f8b_pers_credcard_bal ==.
replace Tot_Pers_Debt_Owed_Resp  =. if f8b_bus_credcard_bal ==.
replace Tot_Pers_Debt_Owed_Resp  =. if f8d_pers_loan_bank_owed ==.
replace Tot_Pers_Debt_Owed_Resp  =. if f8d_pers_loan_fam_owed ==.
replace Tot_Pers_Debt_Owed_Resp  =. if f8d_pers_loan_other_owed ==.
replace Tot_Pers_Debt_Owed_Resp  =. if f8d_pers_other_owed ==.

egen    Tot_Pers_Debt_Other_Owners  =rowtotal(f10b_pers_credcard_bal    f10c_pers_loan_bank_amt    f10b_bus_credcard_bal
f10c_pers_loan_fam_amt  f10c_pers_loan_other_amt   f10c_pers_other_amt ),missing
```

```
egen   Tot_Pers_Debt_Owed_OthrOwnrs =rowtotal(f10b_pers_credcard_bal   f10b_bus_credcard_bal   f10d_pers_loan_bank_owed
f10d_pers_loan_fam_owed  f10d_pers_loan_other_owed  f10d_pers_other_owed ),missing
/* Recode legitimate (hard) missing values */
replace Tot_Pers_Debt_Other_Owners   =.a if classf <6
replace Tot_Pers_Debt_Owed_OthrOwnrs =.a if classf <6
replace Tot_Pers_Debt_Other_Owners   =.a if c4_numowners_confirm <2
replace Tot_Pers_Debt_Owed_OthrOwnrs =.a if c4_numowners_confirm <2
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Pers_Debt_Other_Owners  =. if f10b_pers_credcard_bal ==.
replace Tot_Pers_Debt_Other_Owners  =. if f10b_bus_credcard_bal ==.
replace Tot_Pers_Debt_Other_Owners  =. if f10c_pers_loan_bank_amt ==.
replace Tot_Pers_Debt_Other_Owners  =. if f10c_pers_loan_fam_amt ==.
replace Tot_Pers_Debt_Other_Owners  =. if f10c_pers_loan_other_amt ==.
replace Tot_Pers_Debt_Other_Owners  =. if f10c_pers_other_amt ==.
replace Tot_Pers_Debt_Owed_OthrOwnrs  =. if f10b_pers_credcard_bal ==.
replace Tot_Pers_Debt_Owed_OthrOwnrs  =. if f10b_bus_credcard_bal ==.
replace Tot_Pers_Debt_Owed_OthrOwnrs  =. if f10d_pers_loan_bank_owed ==.
replace Tot_Pers_Debt_Owed_OthrOwnrs  =. if f10d_pers_loan_fam_owed ==.
replace Tot_Pers_Debt_Owed_OthrOwnrs  =. if f10d_pers_loan_other_owed ==.
replace Tot_Pers_Debt_Owed_OthrOwnrs  =. if f10d_pers_other_owed ==.

egen  Tot_Debt_Owner_Operators =rowtotal(Tot_Pers_Debt_Resp  Tot_Pers_Debt_Other_Owners ),missing
egen  Tot_Debt_Owed_Owner_Operators =rowtotal(Tot_Pers_Debt_Owed_Resp  Tot_Pers_Debt_Owed_OthrOwnrs ),missing
/* Recode legitimate (hard) missing values */
replace Tot_Debt_Owner_Operators    =.a if classf <6
replace Tot_Debt_Owed_Owner_Operators =.a if classf <6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Debt_Owner_Operators    =. if Tot_Pers_Debt_Resp ==.
replace Tot_Debt_Owner_Operators    =. if Tot_Pers_Debt_Other_Owners ==.
replace Tot_Debt_Owed_Owner_Operators =. if Tot_Pers_Debt_Owed_Resp ==.
replace Tot_Debt_Owed_Owner_Operators =. if Tot_Pers_Debt_Owed_OthrOwnrs ==.

egen     Tot_Debt_Bus     =rowtotal(f12b_bus_credcard_bal         f12c_bus_loans_bank_amt          f12b_bus_cred_line_bal
f12c_bus_loans_nonbank_amt       f12c_bus_loans_fam_amt          f12c_bus_loans_govt_amt          f12c_bus_loans_emp_amt
f12c_bus_loans_other_ind_amt  f12c_bus_loans_owner_amt  f12c_bus_loans_bus_amt  f12c_bus_other_amt ),missing
egen   Tot_Bus_Debt_Owed   =rowtotal(f12b_bus_cred_line_bal          f12b_bus_credcard_bal      f12d_bus_loans_bank_owed
f12d_bus_loans_nonbank_owed      f12d_bus_loans_emp_owed          f12d_bus_loans_fam_owed      f12d_bus_loans_govt_owed
f12d_bus_loans_other_ind_owed  f12d_bus_loans_owner_owed   f12d_bus_loans_bus_owed  f12d_bus_other_owed ),missing
/* Recode legitimate (hard) missing values */
replace Tot_Debt_Bus      =.a if classf <6
replace Tot_Bus_Debt_Owed =.a if classf <6
/* Recode to soft missing value if any of the total's component is soft missing*/
```

```
replace Tot_Debt_Bus    =. if f12b_bus_credcard_bal ==.
replace Tot_Debt_Bus    =. if f12b_bus_cred_line_bal ==.
replace Tot_Debt_Bus    =. if f12c_bus_loans_bank_amt ==.
replace Tot_Debt_Bus    =. if f12c_bus_loans_nonbank_amt ==.
replace Tot_Debt_Bus    =. if f12c_bus_loans_fam_amt ==.
replace Tot_Debt_Bus    =. if f12c_bus_loans_govt_amt ==.
replace Tot_Debt_Bus    =. if f12c_bus_loans_emp_amt ==.
replace Tot_Debt_Bus    =. if f12c_bus_loans_other_ind_amt ==.
replace Tot_Debt_Bus    =. if f12c_bus_loans_owner_amt  ==.
replace Tot_Debt_Bus    =. if f12c_bus_loans_bus_amt ==.
replace Tot_Debt_Bus    =. if f12c_bus_other_amt ==.
replace Tot_Bus_Debt_Owed    =. if f12b_bus_credcard_bal ==.
replace Tot_Bus_Debt_Owed    =. if f12b_bus_cred_line_bal ==.
replace Tot_Bus_Debt_Owed    =. if f12d_bus_loans_bank_owed ==.
replace Tot_Bus_Debt_Owed    =. if f12d_bus_loans_nonbank_owed ==.
replace Tot_Bus_Debt_Owed    =. if f12d_bus_loans_fam_owed ==.
replace Tot_Bus_Debt_Owed    =. if f12d_bus_loans_govt_owed ==.
replace Tot_Bus_Debt_Owed    =. if f12d_bus_loans_emp_owed ==.
replace Tot_Bus_Debt_Owed    =. if f12d_bus_loans_other_ind_owed ==.
replace Tot_Bus_Debt_Owed    =. if f12d_bus_loans_owner_owed  ==.
replace Tot_Bus_Debt_Owed    =. if f12d_bus_loans_bus_owed  ==.
replace Tot_Bus_Debt_Owed    =. if f12d_bus_other_owed ==.
egen  Tot_Debt =rowtotal(Tot_Debt_Owner_Operators  Tot_Debt_Bus ),missing
egen  Tot_Debt_Owed =rowtotal(Tot_Debt_Owed_Owner_Operators  Tot_Bus_Debt_Owed ),missing
/* Recode legitimate (hard) missing values */
replace Tot_Debt       =.a if classf <6
replace Tot_Debt_Owed  =.a if classf <6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Debt    =. if Tot_Debt_Owner_Operators ==.
replace Tot_Debt    =. if Tot_Debt_Bus ==.
replace Tot_Debt_Owed =. if Tot_Debt_Owed_Owner_Operators ==.
replace Tot_Debt_Owed =. if Tot_Bus_Debt_Owed ==.

gen  Net_Profit =f24_profitloss_amt
/*****************************************************************/
/* Merge the file with the primary owner file "primary_owner.dta" */

merge m:1 mprid using "Xxx:\KFS_Manual_and_Data\primary_owner.dta"
drop _merge

*Replacing missing continuous values by the midpoints of the class intervals
```

```
foreach ow in $owners_1_15  {
recode       total_hours_owner_`ow'_r   (0=0) (1=9.5)(2=27.5)(3=40.5)(4=50.5)(5=60.5) (6=70.5)
recode       age_owner_`ow'_r    (1=21)(2=29.5)(3=39.5)(4=49.5)(5=59.5) (6=69.5) (7=79.5)
    }

foreach ow in $owners_1_15  {
replace  g1b1_hours_owner_`ow' =total_hours_owner_`ow'_r    if  g1b1_hours_owner_`ow' ==.
replace  g4_age_owner_`ow' =age_owner_`ow'_r    if  g4_age_owner_`ow' ==.
}

drop *_r

* Primary Owner(PO) Characteristics at the Baseline: PO by Robb's Stata code
gen  PO_emp                    =.
gen  PO_hours                  =.
gen  PO_work_exp           =.
gen  PO_oth_bus_owner       =.
gen  PO_bus_same_ind        =.
gen  PO_age_owner          =.
gen  PO_hisp_origin           =.
gen  PO_race_group            =.
gen  PO_race_amind_owner    =.
gen  PO_race_asian_owner    =.
gen  PO_race_black_owner    =.
gen  PO_race_nathaw_owner   =.
gen  PO_race_other_owner    =.
gen  PO_race_white_owner    =.
gen  PO_native_born           =.
gen  PO_us_cit                =.
gen  PO_education          =.
gen  PO_gender                 =.

sort mprid year
xtset mprid year

forvalues po = 1/6 {
bysort mprid (year):replace PO_emp           =   g1a_emp_owner_0`po'[1]              if  primary_owner==`po'
bysort mprid (year):replace PO_hours        =  g1b1_hours_owner_0`po'[1]                  if  primary_owner==`po'
bysort mprid (year):replace PO_work_exp        =      g2_work_exp_owner_0`po'[1]             if  primary_owner==`po'
bysort mprid (year):replace PO_oth_bus_owner = g3a_oth_bus_owner_0`po'[1]              if  primary_owner==`po'
bysort mprid (year):replace PO_bus_same_ind  = g3b_bus_same_ind_owner_0`po'[1]  if  primary_owner==`po'
bysort mprid (year):replace PO_age_owner       =      g4_age_owner_0`po'[1]                  if  primary_owner==`po'
```

```
bysort mprid (year):replace PO_hisp_origin     =      g5_hisp_origin_owner_0`po'[1]        if  primary_owner==`po'
bysort mprid (year):replace PO_race_group    = g6b_race_group_0`po'[1]               if  primary_owner==`po'
bysort mprid (year):replace PO_race_amind_owner  = g6_race_amind_owner_0`po'[1]      if  primary_owner==`po'
bysort mprid (year):replace PO_race_asian_owner  = g6_race_asian_owner_0`po'[1]      if  primary_owner==`po'
bysort mprid (year):replace PO_race_black_owner  = g6_race_black_owner_0`po'[1]      if  primary_owner==`po'
bysort mprid (year):replace PO_race_nathaw_owner = g6_race_nathaw_owner_0`po'[1]     if  primary_owner==`po'
bysort mprid (year):replace PO_race_other_owner  = g6_race_other_owner_0`po'[1]      if  primary_owner==`po'
bysort mprid (year):replace PO_race_white_owner  = g6_race_white_owner_0`po'[1]      if  primary_owner==`po'
bysort mprid (year):replace PO_native_born     =      g7_native_born_owner_0`po'[1]        if  primary_owner==`po'
bysort mprid (year):replace PO_us_cit       =  g8_us_cit_owner_0`po'[1]             if  primary_owner==`po'
bysort mprid (year):replace PO_education       =      g9_education_owner_0`po'[1]          if  primary_owner==`po'
bysort mprid (year):replace PO_gender       =  g10_gender_owner_0`po'[1]            if  primary_owner==`po'
}

replace  PO_emp                = .a if  classf   <6
replace  PO_hours            = .a if    classf   <6
replace  PO_work_exp           = .a if  classf   <6
replace  PO_oth_bus_owner         = .a if     classf   <6
replace  PO_bus_same_ind     = .a if    classf   <6
replace  PO_age_owner          = .a if  classf   <6
replace  PO_hisp_origin     = .a if     classf   <6
replace  PO_race_group      = .a if     classf   <6
replace  PO_race_amind_owner  = .a if   classf   <6
replace  PO_race_asian_owner  = .a if   classf   <6
replace  PO_race_black_owner  = .a if   classf   <6
replace  PO_race_nathaw_owner = .a if   classf   <6
replace  PO_race_other_owner  = .a if   classf   <6
replace  PO_race_white_owner  = .a if   classf   <6
replace  PO_native_born     = .a if     classf   <6
replace  PO_us_cit             = .a if  classf   <6
replace  PO_education          = .a if  classf   <6
replace  PO_gender             = .a if  classf   <6

* Active-Owner-Operators Characteristics (OO)
egen   OO_emp_owner           =   rowmean(g1a_emp_owner_* )
egen   OO_hours_owner         = rowmean(g1b1_hours_owner_* )
egen   OO_work_exp_owner      =   rowmean(g2_work_exp_owner_*       )
egen   OO_oth_bus_owner       =   rowmean(g3a_oth_bus_owner_*       )
egen   OO_bus_same_ind_owner  =   rowmean(g3b_bus_same_ind_owner_* )
egen   OO_age_owner           =   rowmean(g4_age_owner_* )
egen   OO_hisp_origin_owner   =   rowmean(g5_hisp_origin_owner_* )
egen   OO_race_amind_owner    =   rowmean(g6_race_amind_owner_* )
```

```
egen   OO_race_asian_owner    =   rowmean(g6_race_asian_owner_* )
egen   OO_race_black_owner    =   rowmean(g6_race_black_owner_* )
egen   OO_race_nathaw_owner   =   rowmean(g6_race_nathaw_owner_* )
egen   OO_race_other_owner    =   rowmean(g6_race_other_owner_* )
egen   OO_race_white_owner    =   rowmean(g6_race_white_owner_* )
egen   OO_native_born_owner   =   rowmean(g7_native_born_owner_*    )
egen   OO_us_cit_owner        =   rowmean(g8_us_cit_owner_*  )
egen   OO_education_owner     =   rowmean(g9_education_owner_*     )
egen   md_education_owner     =   rowmedian(g9_education_owner_* )
gen    OO_D_education_owner   =(md_education_owner     >6.99)        if      md_education_owner     <11
egen   OO_gender_owner        =   rowmean(g10_gender_owner_*       )

/* Recode to soft missing value if any of the total's component is soft missing*/
foreach ow in $owners_1_15  {
replace OO_emp_owner             =. if g1a_emp_owner_`ow' ==.
replace OO_hours_owner              =. if g1b1_hours_owner_`ow' ==.
replace OO_work_exp_owner      =. if g2_work_exp_owner_`ow' ==.
replace OO_oth_bus_owner       =. if g3a_oth_bus_owner_`ow' ==.
replace OO_bus_same_ind_owner  =. if g3b_bus_same_ind_owner_`ow' ==.
replace OO_age_owner             =. if g4_age_owner_`ow' ==.
replace OO_hisp_origin_owner   =. if g5_hisp_origin_owner_`ow' ==.
replace OO_race_amind_owner    =. if g6_race_amind_owner_`ow' ==.
replace OO_race_asian_owner    =. if g6_race_asian_owner_`ow' ==.
replace OO_race_black_owner    =. if g6_race_black_owner_`ow' ==.
replace OO_race_nathaw_owner   =. if g6_race_nathaw_owner_`ow' ==.
replace OO_race_other_owner    =. if g6_race_other_owner_`ow' ==.
replace OO_race_white_owner    =. if g6_race_white_owner_`ow' ==.
replace OO_native_born_owner   =. if g7_native_born_owner_`ow' ==.
replace OO_us_cit_owner        =. if g8_us_cit_owner_`ow' ==.
replace OO_education_owner         =. if g9_education_owner_`ow' ==.
replace md_education_owner         =. if g9_education_owner_`ow' ==.
replace OO_D_education_owner   =. if g9_education_owner_`ow' ==.
replace OO_gender_owner        =. if g10_gender_owner_`ow' ==.
}

/* Recode legitimate (hard) missing values */
replace      OO_emp_owner           = .a if   classf   <6
replace      OO_hours_owner         = .a if   classf   <6
replace      OO_work_exp_owner      = .a if   classf   <6
replace      OO_oth_bus_owner       = .a if   classf   <6
replace      OO_bus_same_ind_owner  = .a if   classf   <6
replace      OO_age_owner           = .a if   classf   <6
```

```
replace      OO_hisp_origin_owner   = .a if    classf   <6
replace      OO_race_amind_owner    = .a if    classf   <6
replace      OO_race_asian_owner    = .a if    classf   <6
replace      OO_race_black_owner    = .a if    classf   <6
replace      OO_race_nathaw_owner   = .a if    classf   <6
replace      OO_race_other_owner    = .a if    classf   <6
replace      OO_race_white_owner    = .a if    classf   <6
replace      OO_native_born_owner   = .a if    classf   <6
replace      OO_us_cit_owner        = .a if    classf   <6
replace      OO_education_owner     = .a if    classf   <6
replace      md_education_owner     = .a if    classf   <6
replace      OO_D_education_owner   = .a if    classf   <6

replace      OO_gender_owner        = .a if    classf   <6

/*****************************************************************************/
*Diversity / Similarity index
gen xr1 =    OO_race_amind_owner        *     OO_race_amind_owner
gen xr2 =    OO_race_asian_owner        *     OO_race_asian_owner
gen xr3 =    OO_race_black_owner        *     OO_race_black_owner
gen xr4 =    OO_race_nathaw_owner *    OO_race_nathaw_owner
gen xr5 =    OO_race_other_owner  *    OO_race_other_owner
gen xr6 =    OO_race_white_owner        *     OO_race_white_owner
* Race_similarity
egen Race_similarity =rowtotal(xr1  xr2  xr3  xr4  xr5  xr6 ),missing
drop xr*
* Race_ diversity
gen Race_diversity =1-Race_similarity

foreach ow in $owners_1_15  {
/* Recode to soft missing value if any of the var's component is soft missing*/
replace Race_similarity    =. if g6_race_amind_owner_`ow' ==.
replace Race_similarity    =. if g6_race_asian_owner_`ow' ==.
replace Race_similarity    =. if g6_race_black_owner_`ow' ==.
replace Race_similarity    =. if g6_race_nathaw_owner_`ow' ==.
replace Race_similarity    =. if g6_race_other_owner_`ow' ==.
replace Race_similarity    =. if g6_race_white_owner_`ow' ==.
}

/* Recode legitimate (hard) missing values */
replace Race_similarity    =.a if classf <6
replace Race_diversity    =.a if Race_similarity ==.a
```

```
replace Race_diversity    =.  if Race_similarity ==.

gen fmal =1-OO_gender_owner
gen xr1 =    OO_gender_owner     *      OO_gender_owner
gen xr2 =    fmal   *      fmal

* Gender_similarity
egen Gender_similarity =rowtotal(xr1  xr2   ),missing
drop xr*  fmal
* Gender_diversity
gen Gender_diversity =1-Gender_similarity

foreach ow in $owners_1_15  {
/* Recode to soft missing value if any of the var's component is soft missing*/
replace  Gender_similarity       =. if g10_gender_owner_`ow' ==.
}

/* Recode legitimate (hard) missing values */
replace Gender_similarity    =.a if classf <6
replace Gender_diversity    =.a if Gender_similarity ==.a
replace Gender_diversity    =.  if Gender_similarity ==.


*Business level Characteristics
*Home Based Dummy
recode c8_primary_loc  (1=1 "Home Based") (nonmiss=0 "Non Home Based" ) ,into (Home_Based )
*Sole_Proprietorship Dummy
recode c1z2_legal_status  (1=1 "Sole_Proprietorship") (nonmiss=0 "Limited Liability" ) ,into (Sole_Proprietorship )
rename d2_comp_advantage  Comp_advantage
egen Have_IP =anymatch(d3_a_have_patent  d3_b_have_copyright  d3_c_have_trademark ), values(1)
rename c5_num_employees      Full_Part_Time_Employees
rename c6_num_ft_employees   Full_Time_Employees
rename c7_num_pt_employees   Part_Time_Employees
egen   Employee_Owner           =rowtotal(g1a_emp_owner_* )
egen   Total_Employees =rowtotal(Employee_Owner  Full_Part_Time_Employees )

foreach ow in $owners_1_15  {
/* Recode to soft missing value if any of the var's component is soft missing*/
replace Have_IP =. if d3_a_have_patent ==. | d3_b_have_copyright ==. | d3_c_have_trademark ==.
replace Employee_Owner           =. if g1a_emp_owner_`ow' ==.
replace Total_Employees                =. if g1a_emp_owner_`ow' ==.
replace Total_Employees                =. if  Full_Part_Time_Employees ==.
```

```
/* Recode legitimate (hard) missing values */
replace Have_IP         =.a if classf <6
replace Employee_Owner  =.a if classf <6
replace Total_Employees =.a if classf <6
}
saveold KFS8_LI_CS_L1,replace
  set logtype text , permanently
  log using KFS8_LI_CS_L1.csv, replace
  FR_Sum_L * [pw=cswgt_final]
  log close
```

3.2.2.2.7. Stata Code: Longitudinal in Long Format

The following Stata code will create the new variables using the KFS long format file "KFS8_L7_Long." The code will save the new longitudinal (n= 18286) file under the name "KFS8_L7_L1.dta."

```
clear
clear mata
clear matrix
capture log close
set more off
set maxvar 32767
program drop _all
adopath + " xxx:\KFS_Manual_and_Data\Farhat_Robb_Commands\"
cd "Xxx:\KFS_Manual_and_Data"
  use KFS8_LI_L_Long,clear

drop tot_assets tot_liab tot_equity_nonowneroperators tot_equity_nonownerop_allyrs ///
tot_pers_debt_owed_resp  tot_pers_debt_resp tot_pers_debt_owed_othrownrs ///
tot_pers_debt_other_owners tot_bus_debt_owed tot_debt_bus tot_debt_owner_operators ///
tot_debt_owed_owner_operators tot_debt tot_debt_owed tot_equity_owner_operators  ///
tot_equity_owneroper_allyrs tot_debt_liab_equity tot_equity tot_equity_allyrs


*Replacing missing continuous values by the midpoints of the class intervals

foreach totfin in tot_*_r {
recode `totfin'     (0=0)(1=250)(2=750)(3=2000)(4=4000)(5=7500) (6=17500) (7=62500) (8=550000) (9=1000000)
    }

global owners_1_15 "01 02 03 04 05 06 07 08 09 10 11 12 13 14 15"

foreach ow in $owners_1_15  {
replace  f2_owner_amt_eq_invest_`ow' =tot_equity_owner_`ow'_r    if  f2_owner_amt_eq_invest_`ow' ==.
replace  f2_ownr_amt_eqinvest_allyrs_`ow' =tot_equity_allyrs_owner_`ow'_r    if  f2_ownr_amt_eqinvest_allyrs_`ow' ==.
}

replace  f4_eq_amt_angels =tot_equity_angels_r    if  f4_eq_amt_angels ==.
replace  f4_eq_amt_angels_allyrs =tot_equity_angels_allyrs_r    if  f4_eq_amt_angels_allyrs ==.
replace  f4_eq_amt_companies =tot_equity_companies_r    if  f4_eq_amt_companies ==.
replace  f4_eq_amt_companies_allyrs =tot_equity_companies_allyrs_r    if  f4_eq_amt_companies_allyrs ==.
replace  f4_eq_amt_govt =tot_equity_govt_r    if  f4_eq_amt_govt ==.
```

```
replace  f4_eq_amt_govt_allyrs =tot_equity_govt_allyrs_r    if  f4_eq_amt_govt_allyrs ==.
replace  f4_eq_amt_other =tot_equity_other_r    if  f4_eq_amt_other ==.
replace  f4_eq_amt_other_allyrs =tot_equity_other_allyrs_r    if  f4_eq_amt_other_allyrs ==.
replace  f4_eq_amt_parents =tot_equity_parents_r    if  f4_eq_amt_parents ==.
replace  f4_eq_amt_parents_allyrs =tot_equity_parents_allyrs_r    if  f4_eq_amt_parents_allyrs ==.
replace  f4_eq_amt_spouse =tot_equity_spouse_r    if  f4_eq_amt_spouse ==.
replace  f4_eq_amt_spouse_allyrs =tot_equity_spouse_allyrs_r    if  f4_eq_amt_spouse_allyrs ==.
replace  f4_eq_amt_vent_cap =tot_equity_vent_cap_r    if  f4_eq_amt_vent_cap ==.
replace  f4_eq_amt_vent_cap_allyrs =tot_equity_vent_cap_allyrs_r    if  f4_eq_amt_vent_cap_allyrs ==.
replace  f6b_personal_use_amt =tot_personal_use_r    if  f6b_personal_use_amt ==.
replace  f8a_bus_credcard_line =tot_bus_credcard_line_resp_r    if  f8a_bus_credcard_line ==.
replace  f8b_bus_credcard_bal =tot_bus_credcard_bal_resp_r    if  f8b_bus_credcard_bal ==.
replace  f8a_pers_credcard_line =tot_pers_credcard_line_resp_r    if  f8a_pers_credcard_line ==.
replace  f8b_pers_credcard_bal =tot_pers_credcard_bal_resp_r    if  f8b_pers_credcard_bal ==.
replace  f8c_pers_loan_bank_amt =tot_pers_loan_bank_resp_r    if  f8c_pers_loan_bank_amt ==.
replace  f8d_pers_loan_bank_owed =tot_pers_loan_bank_owed_resp_r    if  f8d_pers_loan_bank_owed ==.
replace  f8c_pers_loan_fam_amt =tot_pers_loan_fam_resp_r    if  f8c_pers_loan_fam_amt ==.
replace  f8d_pers_loan_fam_owed =tot_pers_loan_fam_owed_resp_r    if  f8d_pers_loan_fam_owed ==.
replace  f8c_pers_loan_other_amt =tot_pers_loan_other_resp_r    if  f8c_pers_loan_other_amt ==.
replace  f8d_pers_loan_other_owed =tot_persloan_other_owed_resp_r    if  f8d_pers_loan_other_owed ==.
replace  f8c_pers_other_amt =tot_pers_other_resp_r    if  f8c_pers_other_amt ==.
replace  f8d_pers_other_owed =tot_pers_other_owed_resp_r    if  f8d_pers_other_owed ==.
replace  f10a_bus_credcard_line =tot_bus_credcard_line_others_r    if  f10a_bus_credcard_line ==.
replace  f10b_bus_credcard_bal =tot_bus_credcard_bal_others_r    if  f10b_bus_credcard_bal ==.
replace  f10a_pers_credcard_line =tot_pers_credcard_line_othrs_r    if  f10a_pers_credcard_line ==.
replace  f10b_pers_credcard_bal =tot_pers_credcard_bal_others_r    if  f10b_pers_credcard_bal ==.
replace  f10c_pers_loan_bank_amt =tot_pers_loan_bank_others_r    if  f10c_pers_loan_bank_amt ==.
replace  f10d_pers_loan_bank_owed =tot_persloan_bank_owed_othrs_r    if  f10d_pers_loan_bank_owed ==.
replace  f10c_pers_loan_fam_amt =tot_persloan_fam_othrowners_r    if  f10c_pers_loan_fam_amt ==.
replace  f10d_pers_loan_fam_owed =tot_persloan_fam_owed_othrs_r    if  f10d_pers_loan_fam_owed ==.
replace  f10c_pers_loan_other_amt =tot_pers_loan_other_owners_r    if  f10c_pers_loan_other_amt ==.
replace  f10d_pers_loan_other_owed =tot_persloan_othr_owed_othrs_r    if  f10d_pers_loan_other_owed ==.
replace  f10c_pers_other_amt =tot_pers_other_other_owners_r    if  f10c_pers_other_amt ==.
replace  f10d_pers_other_owed =tot_pers_other_owed_others_r    if  f10d_pers_other_owed ==.
replace  f12a_bus_cred_line =tot_cred_line_bus_line_r    if  f12a_bus_cred_line ==.
replace  f12b_bus_cred_line_bal =tot_cred_line_bus_bal_r    if  f12b_bus_cred_line_bal ==.
replace  f12a_bus_credcard_line =tot_credcard_line_bus_r    if  f12a_bus_credcard_line ==.
replace  f12b_bus_credcard_bal =tot_credcard_bal_bus_r    if  f12b_bus_credcard_bal ==.
replace  f12c_bus_loans_bank_amt =tot_loan_bank_bus_r    if  f12c_bus_loans_bank_amt ==.
replace  f12d_bus_loans_bank_owed =tot_bus_loans_bank_owed_r    if  f12d_bus_loans_bank_owed ==.
replace  f12c_bus_loans_nonbank_amt =tot_loan_nonbank_bus_r    if  f12c_bus_loans_nonbank_amt ==.
replace  f12d_bus_loans_nonbank_owed =tot_bus_loans_nonbank_owed_r    if  f12d_bus_loans_nonbank_owed ==.
```

```
replace  f12c_bus_loans_emp_amt =tot_loan_emp_bus_r    if  f12c_bus_loans_emp_amt ==.
replace  f12d_bus_loans_emp_owed =tot_bus_loans_emp_owed_r    if  f12d_bus_loans_emp_owed ==.
replace  f12c_bus_loans_fam_amt =tot_loan_fam_bus_r    if  f12c_bus_loans_fam_amt ==.
replace  f12d_bus_loans_fam_owed =tot_bus_loans_fam_owed_r    if  f12d_bus_loans_fam_owed ==.
replace  f12c_bus_loans_govt_amt =tot_loan_govt_bus_r    if  f12c_bus_loans_govt_amt ==.
replace  f12d_bus_loans_govt_owed =tot_bus_loans_govt_owed_r    if  f12d_bus_loans_govt_owed ==.
replace  f12c_bus_loans_other_ind_amt =tot_loan_other_ind_r    if  f12c_bus_loans_other_ind_amt ==.
replace  f12d_bus_loans_other_ind_owed =tot_bus_loans_otherind_owed_r    if  f12d_bus_loans_other_ind_owed ==.
replace  f12c_bus_loans_owner_amt =tot_loan_owner_bus_r    if  f12c_bus_loans_owner_amt ==.
replace  f12d_bus_loans_owner_owed =tot_bus_loans_owner_owed_r    if  f12d_bus_loans_owner_owed ==.
replace  f12c_bus_loans_bus_amt =tot_loan_other_bus_r    if  f12c_bus_loans_bus_amt ==.
replace  f12d_bus_loans_bus_owed =tot_bus_loans_otherbus_owed_r    if  f12d_bus_loans_bus_owed ==.
replace  f12c_bus_other_amt =tot_bus_debt_other_r    if  f12c_bus_other_amt ==.
replace  f12d_bus_other_owed =tot_bus_loans_other_owed_r    if  f12d_bus_other_owed ==.
replace  f14a_trade_fin_amt =tot_trade_finan_r    if  f14a_trade_fin_amt ==.
replace  f16a_rev_amt =tot_revenue_r    if  f16a_rev_amt ==.
replace  f17a_total_exp_amt =tot_expenses_r    if  f17a_total_exp_amt ==.
replace  f18a_wage_exp_amt =tot_wages_r    if  f18a_wage_exp_amt ==.
replace  f19a_res_dev_amt =tot_res_dev_r    if  f19a_res_dev_amt ==.
replace  f19c_a_design_amt =tot_intangassets_design_r    if  f19c_a_design_amt ==.
replace  f19c_b_investments_amt =tot_intangassets_invest_r    if  f19c_b_investments_amt ==.
replace  f19c_c_brand_dev_amt =tot_intangassets_branddev_r    if  f19c_c_brand_dev_amt ==.
replace  f19c_d_org_dev_amt =tot_intangassets_orgdev_r    if  f19c_d_org_dev_amt ==.
replace  f19c_e_worker_training_amt =tot_intangassets_wkrtrng_r    if  f19c_e_worker_training_amt ==.
replace  f19c_f_other_amt =tot_intangassets_other_r    if  f19c_f_other_amt ==.
replace  f19c_intangassets_amt =tot_intang_assets_r    if  f19c_intangassets_amt ==.
replace  f24_profit_amt =tot_profit_r    if  f24_profit_amt ==.
replace  f26_loss_amt =tot_loss_r    if  f26_loss_amt ==.
replace  f29_assetval_cash =tot_asset_cash_r    if  f29_assetval_cash ==.
replace  f29_assetval_acctrec =tot_asset_acct_rec_r    if  f29_assetval_acctrec ==.
replace  f29_assetval_inv =tot_asset_inv_r    if  f29_assetval_inv ==.
replace  f29_assetval_equip =tot_asset_equip_r    if  f29_assetval_equip ==.
replace  f29_assetval_landbuild =tot_asset_landbuild_r    if  f29_assetval_landbuild ==.
replace  f29_assetval_veh =tot_asset_veh_r    if  f29_assetval_veh ==.
replace  f29_assetval_othbusprop =tot_asset_other_bus_prop_r    if  f29_assetval_othbusprop ==.
replace  f29_assetval_other =tot_asset_other_r    if  f29_assetval_other ==.
replace  f31_value_acctpay =tot_liab_acct_pay_r    if  f31_value_acctpay ==.
replace  f31_value_pension =tot_liab_pension_r    if  f31_value_pension ==.
replace  f31_value_other =tot_liab_other_r    if  f31_value_other ==.

egen   Tot_Equity_Owner_Operators =     rowtotal(    f2_owner_amt_eq_invest_*   )      , missing
egen   Tot_Equity_OwnerOper_AllYrs =rowtotal(  f2_ownr_amt_eqinvest_allyrs_*    ), missing
```

```
/* Recode legitimate (hard) missing values */
replace Tot_Equity_Owner_Operators  =.a if classf <6
replace Tot_Equity_OwnerOper_AllYrs =.a if classf <6

foreach ow in $owners_1_15  {
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Equity_Owner_Operators  =.   if  f2_owner_amt_eq_invest_`ow' ==.
replace Tot_Equity_OwnerOper_AllYrs =.   if  f2_ownr_amt_eqinvest_allyrs_`ow' ==.
}

global List1 "spouse parents angels companies govt vent_cap other"

egen  Tot_Equity_NonOwnerOperators =rowtotal(f4_eq_amt_angels  f4_eq_amt_companies  f4_eq_amt_govt  ///
f4_eq_amt_other  f4_eq_amt_parents  f4_eq_amt_spouse   f4_eq_amt_vent_cap )     , missing
egen  Tot_Equity_NonOwnerOp_AllYrs =rowtotal(f4_eq_amt_*_allyrs )  , missing
/* Recode legitimate (hard) missing values */
replace Tot_Equity_NonOwnerOperators  =.a if classf <6
replace Tot_Equity_NonOwnerOperators  =.a if  c1z2_legal_status ==1
replace Tot_Equity_NonOwnerOperators  =.a if  one_owner ==1
replace Tot_Equity_NonOwnerOp_AllYrs  =.a if classf <6
replace Tot_Equity_NonOwnerOp_AllYrs  =.a if  c1z2_legal_status ==1
replace Tot_Equity_NonOwnerOp_AllYrs  =.a if  one_owner ==1

foreach name in $List1  {
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Equity_NonOwnerOperators  =. if f4_eq_amt_`name' ==.
replace Tot_Equity_NonOwnerOp_AllYrs  =. if f4_eq_amt_`name'_allyrs ==.
}

egen  Tot_Equity        =rowtotal(Tot_Equity_Owner_Operators  Tot_Equity_NonOwnerOperators ), missing
egen  Tot_Equity_AllYrs =rowtotal(Tot_Equity_OwnerOper_AllYrs  Tot_Equity_NonOwnerOp_AllYrs ), missing
/* Recode legitimate (hard) missing values */
replace Tot_Equity        =.a if classf <6
replace Tot_Equity_AllYrs =.a if classf <6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Equity        =. if Tot_Equity_Owner_Operators ==.
replace Tot_Equity        =. if Tot_Equity_NonOwnerOperators ==.
replace Tot_Equity_AllYrs =. if Tot_Equity_NonOwnerOperators ==.
replace Tot_Equity_AllYrs =. if Tot_Equity_NonOwnerOp_AllYrs ==.

egen  Tot_Assets =rowtotal(f29_assetval_* )    , missing
```

```
egen  Tot_Liab =rowtotal(f31_value_* ), missing
/* Recode legitimate (hard) missing values */
replace Tot_Assets  =.a if classf <6
replace Tot_Liab    =.a if classf <6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Assets  =. if f29_assetval_acctrec ==.
replace Tot_Assets  =. if f29_assetval_cash ==.
replace Tot_Assets  =. if f29_assetval_equip ==.
replace Tot_Assets  =. if f29_assetval_inv ==.
replace Tot_Assets  =. if f29_assetval_landbuild ==.
replace Tot_Assets  =. if f29_assetval_othbusprop ==.
replace Tot_Assets  =. if f29_assetval_other ==.
replace Tot_Assets  =. if f29_assetval_veh ==.
replace Tot_Liab  =. if  f31_value_acctpay ==.
replace Tot_Liab  =. if  f31_value_other ==.
replace Tot_Liab  =. if  f31_value_pension ==.

egen       Tot_Pers_Debt_Resp    =rowtotal(f8b_pers_credcard_bal       f8b_bus_credcard_bal       f8c_pers_loan_bank_amt
f8c_pers_loan_fam_amt  f8c_pers_loan_other_amt  f8c_pers_other_amt ),missing
egen     Tot_Pers_Debt_Owed_Resp   =rowtotal(f8b_pers_credcard_bal     f8b_bus_credcard_bal     f8d_pers_loan_bank_owed
f8d_pers_loan_fam_owed  f8d_pers_loan_other_owed  f8d_pers_other_owed ),missing
/* Recode legitimate (hard) missing values */
replace Tot_Pers_Debt_Resp       =.a if classf <6
replace Tot_Pers_Debt_Owed_Resp  =.a if classf <6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Pers_Debt_Resp  =. if f8b_pers_credcard_bal ==.
replace Tot_Pers_Debt_Resp  =. if f8b_bus_credcard_bal ==.
replace Tot_Pers_Debt_Resp  =. if f8c_pers_loan_bank_amt ==.
replace Tot_Pers_Debt_Resp  =. if f8c_pers_loan_fam_amt ==.
replace Tot_Pers_Debt_Resp  =. if f8c_pers_loan_other_amt ==.
replace Tot_Pers_Debt_Resp  =. if f8c_pers_other_amt ==.
replace Tot_Pers_Debt_Owed_Resp  =. if f8b_pers_credcard_bal ==.
replace Tot_Pers_Debt_Owed_Resp  =. if f8b_bus_credcard_bal ==.
replace Tot_Pers_Debt_Owed_Resp  =. if f8d_pers_loan_bank_owed ==.
replace Tot_Pers_Debt_Owed_Resp  =. if f8d_pers_loan_fam_owed ==.
replace Tot_Pers_Debt_Owed_Resp  =. if f8d_pers_loan_other_owed ==.
replace Tot_Pers_Debt_Owed_Resp  =. if f8d_pers_other_owed ==.

egen    Tot_Pers_Debt_Other_Owners  =rowtotal(f10b_pers_credcard_bal    f10c_pers_loan_bank_amt    f10b_bus_credcard_bal
f10c_pers_loan_fam_amt  f10c_pers_loan_other_amt   f10c_pers_other_amt ),missing
egen   Tot_Pers_Debt_Owed_OthrOwnrs  =rowtotal(f10b_pers_credcard_bal   f10b_bus_credcard_bal   f10d_pers_loan_bank_owed
f10d_pers_loan_fam_owed  f10d_pers_loan_other_owed  f10d_pers_other_owed ),missing
```

```
/* Recode legitimate (hard) missing values */
replace Tot_Pers_Debt_Other_Owners   =.a if classf <6
replace Tot_Pers_Debt_Owed_OthrOwnrs =.a if classf <6
replace Tot_Pers_Debt_Other_Owners   =.a if c4_numowners_confirm <2
replace Tot_Pers_Debt_Owed_OthrOwnrs =.a if c4_numowners_confirm <2
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Pers_Debt_Other_Owners  =. if f10b_pers_credcard_bal ==.
replace Tot_Pers_Debt_Other_Owners  =. if f10b_bus_credcard_bal ==.
replace Tot_Pers_Debt_Other_Owners  =. if f10c_pers_loan_bank_amt ==.
replace Tot_Pers_Debt_Other_Owners  =. if f10c_pers_loan_fam_amt ==.
replace Tot_Pers_Debt_Other_Owners  =. if f10c_pers_loan_other_amt ==.
replace Tot_Pers_Debt_Other_Owners  =. if f10c_pers_other_amt ==.
replace Tot_Pers_Debt_Owed_OthrOwnrs  =. if f10b_pers_credcard_bal ==.
replace Tot_Pers_Debt_Owed_OthrOwnrs  =. if f10b_bus_credcard_bal ==.
replace Tot_Pers_Debt_Owed_OthrOwnrs  =. if f10d_pers_loan_bank_owed ==.
replace Tot_Pers_Debt_Owed_OthrOwnrs  =. if f10d_pers_loan_fam_owed ==.
replace Tot_Pers_Debt_Owed_OthrOwnrs  =. if f10d_pers_loan_other_owed ==.
replace Tot_Pers_Debt_Owed_OthrOwnrs  =. if f10d_pers_other_owed ==.

egen  Tot_Debt_Owner_Operators =rowtotal(Tot_Pers_Debt_Resp  Tot_Pers_Debt_Other_Owners ),missing
egen  Tot_Debt_Owed_Owner_Operators =rowtotal(Tot_Pers_Debt_Owed_Resp  Tot_Pers_Debt_Owed_OthrOwnrs ),missing
/* Recode legitimate (hard) missing values */
replace Tot_Debt_Owner_Operators    =.a if classf <6
replace Tot_Debt_Owed_Owner_Operators =.a if classf <6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Debt_Owner_Operators    =. if Tot_Pers_Debt_Resp ==.
replace Tot_Debt_Owner_Operators    =. if Tot_Pers_Debt_Other_Owners ==.
replace Tot_Debt_Owed_Owner_Operators =. if Tot_Pers_Debt_Owed_Resp ==.
replace Tot_Debt_Owed_Owner_Operators =. if Tot_Pers_Debt_Owed_OthrOwnrs ==.

egen     Tot_Debt_Bus     =rowtotal(f12b_bus_credcard_bal          f12c_bus_loans_bank_amt          f12b_bus_cred_line_bal
f12c_bus_loans_nonbank_amt       f12c_bus_loans_fam_amt          f12c_bus_loans_govt_amt          f12c_bus_loans_emp_amt
f12c_bus_loans_other_ind_amt  f12c_bus_loans_owner_amt  f12c_bus_loans_bus_amt  f12c_bus_other_amt ),missing
egen   Tot_Bus_Debt_Owed   =rowtotal(f12b_bus_cred_line_bal          f12b_bus_credcard_bal      f12d_bus_loans_bank_owed
f12d_bus_loans_nonbank_owed      f12d_bus_loans_emp_owed          f12d_bus_loans_fam_owed      f12d_bus_loans_govt_owed
f12d_bus_loans_other_ind_owed  f12d_bus_loans_owner_owed   f12d_bus_loans_bus_owed  f12d_bus_other_owed ),missing
/* Recode legitimate (hard) missing values */
replace Tot_Debt_Bus      =.a if classf <6
replace Tot_Bus_Debt_Owed =.a if classf <6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Debt_Bus    =. if f12b_bus_credcard_bal ==.
replace Tot_Debt_Bus    =. if f12b_bus_cred_line_bal ==.
```

```
replace Tot_Debt_Bus    =. if f12c_bus_loans_bank_amt ==.
replace Tot_Debt_Bus    =. if f12c_bus_loans_nonbank_amt ==.
replace Tot_Debt_Bus    =. if f12c_bus_loans_fam_amt ==.
replace Tot_Debt_Bus    =. if f12c_bus_loans_govt_amt ==.
replace Tot_Debt_Bus    =. if f12c_bus_loans_emp_amt ==.
replace Tot_Debt_Bus    =. if f12c_bus_loans_other_ind_amt ==.
replace Tot_Debt_Bus    =. if f12c_bus_loans_owner_amt  ==.
replace Tot_Debt_Bus    =. if f12c_bus_loans_bus_amt ==.
replace Tot_Debt_Bus    =. if f12c_bus_other_amt ==.
replace Tot_Bus_Debt_Owed    =. if f12b_bus_credcard_bal ==.
replace Tot_Bus_Debt_Owed    =. if f12b_bus_cred_line_bal ==.
replace Tot_Bus_Debt_Owed    =. if f12d_bus_loans_bank_owed ==.
replace Tot_Bus_Debt_Owed    =. if f12d_bus_loans_nonbank_owed ==.
replace Tot_Bus_Debt_Owed    =. if f12d_bus_loans_fam_owed ==.
replace Tot_Bus_Debt_Owed    =. if f12d_bus_loans_govt_owed ==.
replace Tot_Bus_Debt_Owed    =. if f12d_bus_loans_emp_owed ==.
replace Tot_Bus_Debt_Owed    =. if f12d_bus_loans_other_ind_owed ==.
replace Tot_Bus_Debt_Owed    =. if f12d_bus_loans_owner_owed  ==.
replace Tot_Bus_Debt_Owed    =. if f12d_bus_loans_bus_owed  ==.
replace Tot_Bus_Debt_Owed    =. if f12d_bus_other_owed ==.

egen  Tot_Debt =rowtotal(Tot_Debt_Owner_Operators  Tot_Debt_Bus ),missing
egen  Tot_Debt_Owed =rowtotal(Tot_Debt_Owed_Owner_Operators  Tot_Bus_Debt_Owed ),missing
/* Recode legitimate (hard) missing values */
replace Tot_Debt       =.a if classf <6
replace Tot_Debt_Owed  =.a if classf <6
/* Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Debt    =. if Tot_Debt_Owner_Operators ==.
replace Tot_Debt    =. if Tot_Debt_Bus ==.
replace Tot_Debt_Owed =. if Tot_Debt_Owed_Owner_Operators ==.
replace Tot_Debt_Owed =. if Tot_Bus_Debt_Owed ==.

gen  Net_Profit =f24_profitloss_amt

/*****************************************************************/
/* Merge the file with the primary owner file "primary_owner.dta" */

merge m:1 mprid using "Xxx:\KFS_Manual_and_Data\primary_owner.dta"
drop _merge

*Replacing missing continuous values by the midpoints of the class intervals
```

```
foreach ow in $owners_1_15  {
recode       total_hours_owner_`ow'_r   (0=0) (1=9.5)(2=27.5)(3=40.5)(4=50.5)(5=60.5) (6=70.5)
recode       age_owner_`ow'_r    (1=21)(2=29.5)(3=39.5)(4=49.5)(5=59.5) (6=69.5) (7=79.5)
    }

foreach ow in $owners_1_15  {
replace  g1b1_hours_owner_`ow' =total_hours_owner_`ow'_r    if  g1b1_hours_owner_`ow' ==.
replace  g4_age_owner_`ow' =age_owner_`ow'_r    if  g4_age_owner_`ow' ==.
}

drop *_r

* Primary Owner(PO) Characteristics at the Baseline: PO by Robb's Stata code

gen  PO_emp                    =.
gen  PO_hours                  =.
gen  PO_work_exp           =.
gen  PO_oth_bus_owner       =.
gen  PO_bus_same_ind        =.
gen  PO_age_owner          =.
gen  PO_hisp_origin           =.
gen  PO_race_group            =.
gen  PO_race_amind_owner    =.
gen  PO_race_asian_owner    =.
gen  PO_race_black_owner    =.
gen  PO_race_nathaw_owner   =.
gen  PO_race_other_owner    =.
gen  PO_race_white_owner    =.
gen  PO_native_born           =.
gen  PO_us_cit                =.
gen  PO_education          =.
gen  PO_gender                 =.

sort mprid year
xtset mprid year

forvalues po = 1/6 {
bysort mprid (year):replace PO_emp           =   g1a_emp_owner_0`po'[1]              if  primary_owner==`po'
bysort mprid (year):replace PO_hours        =  g1b1_hours_owner_0`po'[1]                 if  primary_owner==`po'
bysort mprid (year):replace PO_work_exp       =     g2_work_exp_owner_0`po'[1]              if  primary_owner==`po'
bysort mprid (year):replace PO_oth_bus_owner = g3a_oth_bus_owner_0`po'[1]             if  primary_owner==`po'
bysort mprid (year):replace PO_bus_same_ind  = g3b_bus_same_ind_owner_0`po'[1]  if  primary_owner==`po'
```

```
bysort mprid (year):replace PO_age_owner       =      g4_age_owner_0`po'[1]                    if  primary_owner==`po'
bysort mprid (year):replace PO_hisp_origin     =      g5_hisp_origin_owner_0`po'[1]          if  primary_owner==`po'
bysort mprid (year):replace PO_race_group    = g6b_race_group_0`po'[1]                  if  primary_owner==`po'
bysort mprid (year):replace PO_race_amind_owner  = g6_race_amind_owner_0`po'[1]       if  primary_owner==`po'
bysort mprid (year):replace PO_race_asian_owner  = g6_race_asian_owner_0`po'[1]       if  primary_owner==`po'
bysort mprid (year):replace PO_race_black_owner  = g6_race_black_owner_0`po'[1]       if  primary_owner==`po'
bysort mprid (year):replace PO_race_nathaw_owner = g6_race_nathaw_owner_0`po'[1]      if  primary_owner==`po'
bysort mprid (year):replace PO_race_other_owner  = g6_race_other_owner_0`po'[1]       if  primary_owner==`po'
bysort mprid (year):replace PO_race_white_owner  = g6_race_white_owner_0`po'[1]       if  primary_owner==`po'
bysort mprid (year):replace PO_native_born     =      g7_native_born_owner_0`po'[1]          if  primary_owner==`po'
bysort mprid (year):replace PO_us_cit       =  g8_us_cit_owner_0`po'[1]                if  primary_owner==`po'
bysort mprid (year):replace PO_education       =      g9_education_owner_0`po'[1]              if  primary_owner==`po'
bysort mprid (year):replace PO_gender       =  g10_gender_owner_0`po'[1]               if  primary_owner==`po'
}

replace  PO_emp                = .a if  classf   <6
replace  PO_hours            = .a if    classf   <6
replace  PO_work_exp           = .a if  classf   <6
replace  PO_oth_bus_owner          = .a if     classf   <6
replace  PO_bus_same_ind     = .a if    classf   <6
replace  PO_age_owner          = .a if  classf   <6
replace  PO_hisp_origin     = .a if     classf   <6
replace  PO_race_group      = .a if     classf   <6
replace  PO_race_amind_owner  = .a if   classf   <6
replace  PO_race_asian_owner  = .a if   classf   <6
replace  PO_race_black_owner  = .a if   classf   <6
replace  PO_race_nathaw_owner = .a if   classf   <6
replace  PO_race_other_owner  = .a if   classf   <6
replace  PO_race_white_owner  = .a if   classf   <6
replace  PO_native_born     = .a if     classf   <6
replace  PO_us_cit             = .a if  classf   <6
replace  PO_education          = .a if  classf   <6
replace  PO_gender             = .a if  classf   <6

* Active-Owner-Operators Characteristics (OO)
egen   OO_emp_owner           =   rowmean(g1a_emp_owner_* )
egen   OO_hours_owner         = rowmean(g1b1_hours_owner_* )
egen   OO_work_exp_owner      =   rowmean(g2_work_exp_owner_*       )
egen   OO_oth_bus_owner       =   rowmean(g3a_oth_bus_owner_*       )
egen   OO_bus_same_ind_owner  =   rowmean(g3b_bus_same_ind_owner_* )
egen   OO_age_owner           =   rowmean(g4_age_owner_* )
egen   OO_hisp_origin_owner   =   rowmean(g5_hisp_origin_owner_* )
```

```
egen   OO_race_amind_owner    =   rowmean(g6_race_amind_owner_* )
egen   OO_race_asian_owner    =   rowmean(g6_race_asian_owner_* )
egen   OO_race_black_owner    =   rowmean(g6_race_black_owner_* )
egen   OO_race_nathaw_owner   =   rowmean(g6_race_nathaw_owner_* )
egen   OO_race_other_owner    =   rowmean(g6_race_other_owner_* )
egen   OO_race_white_owner    =   rowmean(g6_race_white_owner_* )
egen   OO_native_born_owner   =   rowmean(g7_native_born_owner_*   )
egen   OO_us_cit_owner        =   rowmean(g8_us_cit_owner_*  )
egen   OO_education_owner     =   rowmean(g9_education_owner_*     )
egen   md_education_owner     =   rowmedian(g9_education_owner_* )
gen    OO_D_education_owner   =(md_education_owner     >6.99)        if      md_education_owner     <11

egen   OO_gender_owner        =   rowmean(g10_gender_owner_*       )

/* Recode to soft missing value if any of the total's component is soft missing*/
foreach ow in $owners_1_15  {
replace OO_emp_owner             =. if g1a_emp_owner_`ow' ==.
replace OO_hours_owner              =. if g1b1_hours_owner_`ow' ==.
replace OO_work_exp_owner      =. if g2_work_exp_owner_`ow' ==.
replace OO_oth_bus_owner       =. if g3a_oth_bus_owner_`ow' ==.
replace OO_bus_same_ind_owner  =. if g3b_bus_same_ind_owner_`ow' ==.
replace OO_age_owner             =. if g4_age_owner_`ow' ==.
replace OO_hisp_origin_owner   =. if g5_hisp_origin_owner_`ow' ==.
replace OO_race_amind_owner    =. if g6_race_amind_owner_`ow' ==.
replace OO_race_asian_owner    =. if g6_race_asian_owner_`ow' ==.
replace OO_race_black_owner    =. if g6_race_black_owner_`ow' ==.
replace OO_race_nathaw_owner   =. if g6_race_nathaw_owner_`ow' ==.
replace OO_race_other_owner    =. if g6_race_other_owner_`ow' ==.
replace OO_race_white_owner    =. if g6_race_white_owner_`ow' ==.
replace OO_native_born_owner   =. if g7_native_born_owner_`ow' ==.
replace OO_us_cit_owner        =. if g8_us_cit_owner_`ow' ==.
replace OO_education_owner         =. if g9_education_owner_`ow' ==.
replace md_education_owner         =. if g9_education_owner_`ow' ==.
replace OO_D_education_owner   =. if g9_education_owner_`ow' ==.
replace OO_gender_owner        =. if g10_gender_owner_`ow' ==.
}

/* Recode legitimate (hard) missing values */
replace      OO_emp_owner           = .a if   classf   <6
replace      OO_hours_owner         = .a if   classf   <6
replace      OO_work_exp_owner      = .a if   classf   <6
replace      OO_oth_bus_owner       = .a if   classf   <6
```

```
replace      OO_bus_same_ind_owner  = .a if   classf   <6
replace      OO_age_owner           = .a if   classf   <6
replace      OO_hisp_origin_owner   = .a if   classf   <6
replace      OO_race_amind_owner    = .a if   classf   <6
replace      OO_race_asian_owner    = .a if   classf   <6
replace      OO_race_black_owner    = .a if   classf   <6
replace      OO_race_nathaw_owner   = .a if   classf   <6
replace      OO_race_other_owner    = .a if   classf   <6
replace      OO_race_white_owner    = .a if   classf   <6
replace      OO_native_born_owner   = .a if   classf   <6
replace      OO_us_cit_owner        = .a if   classf   <6
replace      OO_education_owner     = .a if   classf   <6
replace      md_education_owner     = .a if   classf   <6
replace      OO_D_education_owner   = .a if   classf   <6

replace      OO_gender_owner        = .a if   classf   <6
/*****************************************************************************/
*Diversity / Similarity index
gen xr1 =    OO_race_amind_owner        *     OO_race_amind_owner
gen xr2 =    OO_race_asian_owner        *     OO_race_asian_owner
gen xr3 =    OO_race_black_owner        *     OO_race_black_owner
gen xr4 =    OO_race_nathaw_owner *    OO_race_nathaw_owner
gen xr5 =    OO_race_other_owner  *    OO_race_other_owner
gen xr6 =    OO_race_white_owner        *     OO_race_white_owner
* Race_similarity
egen Race_similarity =rowtotal(xr1  xr2  xr3  xr4  xr5  xr6 ),missing
drop xr*
* Race_ diversity
gen Race_diversity =1-Race_similarity
foreach ow in $owners_1_15  {
/* Recode to soft missing value if any of the var's component is soft missing*/
replace Race_similarity    =. if g6_race_amind_owner_`ow' ==.
replace Race_similarity    =. if g6_race_asian_owner_`ow' ==.
replace Race_similarity    =. if g6_race_black_owner_`ow' ==.
replace Race_similarity    =. if g6_race_nathaw_owner_`ow' ==.
replace Race_similarity    =. if g6_race_other_owner_`ow' ==.
replace Race_similarity    =. if g6_race_white_owner_`ow' ==.
}

/* Recode legitimate (hard) missing values */
replace Race_similarity    =.a if classf <6
replace Race_diversity    =.a if Race_similarity ==.a
```

```
replace Race_diversity    =.  if Race_similarity ==.

gen fmal =1-OO_gender_owner
gen xr1 =    OO_gender_owner     *      OO_gender_owner
gen xr2 =    fmal   *      fmal

* Gender_similarity
egen Gender_similarity =rowtotal(xr1  xr2   ),missing
drop xr*  fmal
* Gender_diversity
gen Gender_diversity =1-Gender_similarity

foreach ow in $owners_1_15  {
/* Recode to soft missing value if any of the var's component is soft missing*/
replace  Gender_similarity       =. if g10_gender_owner_`ow' ==.
}

/* Recode legitimate (hard) missing values */
replace Gender_similarity    =.a if classf <6
replace Gender_diversity    =.a if Gender_similarity ==.a
replace Gender_diversity    =.  if Gender_similarity ==.

*Business level Characteristics
*Home Based Dummy
recode c8_primary_loc  (1=1 "Home Based") (nonmiss=0 "Non Home Based" ) ,into (Home_Based )
*Sole_Proprietorship Dummy
recode c1z2_legal_status  (1=1 "Sole_Proprietorship") (nonmiss=0 "Limited Liability" ) ,into (Sole_Proprietorship )
rename d2_comp_advantage  Comp_advantage
egen Have_IP =anymatch(d3_a_have_patent  d3_b_have_copyright  d3_c_have_trademark ), values(1)
rename c5_num_employees      Full_Part_Time_Employees
rename c6_num_ft_employees   Full_Time_Employees
rename c7_num_pt_employees   Part_Time_Employees
egen   Employee_Owner           =rowtotal(g1a_emp_owner_* )
egen   Total_Employees =rowtotal(Employee_Owner  Full_Part_Time_Employees )

foreach ow in $owners_1_15  {
/* Recode to soft missing value if any of the var's component is soft missing*/
replace Have_IP =. if d3_a_have_patent ==. | d3_b_have_copyright ==. | d3_c_have_trademark ==.
replace Employee_Owner           =. if g1a_emp_owner_`ow' ==.
replace Total_Employees                =. if g1a_emp_owner_`ow' ==.
replace Total_Employees                =. if  Full_Part_Time_Employees ==.
/* Recode legitimate (hard) missing values */
```

```
replace Have_IP         =.a if classf <6
replace Employee_Owner  =.a if classf <6
replace Total_Employees =.a if classf <6
}

drop cswgt_final wgt_1_long wgt_2_long wgt_3_long wgt_4_long wgt_5_long wgt_6_long

saveold KFS8_LI_L_L1,replace
  set logtype text , permanently
  log using KFS8_LI_L_L1.csv, replace
  FR_Sum_L * [pw=wgt_7_long]
  log close
```

3.2.3. The KFS Multiply Imputed Data Files

Multiple imputations involve generating "m" substitute sets for the missing values, which allows for the uncertainty due to imputation to be reflected in the analysis (Rubin, 1978, 1987). The imputed values are ideally independent draws from the predictive distribution of the missing values conditional on the observed values. For the KFS multiply imputed data, there are five completed data sets (m=5).

The multiple imputations for the KFS fixed core set of questions (asked by all businesses in every survey) were created using sequential regression multivariate imputation (SRMI) (Raghunathan et al., 2001), as implemented by the module mi impute chained command in STATA software.

The KFS Multiply Imputed Data files are available in Stata/SAS format at NORC. The following table shows the names of the files, the number of observations in each file, and the file's format:

File name	Original KFS data	KFS Multiply Imputed Data — format
KFS8_Cross_Sectional_wide_MI_Long	Wide (4,928 obs.)	Flong (4928*6=29,568 obs.)
KFS8_Cross_Sectional_Long_MI_Long	Long (4,928*8=39,424 obs.)	Flong (39,424*6=236,544 obs.)
KFS8_Longitudinal_wide_MI_Long	Wide (3,140 obs.)	Flong (3,140*6=18,840 obs.)
KFS8_Longitudinal_Long_MI_Long	Long	Flong (18,286*6=109,716 obs.)

While Stata allows other formats (wide, mlong), we do not recommend converting the data to another format; when you construct a new variable, it could be a super-varying variable. You must use the flong format because in the wide and mlong formats, there is simply no place to store super-varying values.

A variable is said to be super varying if its values in the complete observations differ across m. While the existence of super-varying variables is usually an indication of error, such is not the case for the KFS's super-varying variables. The KFS has a complex skip logic that will produce super-varying variables.

SPSS Users: While you can import the KFS Multiply Imputed Stata files to SPSS, you should be aware that (as of SPSS19) Complex Sampling procedures in SPSS currently do not automatically analyze multiply imputed datasets.

SAS Users: Use MIANALYZE procedures to analyze a multiply imputed data set. Also, SAS-callable SUDAAN includes a built-in option for analyzing multiply imputed data.

Super-varying variables are not an issue that you should be worried about in both SAS and SPSS.

3.2.3.1. The Stata MI Suite of Commands

"The mi suite of commands deals with multiple-imputation data, abbreviated as 'mi data.'" In summary,

1. mi data may be stored in one of four formats—flongsep, flong, mlong, and wide—known as styles.
2. mi data contain M imputations numbered m = 1, 2, : : : , M, and contain m = 0, the original data with missing values.
3. Each variable in mi data is registered as imputed, passive, or regular, or it is unregistered.
    a. Unregistered variables are mostly treated like regular variables.
    b. Regular variables usually do not contain missing values, but if they do, the missing values are not imputed in m > 0.
    c. Imputed variables contain missing values in m = 0, and those values are imputed in m > 0.
    d. Passive variables are algebraic combinations of imputed, regular, or other passive variables.
4. If an imputed variable contains a value greater than "." in m = 0, it contains .a, .b, : : , .z—then that value is considered a hard missing and the missing value persists in m > 0."

In Stata®, the definitions of MI variables are:

1. A regular variable is a variable that is neither imputed nor passive and that has the same values, whether missing or not, in all m (e.g., c4_numowners_confirm). New variables that are functions of existing regular variables are also regular variables.
2. An imputed variable is a variable that has missing values for which there are imputations. An imputed variable will have missing values in m = 0 and varying values for observations in m > 0.
3. A passive variable is a varying variable that is a function of imputed variables or of other passive variables. A passive variable will have missing values in m = 0 and varying values for observations in m > 0. You can use mi passive with any function that produces values that solely depend on values within the observation. In general, you cannot use mi passive with functions that produce values that depend on groups of observations.

Two other definitions that they use in the manual, the definitions for varying and super varying.

4. A variable is said to be varying if its values in the incomplete observations (missing) differ across imputations. Imputed and passive variables are varying. Regular variables are nonvarying. Unregistered variables can be either.

5. A variable is said to be super varying if its values in the complete observations (no missing in m=0) differ across imputations.

The distinction between varying variables and super-varying variables allows -mi- to detect inconsistencies among complete observations across imputations and to fix such inconsistencies. Variables that are functions of the values of other imputed variables are likely to be super-varying (e.g., skip logic). In KFS, a super-varying variable could be a result of a skip logic where the sample varies across imputations, incorrect flow through prescribed skip patterns, and/or inconsistency in values after editing the raw data. Super-varying variables must not be registered.

Stata users must be sure that they understand the meaning of regular, imputed, passive, varying, and super varying variables and what register, mi passive, mi update, and automatic update do to the data when you create or change variables.

For a better understanding of how Stata deals with mi data, let us consider the following artificial survey data.

midata	mprid	x1_0	x2_0	x3_0	x4_0	x5_0	_mi_m	_mi_id	_mi_miss
0	1	.	.	5000	.	.	0	1	1
0	2	55	.	2500	1	10000	0	2	1
0	3	89	96	3000	0	.a	0	3	0
1	1	45	23	5000	1	15000	1	1	.
1	2	55	85	2500	1	10000	1	2	.
1	3	89	96	3000	0	.a	1	3	.
2	1	52	64	5000	0	.a	2	1	.
2	2	55	27	2500	1	10000	2	2	.
2	3	89	96	3000	0	.a	2	3	.

Where

Var	**Description**	**Type**
midata	A variable identifying the original data (m=0) as well as two multiply imputed data (m=1,2)	Super varying, we should leave it unregistered
mprid	Business ID	Regular variable/Non-varying
x1_0	Total current assets	Imputed variable/Varying
x2_0	Total fixed assets	Imputed variable/Varying
x3_0	Amount of equity invested by respondent	Regular variable/Non-varying
x4_0	If the business has more than one owner (1=yes, 0=no)	Imputed variable/Varying
x5_0	(Skip x5_0 if x4_0 is No) Total amount of equity invested by other owners	Imputed variable/Varying
_mi_m	System variable	DO NOT drop or delete this variable
_mi_id	System variable	DO NOT drop or delete this variable
_mi_miss	System variable	DO NOT drop or delete this variable

Using mi describe, we can see the style of the data, the number of complete and incomplete observations, M (the number of imputations), the registered variables, and the number of missing values in m=0 of the imputed and passive variables.

```
mi describe

  Style:  flong
          last mi update 03jan2014 09:08:57, 0 seconds ago

  Obs.:   complete            1
          incomplete          2  (M = 2 imputations)
          --------------------
          total               3

  Vars.:  imputed:  4; x1_0(1) x2_0(2) x4_0(1) x5_0(1+1)
          passive:  0
          regular:  2; x3_0 mprid
          system:   3; _mi_m _mi_id _mi_miss
         (there is one unregistered variable; midata)
```

Let see what will happen if we mistakenly registered a super varying variable like midata.

```
mi register regular midata
(6 values of regular variable mi in m>0 updated to match values in m=0)
```

midata	mprid	x1_0	x2_0	x3_0	x4_0	x5_0
0	1	.	.	5000	.	.
0	2	55	.	2500	1	10000
0	3	89	96	3000	0	.a
0	1	45	23	5000	1	15000
0	2	55	85	2500	1	10000
0	3	89	96	3000	0	.a
0	1	52	64	5000	0	.a
0	2	55	27	2500	1	10000
0	3	89	96	3000	0	.a

Because a regular variable should have the same values in all m, Stata will update the data in m>0 to match values in m=0, which is wrong; thus, super varying variables should be left unregistered (Stata leaves unregistered super varying variables alone).

Next, we will create new variables: total assets and total equity using egen command[2]. Given that total assets and total equity are a function of imputed variables, they must be passive by definition. Keep in mind that a passive variable is a varying

[2] Most egen functions result in super-varying variables.

variable that is a function of imputed variables or of other passive variables. We will use mi passive command. (If you create passive variables by using mi passive, that command automatically registers them for you.)

```
mi passive : egen Total_assets_0=rowtotal(x1_0 x2_0),missing
mi passive : egen Total_equity_0=rowtotal(x3_0 x5_0),missing
```

midata	mprid	x1_0	x2_0	x3_0	x4_0	x5_0	Total_assets_0	Total_equity_0
0	1	.	.	5000	.	.	.	5000
0	2	55	.	2500	1	10000	55	12500
0	3	89	96	3000	0	.a	185	3000
1	1	45	23	5000	1	15000	68	20000
1	2	55	85	2500	1	10000	140	12500
1	3	89	96	3000	0	.a	185	3000
2	1	52	64	5000	0	.a	116	5000
2	2	55	27	2500	1	10000	82	12500
2	3	89	96	3000	0	.a	185	3000

So, did we have the right results? No, both total assets and total equity are super varying. While Stata did not update the data in m > 0 to match values in m = 0 at this point, it will run the update automatically later and update the data. The distinction between passive (varying) variables and super-varying variables allows Stata to detect inconsistencies among complete observations across imputations and fix such inconsistencies.

How about using gen command instead of egen?

```
mi passive : gen Total_assets_0=x1_0+ x2_0
mi passive : gen Total_equity_0=x3_0+ x5_0
```

midata	mprid	x1_0	x2_0	x3_0	x4_0	x5_0	Total_assets_0	Total_equity_0
0	1	.	.	5000	.		.	.
0	2	55	.	2500	1	10000	.	12500
0	3	89	96	3000	0	.a	185	.
1	1	45	23	5000	1	15000	68	20000
1	2	55	85	2500	1	10000	140	12500
1	3	89	96	3000	0	.a	185	.
2	1	52	64	5000	0	.a	116	.
2	2	55	27	2500	1	10000	82	12500
2	3	89	96	3000	0	.a	185	.

So, did we have the right results? The answer is yes and no. For the total assets variable, we have a passive variable, which is a varying variable. As for total equity, we

do not have the right results. This is because we have a skip logic for the x5_0 variable (coded as hard missing) and the gen command produced missing values in m > 0, which is not correct.

Given that almost every variable in KFS involves some type of skip logic, how can we overcome the above issues to avoid making any mistakes in creating new variables?

The answer is simple: Stata requires imputed variables to be registered (during imputation process). Meanwhile, it is only recommended that you register passive or regular variables. Thus, we can leave all the newly created variables unregistered (some will be super varying). Keep in mind that the purpose of registering the variables in Stata is to verify that the mi data are consistent; thus, having unregistered variables will not affect your analysis.

```
egen Total_assets_0=rowtotal(x1_0 x2_0),missing
egen Total_equity_0=rowtotal(x3_0 x5_0),missing
```

midata	mprid	x1_0	x2_0	x3_0	x4_0	x5_0	Total_assets_0	Total_equity_0
0	1	.	.	5000	.	.	.	5000
0	2	55	.	2500	1	10000	55	12500
0	3	89	96	3000	0	.a	185	3000
1	1	45	23	5000	1	15000	68	20000
1	2	55	85	2500	1	10000	140	12500
1	3	89	96	3000	0	.a	185	3000
2	1	52	64	5000	0	.a	116	5000
2	2	55	27	2500	1	10000	82	12500
2	3	89	96	3000	0	.a	185	3000

Because we did not register total assets and total equity (both are super varying), Stata will never attempt to update them.

If you insist on registering newly created passive variables, more work needs to be done. We can use egen command, but we need to make sure that we will not create a super varying variable. You need to exercise caution when using this approach.

```
egen Total_assets_0=rowtotal(x1_0 x2_0),missing
egen Total_equity_0=rowtotal(x3_0 x5_0),missing
replace Total_assets_0=. if x1_0==. & midata==0
replace Total_assets_0=. if x2_0==. & midata==0
replace  Total_equity_0=. if x3_0==. & midata==0
replace  Total_equity_0=. if x5_0==. & midata==0

mi register passive Total_assets_0 Total_equity_0
```

midata	mprid	x1_0	x2_0	x3_0	x4_0	x5_0	Total_assets_0	Total_equity_0
0	1	.	.	5000	.	.	.	.
0	2	55	.	2500	1	10000	.	12500
0	3	89	96	3000	0	.a	185	3000
1	1	45	23	5000	1	15000	68	20000
1	2	55	85	2500	1	10000	140	12500
1	3	89	96	3000	0	.a	185	3000
2	1	52	64	5000	0	.a	116	5000
2	2	55	27	2500	1	10000	82	12500
2	3	89	96	3000	0	.a	185	3000

Now we have both total assets and total equity as a passive varying variable.

3.2.3.2. Creating or Changing Variables

All KFS multiply imputed data files already are declared to be multiple-imputation data, thus they do not use mi set. Commands like svyset, stset, and xtset also have mi versions; use mi svyset to declare survey data, use mi stset to declare survival data, and use mi xtset to declare panel data.

We already register the varying variables in the KFS MI files. All variables were checked for consistency across imputation and were registered correctly. You can use the command "mi varying" to see the registered/unregistered variables.

Some imputed variables were not registered for various reasons (mainly because they are super-varying variables); any newly created variable that uses any of these unregistered variables should not be registered. The following table shows the names of all unregistered variables in KFS multiply imputed data files.

The major problem we will face when using mi xeq and mi passive with gen/egen/replace/rename commands is that the mi xeq and mi passive are not memory efficient.[3] With a large data set like KFS, some programs could take hours to run. To overcome this issue, we will show how to correctly create/change variables in a memory efficient way.

[3] You can use data-modification commands with mi xeq, but doing so is not especially useful unless you are using flongsep data.

3.2.3.2.1. Stata Code: Cross Sectional in Wide Format

The following Stata code will create the same variables we discussed in sections 3.2.2.2.1, 3.2.2.2.2, and 3.2.2.2.3, using the KFS multiply imputed data file "KFS8_Cross_Sectional_wide_MI_Long." The code will save the new file under the name "Cross_Sectional_wide_MI_Long_w1.dta" (n=4928*6).

```
clear
clear mata
clear matrix
capture log close
set more off
set maxvar 32767
program drop _all
adopath + " xxx:\KFS_Manual_and_Data\Farhat_Robb_Commands\"
cd "Xxx:\KFS_Manual_and_Data"
use KFS8_Cross_Sectional_wide_MI_Long,clear

forvalues i = 0/7 {
egen   Tot_Equity_Owner_Operators_`i'=  rowtotal(    f2_owner_amt_eq_invest_*_`i'      )      , missing
egen   Tot_Equity_OwnerOper_AllYrs_`i'=rowtotal(     f2_ownr_amt_eqinvest_allyrs_*_`i' ), missing
}

forvalues i = 0/7 {
foreach ow in $owners_1_15  {
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Tot_Equity_Owner_Operators_`i' =.a if classf_`i'<6
replace Tot_Equity_OwnerOper_AllYrs_`i'=.a if classf_`i'<6
/*To prevent super-varying variables, Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Equity_Owner_Operators_`i' =.   if  f2_owner_amt_eq_invest_`ow'_`i'==.
replace Tot_Equity_OwnerOper_AllYrs_`i'=.   if  f2_ownr_amt_eqinvest_allyrs_`ow'_`i'==.
}
}

global List1 "spouse parents angels companies govt vent_cap other"

forvalues i = 0/7 {

egen  Tot_Equity_NonOwnerOperators_`i'=rowtotal(f4_eq_amt_angels_`i' f4_eq_amt_companies_`i' f4_eq_amt_govt_`i' ///
f4_eq_amt_other_`i' f4_eq_amt_parents_`i' f4_eq_amt_spouse_`i'  f4_eq_amt_vent_cap_`i')      , missing
```

```
egen  Tot_Equity_NonOwnerOp_AllYrs_`i'=rowtotal(f4_eq_amt_*_allyrs_`i')   , missing
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Tot_Equity_NonOwnerOperators_`i' =.a if classf_`i'<6
replace Tot_Equity_NonOwnerOperators_`i' =.a if  c1z2_legal_status_`i'==1

replace Tot_Equity_NonOwnerOp_AllYrs_`i' =.a if classf_`i'<6
replace Tot_Equity_NonOwnerOp_AllYrs_`i' =.a if  c1z2_legal_status_`i'==1
}

forvalues i = 0/7 {
foreach name in $List1  {
/*To prevent super-varying variables, Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Equity_NonOwnerOperators_`i' =. if f4_eq_amt_`name'_`i'==.
replace Tot_Equity_NonOwnerOp_AllYrs_`i' =. if f4_eq_amt_`name'_allyrs_`i'==.

}
}

forvalues i = 0/7 {
egen  Tot_Equity_`i'       =rowtotal(Tot_Equity_Owner_Operators_`i' Tot_Equity_NonOwnerOperators_`i'), missing
egen  Tot_Equity_AllYrs_`i'=rowtotal(Tot_Equity_OwnerOper_AllYrs_`i' Tot_Equity_NonOwnerOp_AllYrs_`i'), missing
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Tot_Equity_`i'       =.a if classf_`i'<6
replace Tot_Equity_AllYrs_`i'=.a if classf_`i'<6
/*To prevent super-varying variables, Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Equity_`i'       =. if Tot_Equity_Owner_Operators_`i'==.
replace Tot_Equity_`i'       =. if Tot_Equity_NonOwnerOperators_`i'==.
replace Tot_Equity_AllYrs_`i'=. if Tot_Equity_NonOwnerOperators_`i'==.
replace Tot_Equity_AllYrs_`i'=. if Tot_Equity_NonOwnerOp_AllYrs_`i'==.
}

forvalues i = 0/7 {
egen  Tot_Assets_`i'=rowtotal(f29_assetval_*_`i')    , missing
egen  Tot_Liab_`i'=rowtotal(f31_value_*_`i'), missing
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Tot_Assets_`i' =.a if classf_`i'<6
replace Tot_Liab_`i'   =.a if classf_`i'<6
/*To prevent super-varying variables, Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Assets_`i' =. if f29_assetval_acctrec_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_cash_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_equip_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_inv_`i'==.
```

```
replace Tot_Assets_`i' =. if f29_assetval_landbuild_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_othbusprop_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_other_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_veh_`i'==.
replace Tot_Liab_`i' =. if  f31_value_acctpay_`i'==.
replace Tot_Liab_`i' =. if  f31_value_other_`i'==.
replace Tot_Liab_`i' =. if  f31_value_pension_`i'==.
}

forvalues i = 0/7 {
egen  Tot_Pers_Debt_Resp_`i'=rowtotal(f8b_pers_credcard_bal_`i' f8b_bus_credcard_bal_`i' f8c_pers_loan_bank_amt_`i'
f8c_pers_loan_fam_amt_`i' f8c_pers_loan_other_amt_`i' f8c_pers_other_amt_`i'),missing
egen  Tot_Pers_Debt_Owed_Resp_`i'=rowtotal(f8b_pers_credcard_bal_`i' f8b_bus_credcard_bal_`i'
f8d_pers_loan_bank_owed_`i' f8d_pers_loan_fam_owed_`i' f8d_pers_loan_other_owed_`i' f8d_pers_other_owed_`i'),missing
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Tot_Pers_Debt_Resp_`i'      =.a if classf_`i'<6
replace Tot_Pers_Debt_Owed_Resp_`i' =.a if classf_`i'<6
/*To prevent super-varying variables, Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Pers_Debt_Resp_`i' =. if f8b_pers_credcard_bal_`i'==.
replace Tot_Pers_Debt_Resp_`i' =. if f8b_bus_credcard_bal_`i'==.
replace Tot_Pers_Debt_Resp_`i' =. if f8c_pers_loan_bank_amt_`i'==.
replace Tot_Pers_Debt_Resp_`i' =. if f8c_pers_loan_fam_amt_`i'==.
replace Tot_Pers_Debt_Resp_`i' =. if f8c_pers_loan_other_amt_`i'==.
replace Tot_Pers_Debt_Resp_`i' =. if f8c_pers_other_amt_`i'==.
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8b_pers_credcard_bal_`i'==.
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8b_bus_credcard_bal_`i'==.
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8d_pers_loan_bank_owed_`i'==.
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8d_pers_loan_fam_owed_`i'==.
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8d_pers_loan_other_owed_`i'==.
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8d_pers_other_owed_`i'==.
}

forvalues i = 0/7 {
egen  Tot_Pers_Debt_Other_Owners_`i'=rowtotal(f10b_pers_credcard_bal_`i' f10c_pers_loan_bank_amt_`i'
f10b_bus_credcard_bal_`i' f10c_pers_loan_fam_amt_`i' f10c_pers_loan_other_amt_`i'  f10c_pers_other_amt_`i'),missing
egen  Tot_Pers_Debt_Owed_OthrOwnrs_`i'=rowtotal(f10b_pers_credcard_bal_`i' f10b_bus_credcard_bal_`i'
f10d_pers_loan_bank_owed_`i' f10d_pers_loan_fam_owed_`i' f10d_pers_loan_other_owed_`i'
f10d_pers_other_owed_`i'),missing
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Tot_Pers_Debt_Other_Owners_`i'  =.a if classf_`i'<6
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i'=.a if classf_`i'<6
replace Tot_Pers_Debt_Other_Owners_`i'  =.a if c4_numowners_confirm_`i'<2
```

```
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i'=.a if c4_numowners_confirm_`i'<2
/*To prevent super-varying variables, Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10b_pers_credcard_bal_`i'==.
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10b_bus_credcard_bal_`i'==.
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10c_pers_loan_bank_amt_`i'==.
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10c_pers_loan_fam_amt_`i'==.
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10c_pers_loan_other_amt_`i'==.
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10c_pers_other_amt_`i'==.
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10b_pers_credcard_bal_`i'==.
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10b_bus_credcard_bal_`i'==.
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10d_pers_loan_bank_owed_`i'==.
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10d_pers_loan_fam_owed_`i'==.
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10d_pers_loan_other_owed_`i'==.
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10d_pers_other_owed_`i'==.
}

forvalues i = 0/7 {
egen  Tot_Debt_Owner_Operators_`i'=rowtotal(Tot_Pers_Debt_Resp_`i' Tot_Pers_Debt_Other_Owners_`i'),missing
egen  Tot_Debt_Owed_Owner_Operators_`i'=rowtotal(Tot_Pers_Debt_Owed_Resp_`i' Tot_Pers_Debt_Owed_OthrOwnrs_`i'),missing
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Tot_Debt_Owner_Operators_`i'   =.a if classf_`i'<6
replace Tot_Debt_Owed_Owner_Operators_`i'=.a if classf_`i'<6
/*To prevent super-varying variables, Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Debt_Owner_Operators_`i'   =. if Tot_Pers_Debt_Resp_`i'==.
replace Tot_Debt_Owner_Operators_`i'   =. if Tot_Pers_Debt_Other_Owners_`i'==.
replace Tot_Debt_Owed_Owner_Operators_`i'=. if Tot_Pers_Debt_Owed_Resp_`i'==.
replace Tot_Debt_Owed_Owner_Operators_`i'=. if Tot_Pers_Debt_Owed_OthrOwnrs_`i'==.
}

forvalues i = 0/7 {
egen Tot_Debt_Bus_`i'=rowtotal(f12b_bus_credcard_bal_`i' f12c_bus_loans_bank_amt_`i' f12b_bus_cred_line_bal_`i'
f12c_bus_loans_nonbank_amt_`i'    f12c_bus_loans_fam_amt_`i' f12c_bus_loans_govt_amt_`i' f12c_bus_loans_emp_amt_`i'
f12c_bus_loans_other_ind_amt_`i' f12c_bus_loans_owner_amt_`i' f12c_bus_loans_bus_amt_`i'
      f12c_bus_other_amt_`i'),missing
egen Tot_Bus_Debt_Owed_`i'=rowtotal(f12b_bus_cred_line_bal_`i'  f12b_bus_credcard_bal_`i' f12d_bus_loans_bank_owed_`i'
f12d_bus_loans_nonbank_owed_`i' f12d_bus_loans_emp_owed_`i'  f12d_bus_loans_fam_owed_`i' f12d_bus_loans_govt_owed_`i'
f12d_bus_loans_other_ind_owed_`i' f12d_bus_loans_owner_owed_`i'  f12d_bus_loans_bus_owed_`i'
f12d_bus_other_owed_`i'),missing
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Tot_Debt_Bus_`i'     =.a if classf_`i'<6
replace Tot_Bus_Debt_Owed_`i'=.a if classf_`i'<6
/*To prevent super-varying variables, Recode to soft missing value if any of the total's component is soft missing*/
```

```
replace Tot_Debt_Bus_`i'   =. if f12b_bus_credcard_bal_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12b_bus_cred_line_bal_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_bank_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_nonbank_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_fam_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_govt_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_emp_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_other_ind_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_owner_amt_`i' ==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_bus_amt_`i' ==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_other_amt_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12b_bus_credcard_bal_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12b_bus_cred_line_bal_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_bank_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_nonbank_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_fam_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_govt_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_emp_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_other_ind_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_owner_owed_`i' ==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_bus_owed_`i'  ==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_other_owed_`i'==.
}

forvalues i = 0/7 {
egen  Tot_Debt_`i'=rowtotal(Tot_Debt_Owner_Operators_`i' Tot_Debt_Bus_`i'),missing
egen  Tot_Debt_Owed_`i'=rowtotal(Tot_Debt_Owed_Owner_Operators_`i' Tot_Bus_Debt_Owed_`i'),missing
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Tot_Debt_`i'      =.a if classf_`i'<6
replace Tot_Debt_Owed_`i' =.a if classf_`i'<6
/*To prevent super-varying variables, Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Debt_`i'   =. if Tot_Debt_Owner_Operators_`i'==.
replace Tot_Debt_`i'   =. if Tot_Debt_Bus_`i'==.
replace Tot_Debt_Owed_`i'=. if Tot_Debt_Owed_Owner_Operators_`i'==.
replace Tot_Debt_Owed_`i'=. if Tot_Bus_Debt_Owed_`i'==.
}

forvalues i = 0/7 {
gen  Net_Profit_`i'=f24_profitloss_amt_`i'
}

rename Tot_* *
```

```
 mi register passive Net_Profit_0 Net_Profit_1 Net_Profit_2 Net_Profit_3 Net_Profit_4 Net_Profit_5 Net_Profit_6
Net_Profit_7 ///
Assets_0 Assets_1 Assets_2 Assets_3 Assets_4 Assets_5 Assets_6 Assets_7  ///
Liab_0 Liab_1 Liab_2 Liab_3 Liab_4 Liab_5 Liab_6 Liab_7  ///
Pers_Debt_Owed_Resp_0 Pers_Debt_Owed_Resp_1 Pers_Debt_Owed_Resp_2 Pers_Debt_Owed_Resp_3 Pers_Debt_Owed_Resp_4
Pers_Debt_Owed_Resp_5 Pers_Debt_Owed_Resp_6 Pers_Debt_Owed_Resp_7  ///
Pers_Debt_Resp_0 Pers_Debt_Resp_1 Pers_Debt_Resp_2 Pers_Debt_Resp_3 Pers_Debt_Resp_4 Pers_Debt_Resp_5 Pers_Debt_Resp_6
Pers_Debt_Resp_7

/*
Bus_Debt_Owed_0 Bus_Debt_Owed_1 Bus_Debt_Owed_2 Bus_Debt_Owed_3 Bus_Debt_Owed_4 Bus_Debt_Owed_5 Bus_Debt_Owed_6
Bus_Debt_Owed_7
Debt_0 Debt_1 Debt_2 Debt_3 Debt_4 Debt_5 Debt_6 Debt_7 ///
Debt_Bus_0 Debt_Bus_1 Debt_Bus_2 Debt_Bus_3 Debt_Bus_4 Debt_Bus_5 Debt_Bus_6 Debt_Bus_7
Debt_Owed_0 Debt_Owed_1 Debt_Owed_2 Debt_Owed_3 Debt_Owed_4 Debt_Owed_5 Debt_Owed_6 Debt_Owed_7
Debt_Owed_Owner_Operators_0 Debt_Owed_Owner_Operators_1 Debt_Owed_Owner_Operators_2 Debt_Owed_Owner_Operators_3
Debt_Owed_Owner_Operators_4 Debt_Owed_Owner_Operators_5 Debt_Owed_Owner_Operators_6
Debt_Owed_Owner_Operators_7 Debt_Owner_Operators_0 Debt_Owner_Operators_1 Debt_Owner_Operators_2 Debt_Owner_Operators_3
Debt_Owner_Operators_4 Debt_Owner_Operators_5 Debt_Owner_Operators_6 Debt_Owner_Operators_7
Equity_OwnerOper_AllYrs_0 Equity_OwnerOper_AllYrs_1Equity_OwnerOper_AllYrs_2 Equity_OwnerOper_AllYrs_3
Equity_OwnerOper_AllYrs_4 Equity_OwnerOper_AllYrs_5 Equity_OwnerOper_AllYrs_6 Equity_OwnerOper_AllYrs_7
Equity_Owner_Operators_0 Equity_Owner_Operators_1 Equity_Owner_Operators_2 Equity_Owner_Operators_3
Equity_Owner_Operators_4 Equity_Owner_Operators_5 Equity_Owner_Operators_6 Equity_Owner_Operators_7
Pers_Debt_Other_Owners_0 Pers_Debt_Other_Owners_1 Pers_Debt_Other_Owners_2 Pers_Debt_Other_Owners_3
Pers_Debt_Other_Owners_4 Pers_Debt_Other_Owners_5 Pers_Debt_Other_Owners_6 Pers_Debt_Other_Owners_7
Pers_Debt_Owed_OthrOwnrs_0 Pers_Debt_Owed_OthrOwnrs_1 Pers_Debt_Owed_OthrOwnrs_2 Pers_Debt_Owed_OthrOwnrs_3
Pers_Debt_Owed_OthrOwnrs_4 Pers_Debt_Owed_OthrOwnrs_5 Pers_Debt_Owed_OthrOwnrs_6 Pers_Debt_Owed_OthrOwnrs_7
Equity_0 Equity_1 Equity_2 Equity_3 Equity_4 Equity_5 Equity_6 Equity_7
Equity_AllYrs_0 Equity_AllYrs_1 Equity_AllYrs_2 Equity_AllYrs_3 Equity_AllYrs_4 Equity_AllYrs_5 Equity_AllYrs_6
Equity_AllYrs_7
Equity_NonOwnerOp_AllYrs_0 Equity_NonOwnerOp_AllYrs_1 Equity_NonOwnerOp_AllYrs_2 Equity_NonOwnerOp_AllYrs_3
Equity_NonOwnerOp_AllYrs_4 Equity_NonOwnerOp_AllYrs_5 Equity_NonOwnerOp_AllYrs_6 Equity_NonOwnerOp_AllYrs_7
Equity_NonOwnerOperators_0 Equity_NonOwnerOperators_1 Equity_NonOwnerOperators_2 Equity_NonOwnerOperators_3
Equity_NonOwnerOperators_4 Equity_NonOwnerOperators_5 Equity_NonOwnerOperators_6 Equity_NonOwnerOperators_7
*/
/*****************************************************************/
/* Merge the file with the primary owner file “primary_owner.dta” */

merge m:1 mprid using "Xxx:\KFS_Manual_and_Data\primary_owner.dta"
drop _merge
* Make sure that the race do not change over time due to entry errors
forvalues i = 1/7 {
```

```
foreach ow in $owners_1_15  {

replace      g6_race_amind_owner_`ow'_`i'      =     g6_race_amind_owner_`ow'_0 if     g6_race_amind_owner_`ow'_`i'
      <.     &      g6_race_amind_owner_`ow'_0<.
replace      g6_race_asian_owner_`ow'_`i'      =     g6_race_asian_owner_`ow'_0 if     g6_race_asian_owner_`ow'_`i'
      <.     &      g6_race_asian_owner_`ow'_0<.
replace      g6_race_black_owner_`ow'_`i'      =     g6_race_black_owner_`ow'_0 if     g6_race_black_owner_`ow'_`i'
      <.     &      g6_race_black_owner_`ow'_0<.
replace      g6_race_nathaw_owner_`ow'_`i'     =     g6_race_nathaw_owner_`ow'_0      if
      g6_race_nathaw_owner_`ow'_`i'     <.     &     g6_race_nathaw_owner_`ow'_0<.
replace      g6_race_other_owner_`ow'_`i'      =     g6_race_other_owner_`ow'_0 if     g6_race_other_owner_`ow'_`i'
      <.     &      g6_race_other_owner_`ow'_0<.
replace      g6_race_white_owner_`ow'_`i'      =     g6_race_white_owner_`ow'_0 if     g6_race_white_owner_`ow'_`i'
      <.     &      g6_race_white_owner_`ow'_0<.
}
}

* Primary Owner(PO) Characteristics at the Baseline: PO by Robb's Stata code
forvalues i = 0/0 {
gen  PO_emp_`i'               =.
gen  PO_hours_`i'             =.
gen  PO_work_exp_`i'             =.
gen  PO_oth_bus_owner_`i'     =.
gen  PO_bus_same_ind_`i'      =.
gen  PO_age_owner_`i'            =.
gen  PO_hisp_origin_`i'      =.
gen  PO_race_group_`i'       =.
gen  PO_race_amind_owner_`i'   =.
gen  PO_race_asian_owner_`i'   =.
gen  PO_race_black_owner_`i'   =.
gen  PO_race_nathaw_owner_`i'  =.
gen  PO_race_other_owner_`i'   =.
gen  PO_race_white_owner_`i'   =.
gen  PO_native_born_`i'      =.
gen  PO_us_cit_`i'             =.
gen  PO_education_`i'             =.
gen  PO_gender_`i'             =.
}

forvalues i = 0/0 {
forvalues po = 1/6 {
replace PO_emp_`i'          =   g1a_emp_owner_0`po'_`i'            if  primary_owner==`po'
```

```
replace PO_hours_`i'          =  g1b1_hours_owner_0`po'_`i'                 if  primary_owner==`po'
replace PO_work_exp_`i'    =     g2_work_exp_owner_0`po'_`i'              if  primary_owner==`po'
replace PO_oth_bus_owner_`i'=    g3a_oth_bus_owner_0`po'_`i'              if  primary_owner==`po'
replace PO_bus_same_ind_`i' =    g3b_bus_same_ind_owner_0`po'_`i'  if  primary_owner==`po'
replace PO_age_owner_`i'   =     g4_age_owner_0`po'_`i'                     if  primary_owner==`po'
replace PO_hisp_origin_`i'       =     g5_hisp_origin_owner_0`po'_`i'       if  primary_owner==`po'
replace PO_race_group_`i'   =    g6b_race_group_0`po'_`i'              if  primary_owner==`po'
replace PO_race_amind_owner_`i' = g6_race_amind_owner_0`po'_`i'     if  primary_owner==`po'
replace PO_race_asian_owner_`i' = g6_race_asian_owner_0`po'_`i'     if  primary_owner==`po'
replace PO_race_black_owner_`i' = g6_race_black_owner_0`po'_`i'     if  primary_owner==`po'
replace PO_race_nathaw_owner_`i'= g6_race_nathaw_owner_0`po'_`i'    if  primary_owner==`po'
replace PO_race_other_owner_`i' = g6_race_other_owner_0`po'_`i'     if  primary_owner==`po'
replace PO_race_white_owner_`i' = g6_race_white_owner_0`po'_`i'     if  primary_owner==`po'
replace PO_native_born_`i'       =     g7_native_born_owner_0`po'_`i'       if  primary_owner==`po'
replace PO_us_cit_`i'          =  g8_us_cit_owner_0`po'_`i'             if  primary_owner==`po'
replace PO_education_`i'   =     g9_education_owner_0`po'_`i'           if  primary_owner==`po'
replace PO_gender_`i'          =  g10_gender_owner_0`po'_`i'           if  primary_owner==`po'
}
}

forvalues i = 1/7 {
gen  PO_emp_`i'          = PO_emp_0
gen  PO_hours_`i'        = PO_hours_0
gen  PO_work_exp_`i'        = PO_work_exp_0
gen  PO_oth_bus_owner_`i'= PO_oth_bus_owner_0
gen  PO_bus_same_ind_`i' = PO_bus_same_ind_0
gen  PO_age_owner_`i'       = PO_age_owner_0
gen  PO_hisp_origin_`i'  = PO_hisp_origin_0
gen  PO_race_group_`i'   = PO_race_group_0
gen  PO_race_amind_owner_`i' =PO_race_amind_owner_0
gen  PO_race_asian_owner_`i' =PO_race_asian_owner_0
gen  PO_race_black_owner_`i' =PO_race_black_owner_0
gen  PO_race_nathaw_owner_`i'=PO_race_nathaw_owner_0
gen  PO_race_other_owner_`i' =PO_race_other_owner_0
gen  PO_race_white_owner_`i' =PO_race_white_owner_0
gen  PO_native_born_`i'  = PO_native_born_0
gen  PO_us_cit_`i'          = PO_us_cit_0
gen  PO_education_`i'       = PO_education_0
gen  PO_gender_`i'       = PO_gender_0
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace  PO_emp_`i'               = .a if    classf_`i'  <6
replace  PO_hours_`i'             = .a if    classf_`i'  <6
```

```
replace  PO_work_exp_`i'        = .a if classf_`i'  <6
replace  PO_oth_bus_owner_`i'      = .a if     classf_`i'  <6
replace  PO_bus_same_ind_`i'       = .a if     classf_`i'  <6
replace  PO_age_owner_`i'       = .a if classf_`i'  <6
replace  PO_hisp_origin_`i'        = .a if     classf_`i'  <6
replace  PO_race_group_`i'         = .a if     classf_`i'  <6
replace  PO_race_amind_owner_`i' = .a if       classf_`i'  <6
replace  PO_race_asian_owner_`i' = .a if       classf_`i'  <6
replace  PO_race_black_owner_`i' = .a if       classf_`i'  <6
replace  PO_race_nathaw_owner_`i'= .a if       classf_`i'  <6
replace  PO_race_other_owner_`i' = .a if       classf_`i'  <6
replace  PO_race_white_owner_`i' = .a if       classf_`i'  <6
replace  PO_native_born_`i'        = .a if     classf_`i'  <6
replace  PO_us_cit_`i'               = .a if    classf_`i'  <6
replace  PO_education_`i'       = .a if classf_`i'  <6
replace  PO_gender_`i'              = .a if    classf_`i'  <6

}
* Active-Owner-Operators Characteristics (OO)
forvalues i = 0/7 {
egen   OO_emp_owner_`i'          =      rowmean(g1a_emp_owner_*_`i')
egen   OO_hours_owner_`i'        = rowmean(g1b1_hours_owner_*_`i')
egen   OO_work_exp_owner_`i'     =      rowmean(g2_work_exp_owner_*_`i'   )
egen   OO_oth_bus_owner_`i'      =      rowmean(g3a_oth_bus_owner_*_`i'   )
egen   OO_bus_same_ind_owner_`i' =      rowmean(g3b_bus_same_ind_owner_*_`i')
egen   OO_age_owner_`i'          =      rowmean(g4_age_owner_*_`i')
egen   OO_hisp_origin_owner_`i'  =      rowmean(g5_hisp_origin_owner_*_`i')
egen   OO_race_amind_owner_`i'   =      rowmean(g6_race_amind_owner_*_`i')
egen   OO_race_asian_owner_`i'   =      rowmean(g6_race_asian_owner_*_`i')
egen   OO_race_black_owner_`i'   =      rowmean(g6_race_black_owner_*_`i')
egen   OO_race_nathaw_owner_`i'  =      rowmean(g6_race_nathaw_owner_*_`i')
egen   OO_race_other_owner_`i'   =      rowmean(g6_race_other_owner_*_`i')
egen   OO_race_white_owner_`i'   =      rowmean(g6_race_white_owner_*_`i')
egen   OO_native_born_owner_`i'  =      rowmean(g7_native_born_owner_*_`i'      )
egen   OO_us_cit_owner_`i'       =      rowmean(g8_us_cit_owner_*_`i'     )
egen   OO_education_owner_`i'    =      rowmean(g9_education_owner_*_`i'  )
egen   md_education_owner_`i'    =      rowmedian(g9_education_owner_*_`i')
gen    OO_D_education_owner_`i'  =(md_education_owner_`i'    >6.99)        if     md_education_owner_`i'    <11

egen   OO_gender_owner_`i'       =      rowmean(g10_gender_owner_*_`i'    )
}
/*To prevent super-varying variables, Recode to soft missing value if any of the total's component is soft missing*/
```

```
forvalues i = 0/7 {
foreach ow in $owners_1_15  {
replace OO_emp_owner_`i'              =. if g1a_emp_owner_`ow'_`i'==.
replace OO_hours_owner_`i'          =. if g1b1_hours_owner_`ow'_`i'==.
replace OO_work_exp_owner_`i'     =. if g2_work_exp_owner_`ow'_`i'==.
replace OO_oth_bus_owner_`i'      =. if g3a_oth_bus_owner_`ow'_`i'==.
replace OO_bus_same_ind_owner_`i' =. if g3b_bus_same_ind_owner_`ow'_`i'==.
replace OO_age_owner_`i'              =. if g4_age_owner_`ow'_`i'==.
replace OO_hisp_origin_owner_`i'  =. if g5_hisp_origin_owner_`ow'_`i'==.
replace OO_race_amind_owner_`i'   =. if g6_race_amind_owner_`ow'_`i'==.
replace OO_race_asian_owner_`i'   =. if g6_race_asian_owner_`ow'_`i'==.
replace OO_race_black_owner_`i'   =. if g6_race_black_owner_`ow'_`i'==.
replace OO_race_nathaw_owner_`i'  =. if g6_race_nathaw_owner_`ow'_`i'==.
replace OO_race_other_owner_`i'   =. if g6_race_other_owner_`ow'_`i'==.
replace OO_race_white_owner_`i'   =. if g6_race_white_owner_`ow'_`i'==.
replace OO_native_born_owner_`i'  =. if g7_native_born_owner_`ow'_`i'==.
replace OO_us_cit_owner_`i'       =. if g8_us_cit_owner_`ow'_`i'==.
replace OO_education_owner_`i'     =. if g9_education_owner_`ow'_`i'==.
replace md_education_owner_`i'     =. if g9_education_owner_`ow'_`i'==.
replace OO_D_education_owner_`i'  =. if g9_education_owner_`ow'_`i'==.
replace OO_gender_owner_`i'       =. if g10_gender_owner_`ow'_`i'==.
}
}
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
forvalues i = 0/7 {
replace      OO_emp_owner_`i'          = .a if classf_`i'  <6
replace      OO_hours_owner_`i'        = .a if classf_`i'  <6
replace      OO_work_exp_owner_`i'     = .a if classf_`i'  <6
replace      OO_oth_bus_owner_`i'      = .a if classf_`i'  <6
replace      OO_bus_same_ind_owner_`i' = .a if classf_`i'  <6
replace      OO_age_owner_`i'          = .a if classf_`i'  <6
replace      OO_hisp_origin_owner_`i'  = .a if classf_`i'  <6
replace      OO_race_amind_owner_`i'   = .a if classf_`i'  <6
replace      OO_race_asian_owner_`i'   = .a if classf_`i'  <6
replace      OO_race_black_owner_`i'   = .a if classf_`i'  <6
replace      OO_race_nathaw_owner_`i'  = .a if classf_`i'  <6
replace      OO_race_other_owner_`i'   = .a if classf_`i'  <6
replace      OO_race_white_owner_`i'   = .a if classf_`i'  <6
replace      OO_native_born_owner_`i'  = .a if classf_`i'  <6
replace      OO_us_cit_owner_`i'       = .a if classf_`i'  <6
replace      OO_education_owner_`i'    = .a if classf_`i'  <6
replace      md_education_owner_`i'    = .a if classf_`i'  <6
```

```
replace      OO_D_education_owner_`i'  = .a if classf_`i'  <6

replace      OO_gender_owner_`i'       = .a if classf_`i'  <6
}
/*****************************************************************************/
*Diversity / Similarity index
forvalues i = 0/7 {
gen xr1_`i'= OO_race_amind_owner_`i'    *      OO_race_amind_owner_`i'
gen xr2_`i'= OO_race_asian_owner_`i'    *      OO_race_asian_owner_`i'
gen xr3_`i'= OO_race_black_owner_`i'    *      OO_race_black_owner_`i'
gen xr4_`i'= OO_race_nathaw_owner_`i'*  OO_race_nathaw_owner_`i'
gen xr5_`i'= OO_race_other_owner_`i' *  OO_race_other_owner_`i'
gen xr6_`i'= OO_race_white_owner_`i'    *      OO_race_white_owner_`i'
* Race_similarity
egen Race_similarity_`i'=rowtotal(xr1_`i' xr2_`i' xr3_`i' xr4_`i' xr5_`i' xr6_`i'),missing
drop xr*_`i'
* Race_ diversity
gen Race_diversity_`i'=1-Race_similarity_`i'
}

forvalues i = 0/7 {
foreach ow in $owners_1_15  {
/* Recode to soft missing value if any of the var's component is soft missing*/
replace Race_similarity_`i'   =. if g6_race_amind_owner_`ow'_`i'==.
replace Race_similarity_`i'   =. if g6_race_asian_owner_`ow'_`i'==.
replace Race_similarity_`i'   =. if g6_race_black_owner_`ow'_`i'==.
replace Race_similarity_`i'   =. if g6_race_nathaw_owner_`ow'_`i'==.
replace Race_similarity_`i'   =. if g6_race_other_owner_`ow'_`i'==.
replace Race_similarity_`i'   =. if g6_race_white_owner_`ow'_`i'==.

}
}
forvalues i = 0/7 {
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Race_similarity_`i'   =.a if classf_`i'<6
replace Race_diversity_`i'   =.a if Race_similarity_`i'==.a
replace Race_diversity_`i'   =.  if Race_similarity_`i'==.
}

forvalues i = 0/7 {
gen fmal_`i'=1-OO_gender_owner_`i'
gen xr1_`i'= OO_gender_owner_`i' *      OO_gender_owner_`i'
```

```
gen xr2_`i'= fmal_`i'     *      fmal_`i'

* Gender_similarity
egen Gender_similarity_`i'=rowtotal(xr1_`i' xr2_`i'  ),missing
drop xr*_`i' fmal_`i'
* Gender_diversity
gen Gender_diversity_`i'=1-Gender_similarity_`i'
}

forvalues i = 0/7 {
foreach ow in $owners_1_15  {
/* Recode to soft missing value if any of the var's component is soft missing*/
replace  Gender_similarity_`i'      =. if g10_gender_owner_`ow'_`i'==.
}
}

forvalues i = 0/7 {
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Gender_similarity_`i'   =.a if classf_`i'<6
replace Gender_diversity_`i'   =.a if Gender_similarity_`i'==.a
replace Gender_diversity_`i'   =.  if Gender_similarity_`i'==.
}

forvalues i = 0/7 {
mean Race_similarity_`i' Gender_similarity_`i' if c4_numowners_confirm_`i'>1 & c4_numowners_confirm_`i'<.
}

*Business level Characteristics
forvalues i = 0/7 {
*Home Based Dummy
recode c8_primary_loc_`i' (1=1 "Home Based") (nonmiss=0 "Non Home Based" ) ,into (Home_Based_`i')
*Sole_Proprietorship Dummy
recode c1z2_legal_status_`i' (1=1 "Sole_Proprietorship") (nonmiss=0 "Limited Liability" ) ,into
(Sole_Proprietorship_`i')
gen Comp_advantage_`i' = d2_comp_advantage_`i'
drop d2_comp_advantage_`i'
egen Have_IP_`i'=anymatch(d3_a_have_patent_`i' d3_b_have_copyright_`i' d3_c_have_trademark_`i'), values(1)
gen Full_Part_Time_Employees_`i'= c5_num_employees_`i'
gen Full_Time_Employees_`i' =c6_num_ft_employees_`i'
gen Part_Time_Employees_`i'=c7_num_pt_employees_`i'
drop c5_num_employees_`i'  c6_num_ft_employees_`i'  c7_num_pt_employees_`i'
egen   Employee_Owner_`i'          =    rowtotal(g1a_emp_owner_*_`i')
```

```
egen   Total_Employees_`i'=rowtotal(Employee_Owner_`i' Full_Part_Time_Employees_`i')
}

forvalues i = 0/7 {
foreach ow in $owners_1_15  {
/* Recode to soft missing value if any of the var's component is soft missing*/
replace Have_IP_`i'=. if d3_a_have_patent_`i'==.   | d3_b_have_copyright_`i'==.   | d3_c_have_trademark_`i'==.
replace Employee_Owner_`i'             =. if g1a_emp_owner_`ow'_`i'==.
replace Total_Employees_`i'            =. if g1a_emp_owner_`ow'_`i'==.
replace Total_Employees_`i'            =. if  Full_Part_Time_Employees_`i'==.
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Have_IP_`i'        =.a if classf_`i'<6
replace Employee_Owner_`i' =.a if classf_`i'<6
replace Total_Employees_`i'=.a if classf_`i'<6
}
}

mi register regular Home_Based* Sole_Proprietorship_*
mi register imputed Full_Part_Time_Employees* Full_Time_Employees* Part_Time_Employees*  Comp_advantage*
mi register passive Employee_Owner_0 Employee_Owner_1 Employee_Owner_2 Employee_Owner_3 Employee_Owner_4
Employee_Owner_5 ///
Employee_Owner_6 Employee_Owner_7  Gender_diversity_0 Gender_diversity_1 Gender_diversity_2 ///
Gender_diversity_3 Gender_diversity_4 Gender_diversity_5 Gender_diversity_6 Gender_diversity_7 Gender_similarity_0
Gender_similarity_1 Gender_similarity_2 ///
Gender_similarity_3 Gender_similarity_4 Gender_similarity_5 Gender_similarity_6 Gender_similarity_7 Have_IP_0 Have_IP_1
Have_IP_2 Have_IP_3 Have_IP_4 Have_IP_5 ///
Have_IP_6 Have_IP_7 OO_D_education_owner_0 OO_D_education_owner_1 OO_D_education_owner_2 OO_D_education_owner_3
OO_D_education_owner_4 OO_D_education_owner_5 ///
OO_D_education_owner_6 OO_D_education_owner_7 OO_age_owner_0 OO_age_owner_1 OO_age_owner_2 OO_age_owner_3
OO_age_owner_4 OO_age_owner_5 OO_age_owner_6 ///
OO_age_owner_7 OO_bus_same_ind_owner_0 OO_bus_same_ind_owner_1 OO_bus_same_ind_owner_2 OO_bus_same_ind_owner_3
OO_bus_same_ind_owner_4 OO_bus_same_ind_owner_5 ///
OO_bus_same_ind_owner_6 OO_bus_same_ind_owner_7 OO_education_owner_0 OO_education_owner_1 OO_education_owner_2
OO_education_owner_3 OO_education_owner_4 ///
OO_education_owner_5 OO_education_owner_6 OO_education_owner_7 OO_emp_owner_0 OO_emp_owner_1 OO_emp_owner_2
OO_emp_owner_3 OO_emp_owner_4 OO_emp_owner_5 ///
OO_emp_owner_6 OO_emp_owner_7 OO_gender_owner_0 OO_gender_owner_1 OO_gender_owner_2 OO_gender_owner_3 OO_gender_owner_4
OO_gender_owner_5 OO_gender_owner_6 ///
OO_gender_owner_7 OO_hisp_origin_owner_0 OO_hisp_origin_owner_1 OO_hisp_origin_owner_2 OO_hisp_origin_owner_3
OO_hisp_origin_owner_4 OO_hisp_origin_owner_5 ///
OO_hisp_origin_owner_6 OO_hisp_origin_owner_7 OO_hours_owner_0 OO_hours_owner_1 OO_hours_owner_2 OO_hours_owner_3
OO_hours_owner_4 OO_hours_owner_5 ///
```

```
OO_hours_owner_6 OO_hours_owner_7 OO_native_born_owner_0 OO_native_born_owner_1 OO_native_born_owner_2
OO_native_born_owner_3 OO_native_born_owner_4 ///
OO_native_born_owner_5 OO_native_born_owner_6 OO_native_born_owner_7 OO_oth_bus_owner_0 OO_oth_bus_owner_1
OO_oth_bus_owner_2 OO_oth_bus_owner_3 ///
OO_oth_bus_owner_4 OO_oth_bus_owner_5 OO_oth_bus_owner_6 OO_oth_bus_owner_7 OO_race_amind_owner_0 OO_race_amind_owner_1
OO_race_amind_owner_2 ///
OO_race_amind_owner_3 OO_race_amind_owner_4 OO_race_amind_owner_5 OO_race_amind_owner_6 OO_race_amind_owner_7
OO_race_asian_owner_0 OO_race_asian_owner_1 ///
OO_race_asian_owner_2 OO_race_asian_owner_3 OO_race_asian_owner_4 OO_race_asian_owner_5 OO_race_asian_owner_6
OO_race_asian_owner_7 OO_race_black_owner_0 ///
OO_race_black_owner_1 OO_race_black_owner_2 OO_race_black_owner_3 OO_race_black_owner_4 OO_race_black_owner_5
OO_race_black_owner_6 OO_race_black_owner_7 ///
OO_race_nathaw_owner_0 OO_race_nathaw_owner_1 OO_race_nathaw_owner_2 OO_race_nathaw_owner_3 OO_race_nathaw_owner_4
OO_race_nathaw_owner_5 ///
OO_race_nathaw_owner_6 OO_race_nathaw_owner_7 OO_race_other_owner_0 OO_race_other_owner_1 OO_race_other_owner_2
OO_race_other_owner_3 OO_race_other_owner_4 ///
OO_race_other_owner_5 OO_race_other_owner_6 OO_race_other_owner_7 OO_race_white_owner_0 OO_race_white_owner_1
OO_race_white_owner_2 OO_race_white_owner_3 ///
OO_race_white_owner_4 OO_race_white_owner_5 OO_race_white_owner_6 OO_race_white_owner_7 OO_us_cit_owner_0
OO_us_cit_owner_1 OO_us_cit_owner_2 OO_us_cit_owner_3 ///
OO_us_cit_owner_4 OO_us_cit_owner_5 OO_us_cit_owner_6 OO_us_cit_owner_7 OO_work_exp_owner_0 OO_work_exp_owner_1
OO_work_exp_owner_2 OO_work_exp_owner_3 ///
OO_work_exp_owner_4 OO_work_exp_owner_5 OO_work_exp_owner_7 PO_age_owner_0 PO_age_owner_1 PO_age_owner_2
Race_diversity_0 Race_diversity_1 Race_diversity_2 Race_diversity_3 ///
Race_diversity_4 Race_diversity_5 Race_diversity_6 Race_diversity_7 Race_similarity_0 Race_similarity_1
Race_similarity_2 Race_similarity_3 Race_similarity_4 ///
Race_similarity_5 Race_similarity_6 Race_similarity_7 Total_Employees_0 Total_Employees_1 Total_Employees_2
Total_Employees_3 Total_Employees_4 ///
Total_Employees_5 Total_Employees_6 Total_Employees_7 md_education_owner_0 md_education_owner_1 md_education_owner_2
md_education_owner_3 md_education_owner_4 ///
md_education_owner_5 md_education_owner_6 md_education_owner_7 PO_age_owner_3 PO_age_owner_4 PO_age_owner_5 ///
PO_age_owner_6 PO_age_owner_7 PO_bus_same_ind_0 PO_bus_same_ind_1 PO_bus_same_ind_2 PO_bus_same_ind_3 PO_bus_same_ind_4
PO_bus_same_ind_5 PO_bus_same_ind_6 ///
PO_bus_same_ind_7 PO_education_0 PO_education_1 PO_education_2 PO_education_3 PO_education_4 PO_education_5 PO_emp_0
PO_emp_1 PO_emp_2 PO_emp_3 PO_emp_4 ///
PO_emp_5 PO_emp_6 PO_emp_7 PO_gender_0 PO_gender_1 PO_gender_2 PO_gender_3 PO_gender_4 PO_gender_5 PO_hisp_origin_0
PO_hisp_origin_1 PO_hisp_origin_2 ///
PO_hisp_origin_3 PO_hisp_origin_4 PO_hisp_origin_5 PO_hisp_origin_6 PO_hisp_origin_7 PO_hours_0 PO_hours_1 PO_hours_2
PO_hours_3 PO_hours_4 PO_hours_5 ///
PO_hours_6 PO_hours_7 PO_native_born_0 PO_native_born_1 PO_native_born_2 PO_native_born_3 PO_native_born_4
PO_native_born_5 PO_native_born_6 PO_native_born_7 ///
PO_oth_bus_owner_0 PO_oth_bus_owner_1 PO_oth_bus_owner_2 PO_oth_bus_owner_3 PO_oth_bus_owner_4 PO_oth_bus_owner_5
```

```
PO_oth_bus_owner_6 PO_oth_bus_owner_7 ///
PO_race_amind_owner_0 PO_race_amind_owner_7 PO_race_asian_owner_0 PO_race_asian_owner_7 PO_race_black_owner_0
PO_race_black_owner_7 PO_race_group_0 ///
PO_race_group_7 PO_race_nathaw_owner_0 PO_race_nathaw_owner_7 PO_race_other_owner_0 PO_race_other_owner_7
PO_race_white_owner_0 PO_race_white_owner_7 ///
PO_us_cit_0 PO_us_cit_1 PO_us_cit_2 PO_us_cit_3 PO_work_exp_0

mi varying

saveold Cross_Sectional_wide_MI_Long_w1,replace
```

3.2.3.2.2. Stata Code: Longitudinal in Wide Format

The following Stata code will create the same variables we discussed in sections 3.2.2.2.1, 3.2.2.2.2, and 3.2.2.2.3, using the KFS multiply imputed data file "KFS8_Longitudinal_wide_MI_Long." The code will save the new file under the name "KFS8_Longitudinal_wide_MI_Long_w1.dta" (n=3140*6).

For the Longitudinal file, last observation carried forward is used to fill in all the fixed core set of questions for businesses reported as temporarily stopped or located (classf=0 or 5); thus, those businesses will have data for all the fixed core set of questions even when classf=0 or 5.

```
clear
clear mata
clear matrix
capture log close
set more off
set maxvar 32767
program drop _all
adopath + " xxx:\KFS_Manual_and_Data\Farhat_Robb_Commands\"
cd "Xxx:\KFS_Manual_and_Data"
use KFS8_Longitudinal_wide_MI_Long,clear

/*For the Longitudinal file, last observation carried forward is used to fill in all the fixed core set of questions
for businesses reported being
temporarily stopped or located (classf=0 or 5), thus those businesses will have data for all the fixed core set of
questions even when classf=0 or 5.*/

forvalues i = 0/7 {
recode classf_`i' (0=6 ) (5=6) ,into (classf_L_`i')
}
forvalues i = 0/7 {
egen   Tot_Equity_Owner_Operators_`i'=  rowtotal(     f2_owner_amt_eq_invest_*_`i'      )      , missing
egen   Tot_Equity_OwnerOper_AllYrs_`i'=rowtotal(      f2_ownr_amt_eqinvest_allyrs_*_`i' ), missing
}

forvalues i = 0/7 {
foreach ow in $owners_1_15  {
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Tot_Equity_Owner_Operators_`i' =.a if classf_L_`i'<6
```

```
replace Tot_Equity_OwnerOper_AllYrs_`i'=.a if classf_L_`i'<6
/*To prevent super-varying variables, Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Equity_Owner_Operators_`i' =.   if  f2_owner_amt_eq_invest_`ow'_`i'==.
replace Tot_Equity_OwnerOper_AllYrs_`i'=.   if  f2_ownr_amt_eqinvest_allyrs_`ow'_`i'==.
}
}

global List1 "spouse parents angels companies govt vent_cap other"

forvalues i = 0/7 {

egen  Tot_Equity_NonOwnerOperators_`i'=rowtotal(f4_eq_amt_angels_`i' f4_eq_amt_companies_`i' f4_eq_amt_govt_`i' ///
f4_eq_amt_other_`i' f4_eq_amt_parents_`i' f4_eq_amt_spouse_`i'  f4_eq_amt_vent_cap_`i')      , missing
egen  Tot_Equity_NonOwnerOp_AllYrs_`i'=rowtotal(f4_eq_amt_*_allyrs_`i')  , missing
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Tot_Equity_NonOwnerOperators_`i' =.a if classf_L_`i'<6
replace Tot_Equity_NonOwnerOperators_`i' =.a if  c1z2_legal_status_`i'==1

replace Tot_Equity_NonOwnerOp_AllYrs_`i' =.a if classf_L_`i'<6
replace Tot_Equity_NonOwnerOp_AllYrs_`i' =.a if  c1z2_legal_status_`i'==1
}

forvalues i = 0/7 {
foreach name in $List1  {
/*To prevent super-varying variables, Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Equity_NonOwnerOperators_`i' =. if f4_eq_amt_`name'_`i'==.
replace Tot_Equity_NonOwnerOp_AllYrs_`i' =. if f4_eq_amt_`name'_allyrs_`i'==.

}
}

forvalues i = 0/7 {
egen  Tot_Equity_`i'       =rowtotal(Tot_Equity_Owner_Operators_`i' Tot_Equity_NonOwnerOperators_`i'), missing
egen  Tot_Equity_AllYrs_`i'=rowtotal(Tot_Equity_OwnerOper_AllYrs_`i' Tot_Equity_NonOwnerOp_AllYrs_`i'), missing
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Tot_Equity_`i'       =.a if classf_L_`i'<6
replace Tot_Equity_AllYrs_`i'=.a if classf_L_`i'<6
/*To prevent super-varying variables, Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Equity_`i'       =. if Tot_Equity_Owner_Operators_`i'==.
replace Tot_Equity_`i'       =. if Tot_Equity_NonOwnerOperators_`i'==.
replace Tot_Equity_AllYrs_`i'=. if Tot_Equity_NonOwnerOperators_`i'==.
replace Tot_Equity_AllYrs_`i'=. if Tot_Equity_NonOwnerOp_AllYrs_`i'==.
```

```
}

forvalues i = 0/7 {
egen  Tot_Assets_`i'=rowtotal(f29_assetval_*_`i')    , missing
egen  Tot_Liab_`i'=rowtotal(f31_value_*_`i'), missing
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Tot_Assets_`i' =.a if classf_L_`i'<6
replace Tot_Liab_`i'   =.a if classf_L_`i'<6
/*To prevent super-varying variables, Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Assets_`i' =. if f29_assetval_acctrec_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_cash_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_equip_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_inv_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_landbuild_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_othbusprop_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_other_`i'==.
replace Tot_Assets_`i' =. if f29_assetval_veh_`i'==.
replace Tot_Liab_`i' =. if  f31_value_acctpay_`i'==.
replace Tot_Liab_`i' =. if  f31_value_other_`i'==.
replace Tot_Liab_`i' =. if  f31_value_pension_`i'==.
}

forvalues i = 0/7 {
egen  Tot_Pers_Debt_Resp_`i'=rowtotal(f8b_pers_credcard_bal_`i' f8b_bus_credcard_bal_`i' f8c_pers_loan_bank_amt_`i'
f8c_pers_loan_fam_amt_`i' f8c_pers_loan_other_amt_`i' f8c_pers_other_amt_`i'),missing
egen  Tot_Pers_Debt_Owed_Resp_`i'=rowtotal(f8b_pers_credcard_bal_`i' f8b_bus_credcard_bal_`i'
f8d_pers_loan_bank_owed_`i' f8d_pers_loan_fam_owed_`i' f8d_pers_loan_other_owed_`i' f8d_pers_other_owed_`i'),missing
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Tot_Pers_Debt_Resp_`i'      =.a if classf_L_`i'<6
replace Tot_Pers_Debt_Owed_Resp_`i' =.a if classf_L_`i'<6
/*To prevent super-varying variables, Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Pers_Debt_Resp_`i' =. if f8b_pers_credcard_bal_`i'==.
replace Tot_Pers_Debt_Resp_`i' =. if f8b_bus_credcard_bal_`i'==.
replace Tot_Pers_Debt_Resp_`i' =. if f8c_pers_loan_bank_amt_`i'==.
replace Tot_Pers_Debt_Resp_`i' =. if f8c_pers_loan_fam_amt_`i'==.
replace Tot_Pers_Debt_Resp_`i' =. if f8c_pers_loan_other_amt_`i'==.
replace Tot_Pers_Debt_Resp_`i' =. if f8c_pers_other_amt_`i'==.
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8b_pers_credcard_bal_`i'==.
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8b_bus_credcard_bal_`i'==.
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8d_pers_loan_bank_owed_`i'==.
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8d_pers_loan_fam_owed_`i'==.
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8d_pers_loan_other_owed_`i'==.
```

```
replace Tot_Pers_Debt_Owed_Resp_`i' =. if f8d_pers_other_owed_`i'==.
}

forvalues i = 0/7 {
egen  Tot_Pers_Debt_Other_Owners_`i'=rowtotal(f10b_pers_credcard_bal_`i' f10c_pers_loan_bank_amt_`i'
f10b_bus_credcard_bal_`i' f10c_pers_loan_fam_amt_`i' f10c_pers_loan_other_amt_`i'  f10c_pers_other_amt_`i'),missing
egen  Tot_Pers_Debt_Owed_OthrOwnrs_`i'=rowtotal(f10b_pers_credcard_bal_`i' f10b_bus_credcard_bal_`i'
f10d_pers_loan_bank_owed_`i' f10d_pers_loan_fam_owed_`i' f10d_pers_loan_other_owed_`i'
f10d_pers_other_owed_`i'),missing
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Tot_Pers_Debt_Other_Owners_`i'  =.a if classf_L_`i'<6
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i'=.a if classf_L_`i'<6
replace Tot_Pers_Debt_Other_Owners_`i'  =.a if c4_numowners_confirm_`i'<2
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i'=.a if c4_numowners_confirm_`i'<2
/*To prevent super-varying variables, Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10b_pers_credcard_bal_`i'==.
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10b_bus_credcard_bal_`i'==.
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10c_pers_loan_bank_amt_`i'==.
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10c_pers_loan_fam_amt_`i'==.
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10c_pers_loan_other_amt_`i'==.
replace Tot_Pers_Debt_Other_Owners_`i' =. if f10c_pers_other_amt_`i'==.
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10b_pers_credcard_bal_`i'==.
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10b_bus_credcard_bal_`i'==.
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10d_pers_loan_bank_owed_`i'==.
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10d_pers_loan_fam_owed_`i'==.
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10d_pers_loan_other_owed_`i'==.
replace Tot_Pers_Debt_Owed_OthrOwnrs_`i' =. if f10d_pers_other_owed_`i'==.
}

forvalues i = 0/7 {
egen  Tot_Debt_Owner_Operators_`i'=rowtotal(Tot_Pers_Debt_Resp_`i' Tot_Pers_Debt_Other_Owners_`i'),missing
egen  Tot_Debt_Owed_Owner_Operators_`i'=rowtotal(Tot_Pers_Debt_Owed_Resp_`i' Tot_Pers_Debt_Owed_OthrOwnrs_`i'),missing
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Tot_Debt_Owner_Operators_`i'   =.a if classf_L_`i'<6
replace Tot_Debt_Owed_Owner_Operators_`i'=.a if classf_L_`i'<6
/*To prevent super-varying variables, Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Debt_Owner_Operators_`i'   =. if Tot_Pers_Debt_Resp_`i'==.
replace Tot_Debt_Owner_Operators_`i'   =. if Tot_Pers_Debt_Other_Owners_`i'==.
replace Tot_Debt_Owed_Owner_Operators_`i'=. if Tot_Pers_Debt_Owed_Resp_`i'==.
replace Tot_Debt_Owed_Owner_Operators_`i'=. if Tot_Pers_Debt_Owed_OthrOwnrs_`i'==.
}
```

```
forvalues i = 0/7 {
egen Tot_Debt_Bus_`i'=rowtotal(f12b_bus_credcard_bal_`i' f12c_bus_loans_bank_amt_`i' f12b_bus_cred_line_bal_`i'
f12c_bus_loans_nonbank_amt_`i'   f12c_bus_loans_fam_amt_`i' f12c_bus_loans_govt_amt_`i' f12c_bus_loans_emp_amt_`i'
f12c_bus_loans_other_ind_amt_`i' f12c_bus_loans_owner_amt_`i' f12c_bus_loans_bus_amt_`i'
      f12c_bus_other_amt_`i'),missing
egen Tot_Bus_Debt_Owed_`i'=rowtotal(f12b_bus_cred_line_bal_`i'  f12b_bus_credcard_bal_`i' f12d_bus_loans_bank_owed_`i'
f12d_bus_loans_nonbank_owed_`i' f12d_bus_loans_emp_owed_`i'  f12d_bus_loans_fam_owed_`i' f12d_bus_loans_govt_owed_`i'
f12d_bus_loans_other_ind_owed_`i' f12d_bus_loans_owner_owed_`i'  f12d_bus_loans_bus_owed_`i'
f12d_bus_other_owed_`i'),missing
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Tot_Debt_Bus_`i'     =.a if classf_L_`i'<6
replace Tot_Bus_Debt_Owed_`i'=.a if classf_L_`i'<6
/*To prevent super-varying variables, Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Debt_Bus_`i'   =. if f12b_bus_credcard_bal_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12b_bus_cred_line_bal_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_bank_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_nonbank_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_fam_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_govt_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_emp_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_other_ind_amt_`i'==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_owner_amt_`i' ==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_loans_bus_amt_`i' ==.
replace Tot_Debt_Bus_`i'   =. if f12c_bus_other_amt_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12b_bus_credcard_bal_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12b_bus_cred_line_bal_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_bank_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_nonbank_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_fam_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_govt_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_emp_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_other_ind_owed_`i'==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_owner_owed_`i' ==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_loans_bus_owed_`i'  ==.
replace Tot_Bus_Debt_Owed_`i'   =. if f12d_bus_other_owed_`i'==.
}

forvalues i = 0/7 {
egen  Tot_Debt_`i'=rowtotal(Tot_Debt_Owner_Operators_`i' Tot_Debt_Bus_`i'),missing
egen  Tot_Debt_Owed_`i'=rowtotal(Tot_Debt_Owed_Owner_Operators_`i' Tot_Bus_Debt_Owed_`i'),missing
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Tot_Debt_`i'      =.a if classf_L_`i'<6
```

```
replace Tot_Debt_Owed_`i' =.a if classf_L_`i'<6
/*To prevent super-varying variables, Recode to soft missing value if any of the total's component is soft missing*/
replace Tot_Debt_`i'   =. if Tot_Debt_Owner_Operators_`i'==.
replace Tot_Debt_`i'   =. if Tot_Debt_Bus_`i'==.
replace Tot_Debt_Owed_`i'=. if Tot_Debt_Owed_Owner_Operators_`i'==.
replace Tot_Debt_Owed_`i'=. if Tot_Bus_Debt_Owed_`i'==.
}

forvalues i = 0/7 {
gen  Net_Profit_`i'=f24_profitloss_amt_`i'
}

* Long names create problems in MI data file. Make the names shorter
rename Tot_* *

*Stata required imputed variables to be registered, it is only recommend that you register passive or regular
variables.
/*Optional

 mi register passive Net_Profit_0 Net_Profit_1 Net_Profit_2 Net_Profit_3 Net_Profit_4 Net_Profit_5 Net_Profit_6
Net_Profit_7 ///
Assets_0 Assets_1 Assets_2 Assets_3 Assets_4 Assets_5 Assets_6 Assets_7  ///
Liab_0 Liab_1 Liab_2 Liab_3 Liab_4 Liab_5 Liab_6 Liab_7  ///
Pers_Debt_Owed_Resp_0 Pers_Debt_Owed_Resp_1 Pers_Debt_Owed_Resp_2 Pers_Debt_Owed_Resp_3 Pers_Debt_Owed_Resp_4
Pers_Debt_Owed_Resp_5 Pers_Debt_Owed_Resp_6 Pers_Debt_Owed_Resp_7  ///
Pers_Debt_Resp_0 Pers_Debt_Resp_1 Pers_Debt_Resp_2 Pers_Debt_Resp_3 Pers_Debt_Resp_4 Pers_Debt_Resp_5 Pers_Debt_Resp_6
Pers_Debt_Resp_7
*/

/* Do not register super varying variables*/
/*Super  varying variables

Bus_Debt_Owed_0 Bus_Debt_Owed_1 Bus_Debt_Owed_2 Bus_Debt_Owed_3 Bus_Debt_Owed_4 Bus_Debt_Owed_5 Bus_Debt_Owed_6
Bus_Debt_Owed_7
Debt_0 Debt_1 Debt_2 Debt_3 Debt_4 Debt_5 Debt_6 Debt_7 ///
Debt_Bus_0 Debt_Bus_1 Debt_Bus_2 Debt_Bus_3 Debt_Bus_4 Debt_Bus_5 Debt_Bus_6 Debt_Bus_7
Debt_Owed_0 Debt_Owed_1 Debt_Owed_2 Debt_Owed_3 Debt_Owed_4 Debt_Owed_5 Debt_Owed_6 Debt_Owed_7
Debt_Owed_Owner_Operators_0 Debt_Owed_Owner_Operators_1 Debt_Owed_Owner_Operators_2 Debt_Owed_Owner_Operators_3
Debt_Owed_Owner_Operators_4 Debt_Owed_Owner_Operators_5 Debt_Owed_Owner_Operators_6
Debt_Owed_Owner_Operators_7 Debt_Owner_Operators_0 Debt_Owner_Operators_1 Debt_Owner_Operators_2 Debt_Owner_Operators_3
Debt_Owner_Operators_4 Debt_Owner_Operators_5 Debt_Owner_Operators_6 Debt_Owner_Operators_7
Equity_OwnerOper_AllYrs_0 Equity_OwnerOper_AllYrs_1Equity_OwnerOper_AllYrs_2 Equity_OwnerOper_AllYrs_3
```

```
Equity_OwnerOper_AllYrs_4 Equity_OwnerOper_AllYrs_5 Equity_OwnerOper_AllYrs_6 Equity_OwnerOper_AllYrs_7
Equity_Owner_Operators_0 Equity_Owner_Operators_1 Equity_Owner_Operators_2 Equity_Owner_Operators_3
Equity_Owner_Operators_4 Equity_Owner_Operators_5 Equity_Owner_Operators_6 Equity_Owner_Operators_7
Pers_Debt_Other_Owners_0 Pers_Debt_Other_Owners_1 Pers_Debt_Other_Owners_2 Pers_Debt_Other_Owners_3
Pers_Debt_Other_Owners_4 Pers_Debt_Other_Owners_5 Pers_Debt_Other_Owners_6 Pers_Debt_Other_Owners_7
Pers_Debt_Owed_OthrOwnrs_0 Pers_Debt_Owed_OthrOwnrs_1 Pers_Debt_Owed_OthrOwnrs_2 Pers_Debt_Owed_OthrOwnrs_3
Pers_Debt_Owed_OthrOwnrs_4 Pers_Debt_Owed_OthrOwnrs_5 Pers_Debt_Owed_OthrOwnrs_6 Pers_Debt_Owed_OthrOwnrs_7
Equity_0 Equity_1 Equity_2 Equity_3 Equity_4 Equity_5 Equity_6 Equity_7
Equity_AllYrs_0 Equity_AllYrs_1 Equity_AllYrs_2 Equity_AllYrs_3 Equity_AllYrs_4 Equity_AllYrs_5 Equity_AllYrs_6
Equity_AllYrs_7
Equity_NonOwnerOp_AllYrs_0 Equity_NonOwnerOp_AllYrs_1 Equity_NonOwnerOp_AllYrs_2 Equity_NonOwnerOp_AllYrs_3
Equity_NonOwnerOp_AllYrs_4 Equity_NonOwnerOp_AllYrs_5 Equity_NonOwnerOp_AllYrs_6 Equity_NonOwnerOp_AllYrs_7
Equity_NonOwnerOperators_0 Equity_NonOwnerOperators_1 Equity_NonOwnerOperators_2 Equity_NonOwnerOperators_3
Equity_NonOwnerOperators_4 Equity_NonOwnerOperators_5 Equity_NonOwnerOperators_6 Equity_NonOwnerOperators_7
*/

/*****************************************************************/
/* Merge the file with the primary owner file “primary_owner.dta” */

merge m:1 mprid using "Xxx:\KFS_Manual_and_Data\primary_owner.dta"
drop if _merge!=3
drop _merge


* Make sure that the race do not chnage over time due to entry errors
forvalues i = 1/7 {
foreach ow in $owners_1_15  {

replace      g6_race_amind_owner_`ow'_`i'      =     g6_race_amind_owner_`ow'_0 if     g6_race_amind_owner_`ow'_`i'
      <.     &      g6_race_amind_owner_`ow'_0<.
replace      g6_race_asian_owner_`ow'_`i'      =     g6_race_asian_owner_`ow'_0 if     g6_race_asian_owner_`ow'_`i'
      <.     &      g6_race_asian_owner_`ow'_0<.
replace      g6_race_black_owner_`ow'_`i'      =     g6_race_black_owner_`ow'_0 if     g6_race_black_owner_`ow'_`i'
      <.     &      g6_race_black_owner_`ow'_0<.
replace      g6_race_nathaw_owner_`ow'_`i'     =     g6_race_nathaw_owner_`ow'_0       if
      g6_race_nathaw_owner_`ow'_`i'     <.     &     g6_race_nathaw_owner_`ow'_0<.
replace      g6_race_other_owner_`ow'_`i'      =     g6_race_other_owner_`ow'_0 if     g6_race_other_owner_`ow'_`i'
      <.     &      g6_race_other_owner_`ow'_0<.
replace      g6_race_white_owner_`ow'_`i'      =     g6_race_white_owner_`ow'_0 if     g6_race_white_owner_`ow'_`i'
      <.     &      g6_race_white_owner_`ow'_0<.
}
}
```

```
* Primary Owner(PO) Characteristics at the Baseline: PO by Robb's Stata code

forvalues i = 0/0 {
gen  PO_emp_`i'                =.
gen  PO_hours_`i'              =.
gen  PO_work_exp_`i'              =.
gen  PO_oth_bus_owner_`i'      =.
gen  PO_bus_same_ind_`i'       =.
gen  PO_age_owner_`i'             =.
gen  PO_hisp_origin_`i'       =.
gen  PO_race_group_`i'        =.
gen  PO_race_amind_owner_`i'   =.
gen  PO_race_asian_owner_`i'   =.
gen  PO_race_black_owner_`i'   =.
gen  PO_race_nathaw_owner_`i'  =.
gen  PO_race_other_owner_`i'   =.
gen  PO_race_white_owner_`i'   =.
gen  PO_native_born_`i'       =.
gen  PO_us_cit_`i'             =.
gen  PO_education_`i'             =.
gen  PO_gender_`i'             =.
}

forvalues i = 0/0 {
forvalues po = 1/6 {
replace PO_emp_`i'          =   g1a_emp_owner_0`po'_`i'             if  primary_owner==`po'
replace PO_hours_`i'           =  g1b1_hours_owner_0`po'_`i'                if  primary_owner==`po'
replace PO_work_exp_`i'    =     g2_work_exp_owner_0`po'_`i'             if  primary_owner==`po'
replace PO_oth_bus_owner_`i'=    g3a_oth_bus_owner_0`po'_`i'             if  primary_owner==`po'
replace PO_bus_same_ind_`i' =    g3b_bus_same_ind_owner_0`po'_`i'  if  primary_owner==`po'
replace PO_age_owner_`i'   =     g4_age_owner_0`po'_`i'                  if  primary_owner==`po'
replace PO_hisp_origin_`i'       =     g5_hisp_origin_owner_0`po'_`i'       if  primary_owner==`po'
replace PO_race_group_`i'   =    g6b_race_group_0`po'_`i'              if  primary_owner==`po'
replace PO_race_amind_owner_`i' = g6_race_amind_owner_0`po'_`i'     if  primary_owner==`po'
replace PO_race_asian_owner_`i' = g6_race_asian_owner_0`po'_`i'     if  primary_owner==`po'
replace PO_race_black_owner_`i' = g6_race_black_owner_0`po'_`i'     if  primary_owner==`po'
replace PO_race_nathaw_owner_`i'= g6_race_nathaw_owner_0`po'_`i'    if  primary_owner==`po'
replace PO_race_other_owner_`i' = g6_race_other_owner_0`po'_`i'     if  primary_owner==`po'
replace PO_race_white_owner_`i' = g6_race_white_owner_0`po'_`i'     if  primary_owner==`po'
replace PO_native_born_`i'       =     g7_native_born_owner_0`po'_`i'       if  primary_owner==`po'
replace PO_us_cit_`i'          =  g8_us_cit_owner_0`po'_`i'            if  primary_owner==`po'
```

```
replace PO_education_`i'   =      g9_education_owner_0`po'_`i'            if  primary_owner==`po'
replace PO_gender_`i'          =  g10_gender_owner_0`po'_`i'           if  primary_owner==`po'
}
}

forvalues i = 1/7 {
gen  PO_emp_`i'          = PO_emp_0
gen  PO_hours_`i'        = PO_hours_0
gen  PO_work_exp_`i'       = PO_work_exp_0
gen  PO_oth_bus_owner_`i'= PO_oth_bus_owner_0
gen  PO_bus_same_ind_`i' = PO_bus_same_ind_0
gen  PO_age_owner_`i'      = PO_age_owner_0
gen  PO_hisp_origin_`i'  = PO_hisp_origin_0
gen  PO_race_group_`i'   = PO_race_group_0
gen  PO_race_amind_owner_`i' =PO_race_amind_owner_0
gen  PO_race_asian_owner_`i' =PO_race_asian_owner_0
gen  PO_race_black_owner_`i' =PO_race_black_owner_0
gen  PO_race_nathaw_owner_`i'=PO_race_nathaw_owner_0
gen  PO_race_other_owner_`i' =PO_race_other_owner_0
gen  PO_race_white_owner_`i' =PO_race_white_owner_0
gen  PO_native_born_`i'  = PO_native_born_0
gen  PO_us_cit_`i'         = PO_us_cit_0
gen  PO_education_`i'      = PO_education_0
gen  PO_gender_`i'       = PO_gender_0
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace  PO_emp_`i'               = .a if    classf_L_`i'  <6
replace  PO_hours_`i'             = .a if    classf_L_`i'  <6
replace  PO_work_exp_`i'       = .a if classf_L_`i'  <6
replace  PO_oth_bus_owner_`i'     = .a if     classf_L_`i'  <6
replace  PO_bus_same_ind_`i'      = .a if     classf_L_`i'  <6
replace  PO_age_owner_`i'      = .a if classf_L_`i'  <6
replace  PO_hisp_origin_`i'       = .a if     classf_L_`i'  <6
replace  PO_race_group_`i'        = .a if     classf_L_`i'  <6
replace  PO_race_amind_owner_`i' = .a if      classf_L_`i'  <6
replace  PO_race_asian_owner_`i' = .a if      classf_L_`i'  <6
replace  PO_race_black_owner_`i' = .a if      classf_L_`i'  <6
replace  PO_race_nathaw_owner_`i'= .a if      classf_L_`i'  <6
replace  PO_race_other_owner_`i' = .a if      classf_L_`i'  <6
replace  PO_race_white_owner_`i' = .a if      classf_L_`i'  <6
replace  PO_native_born_`i'       = .a if     classf_L_`i'  <6
replace  PO_us_cit_`i'              = .a if    classf_L_`i'  <6
replace  PO_education_`i'      = .a if classf_L_`i'  <6
```

```
replace  PO_gender_`i'              = .a if    classf_L_`i'  <6

}

* Active-Owner-Operators Characteristics (OO)

forvalues i = 0/7 {
egen   OO_emp_owner_`i'          =      rowmean(g1a_emp_owner_*_`i')
egen   OO_hours_owner_`i'        = rowmean(g1b1_hours_owner_*_`i')
egen   OO_work_exp_owner_`i'     =      rowmean(g2_work_exp_owner_*_`i'   )
egen   OO_oth_bus_owner_`i'      =      rowmean(g3a_oth_bus_owner_*_`i'   )
egen   OO_bus_same_ind_owner_`i' =      rowmean(g3b_bus_same_ind_owner_*_`i')
egen   OO_age_owner_`i'          =      rowmean(g4_age_owner_*_`i')
egen   OO_hisp_origin_owner_`i'  =      rowmean(g5_hisp_origin_owner_*_`i')
egen   OO_race_amind_owner_`i'   =      rowmean(g6_race_amind_owner_*_`i')
egen   OO_race_asian_owner_`i'   =      rowmean(g6_race_asian_owner_*_`i')
egen   OO_race_black_owner_`i'   =      rowmean(g6_race_black_owner_*_`i')
egen   OO_race_nathaw_owner_`i'  =      rowmean(g6_race_nathaw_owner_*_`i')
egen   OO_race_other_owner_`i'   =      rowmean(g6_race_other_owner_*_`i')
egen   OO_race_white_owner_`i'   =      rowmean(g6_race_white_owner_*_`i')
egen   OO_native_born_owner_`i'  =      rowmean(g7_native_born_owner_*_`i'      )
egen   OO_us_cit_owner_`i'       =      rowmean(g8_us_cit_owner_*_`i'     )
egen   OO_education_owner_`i'    =      rowmean(g9_education_owner_*_`i'  )
egen   md_education_owner_`i'    =      rowmedian(g9_education_owner_*_`i')
gen    OO_D_education_owner_`i'  =(md_education_owner_`i'    >6.99)        if      md_education_owner_`i'    <11

egen   OO_gender_owner_`i'       =      rowmean(g10_gender_owner_*_`i'    )
}

/*To prevent super-varying variables, Recode to soft missing value if any of the total's component is soft missing*/
forvalues i = 0/7 {
foreach ow in $owners_1_15  {
replace OO_emp_owner_`i'              =. if g1a_emp_owner_`ow'_`i'==.
replace OO_hours_owner_`i'         =. if g1b1_hours_owner_`ow'_`i'==.
replace OO_work_exp_owner_`i'     =. if g2_work_exp_owner_`ow'_`i'==.
replace OO_oth_bus_owner_`i'      =. if g3a_oth_bus_owner_`ow'_`i'==.
replace OO_bus_same_ind_owner_`i' =. if g3b_bus_same_ind_owner_`ow'_`i'==.
replace OO_age_owner_`i'               =. if g4_age_owner_`ow'_`i'==.
replace OO_hisp_origin_owner_`i'  =. if g5_hisp_origin_owner_`ow'_`i'==.
replace OO_race_amind_owner_`i'   =. if g6_race_amind_owner_`ow'_`i'==.
replace OO_race_asian_owner_`i'   =. if g6_race_asian_owner_`ow'_`i'==.
replace OO_race_black_owner_`i'   =. if g6_race_black_owner_`ow'_`i'==.
```

```
replace OO_race_nathaw_owner_`i'  =. if g6_race_nathaw_owner_`ow'_`i'==.
replace OO_race_other_owner_`i'   =. if g6_race_other_owner_`ow'_`i'==.
replace OO_race_white_owner_`i'   =. if g6_race_white_owner_`ow'_`i'==.
replace OO_native_born_owner_`i'  =. if g7_native_born_owner_`ow'_`i'==.
replace OO_us_cit_owner_`i'       =. if g8_us_cit_owner_`ow'_`i'==.
replace OO_education_owner_`i'     =. if g9_education_owner_`ow'_`i'==.
replace md_education_owner_`i'     =. if g9_education_owner_`ow'_`i'==.
replace OO_D_education_owner_`i'  =. if g9_education_owner_`ow'_`i'==.
replace OO_gender_owner_`i'       =. if g10_gender_owner_`ow'_`i'==.
}
}

/*To prevent super-varying variables, Recode legitimate (hard) missing values */
forvalues i = 0/7 {
replace      OO_emp_owner_`i'          = .a if classf_L_`i'  <6
replace      OO_hours_owner_`i'        = .a if classf_L_`i'  <6
replace      OO_work_exp_owner_`i'     = .a if classf_L_`i'  <6
replace      OO_oth_bus_owner_`i'      = .a if classf_L_`i'  <6
replace      OO_bus_same_ind_owner_`i' = .a if classf_L_`i'  <6
replace      OO_age_owner_`i'          = .a if classf_L_`i'  <6
replace      OO_hisp_origin_owner_`i'  = .a if classf_L_`i'  <6
replace      OO_race_amind_owner_`i'   = .a if classf_L_`i'  <6
replace      OO_race_asian_owner_`i'   = .a if classf_L_`i'  <6
replace      OO_race_black_owner_`i'   = .a if classf_L_`i'  <6
replace      OO_race_nathaw_owner_`i'  = .a if classf_L_`i'  <6
replace      OO_race_other_owner_`i'   = .a if classf_L_`i'  <6
replace      OO_race_white_owner_`i'   = .a if classf_L_`i'  <6
replace      OO_native_born_owner_`i'  = .a if classf_L_`i'  <6
replace      OO_us_cit_owner_`i'       = .a if classf_L_`i'  <6
replace      OO_education_owner_`i'    = .a if classf_L_`i'  <6
replace      md_education_owner_`i'    = .a if classf_L_`i'  <6
replace      OO_D_education_owner_`i'  = .a if classf_L_`i'  <6

replace      OO_gender_owner_`i'       = .a if classf_L_`i'  <6
}

/*******************************************************************************/
*Diversity / Similarity index
forvalues i = 0/7 {
gen xr1_`i'= OO_race_amind_owner_`i'    *      OO_race_amind_owner_`i'
gen xr2_`i'= OO_race_asian_owner_`i'    *      OO_race_asian_owner_`i'
gen xr3_`i'= OO_race_black_owner_`i'    *      OO_race_black_owner_`i'
```

```
gen xr4_`i'= OO_race_nathaw_owner_`i'*  OO_race_nathaw_owner_`i'
gen xr5_`i'= OO_race_other_owner_`i' *  OO_race_other_owner_`i'
gen xr6_`i'= OO_race_white_owner_`i'    *      OO_race_white_owner_`i'
* Race_similarity
egen Race_similarity_`i'=rowtotal(xr1_`i' xr2_`i' xr3_`i' xr4_`i' xr5_`i' xr6_`i'),missing
drop xr*_`i'
* Race_ diversity
gen Race_diversity_`i'=1-Race_similarity_`i'
}

forvalues i = 0/7 {
foreach ow in $owners_1_15  {
/* Recode to soft missing value if any of the var's component is soft missing*/
replace Race_similarity_`i'   =. if g6_race_amind_owner_`ow'_`i'==.
replace Race_similarity_`i'   =. if g6_race_asian_owner_`ow'_`i'==.
replace Race_similarity_`i'   =. if g6_race_black_owner_`ow'_`i'==.
replace Race_similarity_`i'   =. if g6_race_nathaw_owner_`ow'_`i'==.
replace Race_similarity_`i'   =. if g6_race_other_owner_`ow'_`i'==.
replace Race_similarity_`i'   =. if g6_race_white_owner_`ow'_`i'==.

}
}
forvalues i = 0/7 {
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Race_similarity_`i'   =.a if classf_L_`i'<6
replace Race_diversity_`i'   =.a if Race_similarity_`i'==.a
replace Race_diversity_`i'   =.  if Race_similarity_`i'==.
}

forvalues i = 0/7 {
gen fmal_`i'=1-OO_gender_owner_`i'
gen xr1_`i'= OO_gender_owner_`i' *      OO_gender_owner_`i'
gen xr2_`i'= fmal_`i'     *      fmal_`i'

* Gender_similarity
egen Gender_similarity_`i'=rowtotal(xr1_`i' xr2_`i'  ),missing
drop xr*_`i' fmal_`i'
* Gender_diversity
gen Gender_diversity_`i'=1-Gender_similarity_`i'
}

forvalues i = 0/7 {
```

```
foreach ow in $owners_1_15  {
/* Recode to soft missing value if any of the var's component is soft missing*/
replace  Gender_similarity_`i'      =. if g10_gender_owner_`ow'_`i'==.
}
}

forvalues i = 0/7 {
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Gender_similarity_`i'   =.a if classf_L_`i'<6
replace Gender_diversity_`i'   =.a if Gender_similarity_`i'==.a
replace Gender_diversity_`i'   =.  if Gender_similarity_`i'==.
}

forvalues i = 0/7 {
mean Race_similarity_`i' Gender_similarity_`i' if c4_numowners_confirm_`i'>1 & c4_numowners_confirm_`i'<.
}

*Business level Characteristics
forvalues i = 0/7 {
*Home Based Dummy
recode c8_primary_loc_`i' (1=1 "Home Based") (nonmiss=0 "Non Home Based" ) ,into (Home_Based_`i')
*Sole_Proprietorship Dummy
recode c1z2_legal_status_`i' (1=1 "Sole_Proprietorship") (nonmiss=0 "Limited Liability" ) ,into
(Sole_Proprietorship_`i')
gen Comp_advantage_`i' = d2_comp_advantage_`i'
drop d2_comp_advantage_`i'
egen Have_IP_`i'=anymatch(d3_a_have_patent_`i' d3_b_have_copyright_`i' d3_c_have_trademark_`i'), values(1)
gen Full_Part_Time_Employees_`i'= c5_num_employees_`i'
gen Full_Time_Employees_`i' =c6_num_ft_employees_`i'
gen Part_Time_Employees_`i'=c7_num_pt_employees_`i'
drop c5_num_employees_`i'  c6_num_ft_employees_`i'  c7_num_pt_employees_`i'
egen   Employee_Owner_`i'          =    rowtotal(g1a_emp_owner_*_`i')
egen   Total_Employees_`i'=rowtotal(Employee_Owner_`i' Full_Part_Time_Employees_`i')
}

forvalues i = 0/7 {
foreach ow in $owners_1_15  {
/* Recode to soft missing value if any of the var's component is soft missing*/
replace Have_IP_`i'=. if d3_a_have_patent_`i'==.   | d3_b_have_copyright_`i'==.   | d3_c_have_trademark_`i'==.
replace Employee_Owner_`i'             =. if g1a_emp_owner_`ow'_`i'==.
replace Total_Employees_`i'            =. if g1a_emp_owner_`ow'_`i'==.
replace Total_Employees_`i'            =. if  Full_Part_Time_Employees_`i'==.
```

```
/*To prevent super-varying variables, Recode legitimate (hard) missing values */
replace Have_IP_`i'        =.a if classf_L_`i'<6
replace Employee_Owner_`i' =.a if classf_L_`i'<6
replace Total_Employees_`i'=.a if classf_L_`i'<6
}
}

*Stata required imputed variables to be registered, it is only recommend that you register passive or regular
variables.
/*Optional

mi register passive Full_Part_Time_Employees* Full_Time_Employees* Part_Time_Employees*  Comp_advantage*
mi register passive PO_age_owner_3 PO_age_owner_4 PO_age_owner_5 ///
PO_age_owner_6 PO_age_owner_7 PO_bus_same_ind_0 PO_bus_same_ind_1 PO_bus_same_ind_2 PO_bus_same_ind_3 PO_bus_same_ind_4
PO_bus_same_ind_5 PO_bus_same_ind_6 ///
PO_bus_same_ind_7 PO_education_0 PO_education_1 PO_education_2 PO_education_3 PO_education_4 PO_education_5 PO_emp_0
PO_emp_1 PO_emp_2 PO_emp_3 PO_emp_4 ///
PO_emp_5 PO_emp_6 PO_emp_7 PO_gender_0 PO_gender_1 PO_gender_2 PO_gender_3 PO_gender_4 PO_gender_5 PO_hisp_origin_0
PO_hisp_origin_1 PO_hisp_origin_2 ///
PO_hisp_origin_3 PO_hisp_origin_4 PO_hisp_origin_5 PO_hisp_origin_6 PO_hisp_origin_7 PO_hours_0 PO_hours_1 PO_hours_2
PO_hours_3 PO_hours_4 PO_hours_5 ///
PO_hours_6 PO_hours_7 PO_native_born_0 PO_native_born_1 PO_native_born_2 PO_native_born_3 PO_native_born_4
PO_native_born_5 PO_native_born_6 PO_native_born_7 ///
PO_oth_bus_owner_0 PO_oth_bus_owner_1 PO_oth_bus_owner_2 PO_oth_bus_owner_3 PO_oth_bus_owner_4 PO_oth_bus_owner_5
PO_oth_bus_owner_6 PO_oth_bus_owner_7 ///
PO_race_amind_owner_0 PO_race_amind_owner_7 PO_race_asian_owner_0 PO_race_asian_owner_7 PO_race_black_owner_0
PO_race_black_owner_7 PO_race_group_0 ///
PO_race_group_7 PO_race_nathaw_owner_0 PO_race_nathaw_owner_7 PO_race_other_owner_0 PO_race_other_owner_7
PO_race_white_owner_0 PO_race_white_owner_7 ///
PO_us_cit_0 PO_us_cit_1 PO_us_cit_2 PO_us_cit_3 PO_work_exp_0

mi register regular Home_Based* Sole_Proprietorship_*

mi register passive Employee_Owner_0 Employee_Owner_1 Employee_Owner_2 Employee_Owner_3 Employee_Owner_4
Employee_Owner_5 ///
Employee_Owner_6 Employee_Owner_7  Gender_diversity_0 Gender_diversity_1 Gender_diversity_2 ///
Gender_diversity_3 Gender_diversity_4 Gender_diversity_5 Gender_diversity_6 Gender_diversity_7 Gender_similarity_0
Gender_similarity_1 Gender_similarity_2 ///
Gender_similarity_3 Gender_similarity_4 Gender_similarity_5 Gender_similarity_6 Gender_similarity_7 Have_IP_0 Have_IP_1
Have_IP_2 Have_IP_3 Have_IP_4 Have_IP_5 ///
Have_IP_6 Have_IP_7 OO_D_education_owner_0 OO_D_education_owner_1 OO_D_education_owner_2 OO_D_education_owner_3
OO_D_education_owner_4 OO_D_education_owner_5 ///
```

```
OO_D_education_owner_6 OO_D_education_owner_7 OO_age_owner_0 OO_age_owner_1 OO_age_owner_2 OO_age_owner_3
OO_age_owner_4 OO_age_owner_5 OO_age_owner_6 ///
OO_age_owner_7 OO_bus_same_ind_owner_0 OO_bus_same_ind_owner_1 OO_bus_same_ind_owner_2 OO_bus_same_ind_owner_3
OO_bus_same_ind_owner_4 OO_bus_same_ind_owner_5 ///
OO_bus_same_ind_owner_6 OO_bus_same_ind_owner_7 OO_education_owner_0 OO_education_owner_1 OO_education_owner_2
OO_education_owner_3 OO_education_owner_4 ///
OO_education_owner_5 OO_education_owner_6 OO_education_owner_7 OO_emp_owner_0 OO_emp_owner_1 OO_emp_owner_2
OO_emp_owner_3 OO_emp_owner_4 OO_emp_owner_5 ///
OO_emp_owner_6 OO_emp_owner_7 OO_gender_owner_0 OO_gender_owner_1 OO_gender_owner_2 OO_gender_owner_3 OO_gender_owner_4
OO_gender_owner_5 OO_gender_owner_6 ///
OO_gender_owner_7 OO_hisp_origin_owner_0 OO_hisp_origin_owner_1 OO_hisp_origin_owner_2 OO_hisp_origin_owner_3
OO_hisp_origin_owner_4 OO_hisp_origin_owner_5 ///
OO_hisp_origin_owner_6 OO_hisp_origin_owner_7 OO_hours_owner_0 OO_hours_owner_1 OO_hours_owner_2 OO_hours_owner_3
OO_hours_owner_4 OO_hours_owner_5 ///
OO_hours_owner_6 OO_hours_owner_7 OO_native_born_owner_0 OO_native_born_owner_1 OO_native_born_owner_2
OO_native_born_owner_3 OO_native_born_owner_4 ///
OO_native_born_owner_5 OO_native_born_owner_6 OO_native_born_owner_7 OO_oth_bus_owner_0 OO_oth_bus_owner_1
OO_oth_bus_owner_2 OO_oth_bus_owner_3 ///
OO_oth_bus_owner_4 OO_oth_bus_owner_5 OO_oth_bus_owner_6 OO_oth_bus_owner_7 OO_race_amind_owner_0 OO_race_amind_owner_1
OO_race_amind_owner_2 ///
OO_race_amind_owner_3 OO_race_amind_owner_4 OO_race_amind_owner_5 OO_race_amind_owner_6 OO_race_amind_owner_7
OO_race_asian_owner_0 OO_race_asian_owner_1 ///
OO_race_asian_owner_2 OO_race_asian_owner_3 OO_race_asian_owner_4 OO_race_asian_owner_5 OO_race_asian_owner_6
OO_race_asian_owner_7 OO_race_black_owner_0 ///
OO_race_black_owner_1 OO_race_black_owner_2 OO_race_black_owner_3 OO_race_black_owner_4 OO_race_black_owner_5
OO_race_black_owner_6 OO_race_black_owner_7 ///
OO_race_nathaw_owner_0 OO_race_nathaw_owner_1 OO_race_nathaw_owner_2 OO_race_nathaw_owner_3 OO_race_nathaw_owner_4
OO_race_nathaw_owner_5 ///
OO_race_nathaw_owner_6 OO_race_nathaw_owner_7 OO_race_other_owner_0 OO_race_other_owner_1 OO_race_other_owner_2
OO_race_other_owner_3 OO_race_other_owner_4 ///
OO_race_other_owner_5 OO_race_other_owner_6 OO_race_other_owner_7 OO_race_white_owner_0 OO_race_white_owner_1
OO_race_white_owner_2 OO_race_white_owner_3 ///
OO_race_white_owner_4 OO_race_white_owner_5 OO_race_white_owner_6 OO_race_white_owner_7 OO_us_cit_owner_0
OO_us_cit_owner_1 OO_us_cit_owner_2 OO_us_cit_owner_3 ///
OO_us_cit_owner_4 OO_us_cit_owner_5 OO_us_cit_owner_6 OO_us_cit_owner_7 OO_work_exp_owner_0 OO_work_exp_owner_1
OO_work_exp_owner_2 OO_work_exp_owner_3 ///
OO_work_exp_owner_4 OO_work_exp_owner_5 OO_work_exp_owner_7 PO_age_owner_0 PO_age_owner_1 PO_age_owner_2
Race_diversity_0 Race_diversity_1 Race_diversity_2 Race_diversity_3 ///
Race_diversity_4 Race_diversity_5 Race_diversity_6 Race_diversity_7 Race_similarity_0 Race_similarity_1
Race_similarity_2 Race_similarity_3 Race_similarity_4 ///
Race_similarity_5 Race_similarity_6 Race_similarity_7 Total_Employees_0 Total_Employees_1 Total_Employees_2
Total_Employees_3 Total_Employees_4 ///
```

```
Total_Employees_5 Total_Employees_6 Total_Employees_7 md_education_owner_0 md_education_owner_1 md_education_owner_2
md_education_owner_3 md_education_owner_4 ///
md_education_owner_5 md_education_owner_6 md_education_owner_7
*/
mi varying



saveold Longitudinal_wide_MI_Long_w1,replace
```

3.2.3.2.3. Stata Code: Cross Sectional in Long Format

The following Stata code will create the same variables we discussed in sections 3.2.2.2.1, 3.2.2.2.2, and 3.2.2.2.3, using the KFS multiply imputed data file "KFS8_Cross_Sectional_Long_MI_Long." The code will save the new file under the name "KFS8_Cross_Sectional_Long_MI_Long _L1.dta" (4928*8*6=236,544 obs.).

```
clear
clear mata
clear matrix
capture log close
set more off
set maxvar 32767
program drop _all
adopath + " xxx:\KFS_Manual_and_Data\Farhat_Robb_Commands\"
cd "Xxx:\KFS_Manual_and_Data"
  use KFS8_Cross_Sectional_Long_MI_Long,clear


egen   Tot_Equity_Owner_Operators =     rowtotal(    f2_owner_amt_eq_invest_*   )      , missing
egen   Tot_Equity_OwnerOper_AllYrs =rowtotal(  f2_ownr_amt_eqinvest_allyrs_*    ), missing

/* Recode legitimate (hard) missing values */
replace Tot_Equity_Owner_Operators  =.a if classf <6
replace Tot_Equity_OwnerOper_AllYrs =.a if classf <6

global List1 "spouse parents angels companies govt vent_cap other"

egen  Tot_Equity_NonOwnerOperators =rowtotal(f4_eq_amt_angels  f4_eq_amt_companies  f4_eq_amt_govt  ///
f4_eq_amt_other  f4_eq_amt_parents  f4_eq_amt_spouse   f4_eq_amt_vent_cap )     , missing
egen  Tot_Equity_NonOwnerOp_AllYrs =rowtotal(f4_eq_amt_*_allyrs )  , missing
/* Recode legitimate (hard) missing values */
replace Tot_Equity_NonOwnerOperators  =.a if classf <6
replace Tot_Equity_NonOwnerOperators  =.a if  c1z2_legal_status ==1
replace Tot_Equity_NonOwnerOp_AllYrs  =.a if classf <6
replace Tot_Equity_NonOwnerOp_AllYrs  =.a if  c1z2_legal_status ==1

egen  Tot_Equity        =rowtotal(Tot_Equity_Owner_Operators  Tot_Equity_NonOwnerOperators ), missing
egen  Tot_Equity_AllYrs =rowtotal(Tot_Equity_OwnerOper_AllYrs  Tot_Equity_NonOwnerOp_AllYrs ), missing
/* Recode legitimate (hard) missing values */
```

```
replace Tot_Equity        =.a if classf <6
replace Tot_Equity_AllYrs =.a if classf <6

egen  Tot_Assets =rowtotal(f29_assetval_* )    , missing
egen  Tot_Liab =rowtotal(f31_value_* ), missing
/* Recode legitimate (hard) missing values */
replace Tot_Assets  =.a if classf <6
replace Tot_Liab    =.a if classf <6

egen  Tot_Pers_Debt_Resp =rowtotal(f8b_pers_credcard_bal  f8b_bus_credcard_bal  f8c_pers_loan_bank_amt
f8c_pers_loan_fam_amt  f8c_pers_loan_other_amt  f8c_pers_other_amt ),missing
egen  Tot_Pers_Debt_Owed_Resp =rowtotal(f8b_pers_credcard_bal  f8b_bus_credcard_bal  f8d_pers_loan_bank_owed
f8d_pers_loan_fam_owed  f8d_pers_loan_other_owed  f8d_pers_other_owed ),missing
/* Recode legitimate (hard) missing values */
replace Tot_Pers_Debt_Resp       =.a if classf <6
replace Tot_Pers_Debt_Owed_Resp  =.a if classf <6

egen  Tot_Pers_Debt_Other_Owners =rowtotal(f10b_pers_credcard_bal  f10c_pers_loan_bank_amt  f10b_bus_credcard_bal
f10c_pers_loan_fam_amt  f10c_pers_loan_other_amt   f10c_pers_other_amt ),missing
egen  Tot_Pers_Debt_Owed_OthrOwnrs =rowtotal(f10b_pers_credcard_bal  f10b_bus_credcard_bal  f10d_pers_loan_bank_owed
f10d_pers_loan_fam_owed  f10d_pers_loan_other_owed  f10d_pers_other_owed ),missing
/* Recode legitimate (hard) missing values */
replace Tot_Pers_Debt_Other_Owners   =.a if classf <6
replace Tot_Pers_Debt_Owed_OthrOwnrs =.a if classf <6
replace Tot_Pers_Debt_Other_Owners   =.a if c4_numowners_confirm <2
replace Tot_Pers_Debt_Owed_OthrOwnrs =.a if c4_numowners_confirm <2

egen  Tot_Debt_Owner_Operators =rowtotal(Tot_Pers_Debt_Resp  Tot_Pers_Debt_Other_Owners ),missing
egen  Tot_Debt_Owed_Owner_Operators =rowtotal(Tot_Pers_Debt_Owed_Resp  Tot_Pers_Debt_Owed_OthrOwnrs ),missing
/* Recode legitimate (hard) missing values */
replace Tot_Debt_Owner_Operators    =.a if classf <6
replace Tot_Debt_Owed_Owner_Operators =.a if classf <6

egen Tot_Debt_Bus =rowtotal(f12b_bus_credcard_bal  f12c_bus_loans_bank_amt  f12b_bus_cred_line_bal
f12c_bus_loans_nonbank_amt       f12c_bus_loans_fam_amt  f12c_bus_loans_govt_amt  f12c_bus_loans_emp_amt
f12c_bus_loans_other_ind_amt  f12c_bus_loans_owner_amt  f12c_bus_loans_bus_amt  f12c_bus_other_amt ),missing
egen Tot_Bus_Debt_Owed =rowtotal(f12b_bus_cred_line_bal   f12b_bus_credcard_bal  f12d_bus_loans_bank_owed
f12d_bus_loans_nonbank_owed  f12d_bus_loans_emp_owed   f12d_bus_loans_fam_owed  f12d_bus_loans_govt_owed
f12d_bus_loans_other_ind_owed  f12d_bus_loans_owner_owed   f12d_bus_loans_bus_owed  f12d_bus_other_owed ),missing
/* Recode legitimate (hard) missing values */
replace Tot_Debt_Bus      =.a if classf <6
replace Tot_Bus_Debt_Owed =.a if classf <6
```

```
egen  Tot_Debt =rowtotal(Tot_Debt_Owner_Operators  Tot_Debt_Bus ),missing
egen  Tot_Debt_Owed =rowtotal(Tot_Debt_Owed_Owner_Operators  Tot_Bus_Debt_Owed ),missing
/* Recode legitimate (hard) missing values */
replace Tot_Debt       =.a if classf <6
replace Tot_Debt_Owed  =.a if classf <6

gen  Net_Profit =f24_profitloss_amt

* Long names create problems in MI data file. Make the names shorter

rename Tot_* *

/*********************************************************************/
/* Merge the file with the primary owner file “primary_owner.dta” */

merge m:1 mprid using "Xxx:\KFS_Manual_and_Data\primary_owner.dta"
drop _merge

* Primary Owner(PO) Characteristics at the Baseline: PO by Robb's Stata code

gen  PO_emp                    =.
gen  PO_hours                  =.
gen  PO_work_exp           =.
gen  PO_oth_bus_owner       =.
gen  PO_bus_same_ind        =.
gen  PO_age_owner          =.
gen  PO_hisp_origin           =.
gen  PO_race_group            =.
gen  PO_race_amind_owner    =.
gen  PO_race_asian_owner    =.
gen  PO_race_black_owner    =.
gen  PO_race_nathaw_owner   =.
gen  PO_race_other_owner    =.
gen  PO_race_white_owner    =.
gen  PO_native_born           =.
gen  PO_us_cit                 =.
gen  PO_education          =.
gen  PO_gender                 =.

mi xeq:sort master mprid year
mi xtset mprid year
```

```
/*
forvalues po = 1/6 {
mi xeq:bysort mprid (year):replace PO_emp          =   g1a_emp_owner_0`po'[1]              if  primary_owner==`po'
mi xeq:bysort mprid (year):replace PO_hours       =  g1b1_hours_owner_0`po'[1]                  if
primary_owner==`po'
mi xeq:bysort mprid (year):replace PO_work_exp       =     g2_work_exp_owner_0`po'[1]             if
primary_owner==`po'
mi xeq:bysort mprid (year):replace PO_oth_bus_owner =g3a_oth_bus_owner_0`po'[1]              if  primary_owner==`po'
mi xeq:bysort mprid (year):replace PO_bus_same_ind  =g3b_bus_same_ind_owner_0`po'[1]  if  primary_owner==`po'
mi xeq:bysort mprid (year):replace PO_age_owner      =     g4_age_owner_0`po'[1]                   if
primary_owner==`po'
mi xeq:bysort mprid (year):replace PO_hisp_origin    =     g5_hisp_origin_owner_0`po'[1]        if
primary_owner==`po'
mi xeq:bysort mprid (year):replace PO_race_group    =g6b_race_group_0`po'[1]               if  primary_owner==`po'
mi xeq:bysort mprid (year):replace PO_race_amind_owner  = g6_race_amind_owner_0`po'[1]      if  primary_owner==`po'
mi xeq:bysort mprid (year):replace PO_race_asian_owner  = g6_race_asian_owner_0`po'[1]      if  primary_owner==`po'
mi xeq:bysort mprid (year):replace PO_race_black_owner  = g6_race_black_owner_0`po'[1]      if  primary_owner==`po'
mi xeq:bysort mprid (year):replace PO_race_nathaw_owner = g6_race_nathaw_owner_0`po'[1]     if  primary_owner==`po'
mi xeq:bysort mprid (year):replace PO_race_other_owner  = g6_race_other_owner_0`po'[1]      if  primary_owner==`po'
mi xeq:bysort mprid (year):replace PO_race_white_owner  = g6_race_white_owner_0`po'[1]      if  primary_owner==`po'
mi xeq:bysort mprid (year):replace PO_native_born    =     g7_native_born_owner_0`po'[1]        if
primary_owner==`po'
mi xeq:bysort mprid (year):replace PO_us_cit       =  g8_us_cit_owner_0`po'[1]             if  primary_owner==`po'
mi xeq:bysort mprid (year):replace PO_education      =     g9_education_owner_0`po'[1]             if
primary_owner==`po'
mi xeq:bysort mprid (year):replace PO_gender       =  g10_gender_owner_0`po'[1]            if  primary_owner==`po'
}
*/
/* Faster way */
forvalues m = 0/5 {
forvalues po = 1/6 {
bysort mprid (year):replace PO_emp          =   g1a_emp_owner_0`po'[1]              if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_hours        =  g1b1_hours_owner_0`po'[1]                  if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_work_exp        =     g2_work_exp_owner_0`po'[1]             if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_oth_bus_owner =g3a_oth_bus_owner_0`po'[1]              if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_bus_same_ind  =g3b_bus_same_ind_owner_0`po'[1]  if  primary_owner==`po' & master==`m'
bysort mprid (year):replace PO_age_owner        =     g4_age_owner_0`po'[1]                   if  primary_owner==`po' &
master==`m'
```

```
bysort mprid (year):replace PO_hisp_origin   = g5_hisp_origin_owner_0`po'[1]        if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_race_group    = g6b_race_group_0`po'[1]              if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_race_amind_owner  = g6_race_amind_owner_0`po'[1]     if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_race_asian_owner  = g6_race_asian_owner_0`po'[1]     if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_race_black_owner  = g6_race_black_owner_0`po'[1]     if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_race_nathaw_owner = g6_race_nathaw_owner_0`po'[1]    if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_race_other_owner  = g6_race_other_owner_0`po'[1]     if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_race_white_owner  = g6_race_white_owner_0`po'[1]     if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_native_born     =     g7_native_born_owner_0`po'[1]    if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_us_cit       =  g8_us_cit_owner_0`po'[1]             if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_education       =     g9_education_owner_0`po'[1]            if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_gender       =  g10_gender_owner_0`po'[1]            if  primary_owner==`po' &
master==`m'
}
}

replace  PO_emp                = .a if classf   <6
replace  PO_hours            = .a if    classf   <6
replace  PO_work_exp           = .a if classf   <6
replace  PO_oth_bus_owner         = .a if     classf   <6
replace  PO_bus_same_ind     = .a if    classf   <6
replace  PO_age_owner          = .a if classf   <6
replace  PO_hisp_origin     = .a if     classf   <6
replace  PO_race_group      = .a if     classf   <6
replace  PO_race_amind_owner  = .a if   classf   <6
replace  PO_race_asian_owner  = .a if   classf   <6
replace  PO_race_black_owner  = .a if   classf   <6
replace  PO_race_nathaw_owner = .a if   classf   <6
replace  PO_race_other_owner  = .a if   classf   <6
replace  PO_race_white_owner  = .a if   classf   <6
replace  PO_native_born     = .a if     classf   <6
```

```
replace  PO_us_cit              = .a if classf   <6
replace  PO_education           = .a if classf   <6
replace  PO_gender              = .a if classf   <6
* Active-Owner-Operators Characteristics (OO)
egen   OO_emp_owner           =   rowmean(g1a_emp_owner_* )
egen   OO_hours_owner         = rowmean(g1b1_hours_owner_* )
egen   OO_work_exp_owner      =   rowmean(g2_work_exp_owner_*       )
egen   OO_oth_bus_owner       =   rowmean(g3a_oth_bus_owner_*       )
egen   OO_bus_same_ind_owner  =   rowmean(g3b_bus_same_ind_owner_* )
egen   OO_age_owner           =   rowmean(g4_age_owner_* )
egen   OO_hisp_origin_owner   =   rowmean(g5_hisp_origin_owner_* )
egen   OO_race_amind_owner    =   rowmean(g6_race_amind_owner_* )
egen   OO_race_asian_owner    =   rowmean(g6_race_asian_owner_* )
egen   OO_race_black_owner    =   rowmean(g6_race_black_owner_* )
egen   OO_race_nathaw_owner   =   rowmean(g6_race_nathaw_owner_* )
egen   OO_race_other_owner    =   rowmean(g6_race_other_owner_* )
egen   OO_race_white_owner    =   rowmean(g6_race_white_owner_* )
egen   OO_native_born_owner   =   rowmean(g7_native_born_owner_*    )
egen   OO_us_cit_owner        =   rowmean(g8_us_cit_owner_*  )
egen   OO_education_owner     =   rowmean(g9_education_owner_*      )
egen   md_education_owner     =   rowmedian(g9_education_owner_* )
gen    OO_D_education_owner   =(md_education_owner     >6.99)        if      md_education_owner     <11
egen   OO_gender_owner        =   rowmean(g10_gender_owner_*        )

/* Recode legitimate (hard) missing values */
replace     OO_bus_same_ind_owner= .a if   OO_oth_bus_owner==0
replace      OO_emp_owner           = .a if    classf   <6
replace      OO_hours_owner         = .a if    classf   <6
replace      OO_work_exp_owner      = .a if    classf   <6
replace      OO_oth_bus_owner       = .a if    classf   <6
replace      OO_bus_same_ind_owner  = .a if    classf   <6
replace      OO_age_owner           = .a if    classf   <6
replace      OO_hisp_origin_owner   = .a if    classf   <6
replace      OO_race_amind_owner    = .a if    classf   <6
replace      OO_race_asian_owner    = .a if    classf   <6
replace      OO_race_black_owner    = .a if    classf   <6
replace      OO_race_nathaw_owner   = .a if    classf   <6
replace      OO_race_other_owner    = .a if    classf   <6
replace      OO_race_white_owner    = .a if    classf   <6
replace      OO_native_born_owner   = .a if    classf   <6
replace      OO_us_cit_owner        = .a if    classf   <6
replace      OO_education_owner     = .a if    classf   <6
```

```
replace      md_education_owner     = .a if    classf   <6
replace      OO_D_education_owner   = .a if    classf   <6
replace      OO_gender_owner        = .a if    classf   <6
/*****************************************************************************/
*Diversity / Similarity index
gen xr1 =    OO_race_amind_owner        *      OO_race_amind_owner
gen xr2 =    OO_race_asian_owner        *      OO_race_asian_owner
gen xr3 =    OO_race_black_owner        *      OO_race_black_owner
gen xr4 =    OO_race_nathaw_owner *     OO_race_nathaw_owner
gen xr5 =    OO_race_other_owner  *     OO_race_other_owner
gen xr6 =    OO_race_white_owner        *      OO_race_white_owner
* Race_similarity
egen Race_similarity =rowtotal(xr1  xr2  xr3  xr4  xr5  xr6 ),missing
drop xr*
* Race_ diversity
gen Race_diversity =1-Race_similarity
/* Recode legitimate (hard) missing values */
replace Race_similarity    =.a if classf <6
replace Race_diversity    =.a if Race_similarity ==.a
gen fmal =1-OO_gender_owner
gen xr1 =    OO_gender_owner     *      OO_gender_owner
gen xr2 =    fmal   *      fmal
* Gender_similarity
egen Gender_similarity =rowtotal(xr1  xr2   ),missing
drop xr*  fmal
* Gender_diversity
gen Gender_diversity =1-Gender_similarity
/* Recode legitimate (hard) missing values */
replace Gender_similarity    =.a if classf <6
replace Gender_diversity    =.a if Gender_similarity ==.a
*Business level Characteristics
*Home Based Dummy
recode c8_primary_loc  (1=1 "Home Based") (nonmiss=0 "Non Home Based" ) ,into (Home_Based )
*Sole_Proprietorship Dummy
recode c1z2_legal_status  (1=1 "Sole_Proprietorship") (nonmiss=0 "Limited Liability" ) ,into (Sole_Proprietorship )
rename d2_comp_advantage  Comp_advantage
egen Have_IP =anymatch(d3_a_have_patent  d3_b_have_copyright  d3_c_have_trademark ), values(1)
rename c5_num_employees      Full_Part_Time_Employees
rename c6_num_ft_employees   Full_Time_Employees
rename c7_num_pt_employees   Part_Time_Employees
egen   Employee_Owner           =rowtotal(g1a_emp_owner_* )
egen   Total_Employees =rowtotal(Employee_Owner  Full_Part_Time_Employees )
```

```
replace Have_IP         =.a if classf <6
replace Employee_Owner  =.a if classf <6
replace Total_Employees =.a if classf <6

saveold Cross_Sectional_Long_MI_Long_L1,replace
```

3.2.3.2.4. Stata Code: Longitudinal in Long Format

The following Stata code will create the same variables we discussed in sections 3.2.2.2.1, 3.2.2.2.2, and 3.2.2.2.3, using the KFS multiply imputed data file "KFS8_Longitudinal_Long_MI_Long." The code will save the new file under the name "Longitudinal_Long_MI_Long_L1.dta" (n=18,286*6=109,716 obs.).

For the Longitudinal file, the last observation carried forward is used to fill in all the fixed core set of questions for businesses reported as temporarily stopped or located (classf=0 or 5); thus, those businesses will have data for all the fixed core set of questions even when classf=0 or 5.

```
clear
clear mata
clear matrix
capture log close
set more off
set maxvar 32767
program drop _all
adopath + " xxx:\KFS_Manual_and_Data\Farhat_Robb_Commands\"
cd "Xxx:\KFS_Manual_and_Data"
  use KFS8_Longitudinal_Long_MI_Long,clear


/*For the Longitudinal file, last observation carried forward is used to fill in all the fixed core set of questions
for businesses reported being
temporarily stopped or located (classf=0 or 5), thus those businesses will have data for all the fixed core set of
questions even when classf=0 or 5.*/

recode classf  (0=6 ) (5=6) ,into (classf_L)

egen   Tot_Equity_Owner_Operators =     rowtotal(     f2_owner_amt_eq_invest_*   )      , missing
egen   Tot_Equity_OwnerOper_AllYrs =rowtotal(  f2_ownr_amt_eqinvest_allyrs_*     ), missing

/* Recode legitimate (hard) missing values */
replace Tot_Equity_Owner_Operators  =.a if classf_L <6
replace Tot_Equity_OwnerOper_AllYrs =.a if classf_L <6

global List1 "spouse parents angels companies govt vent_cap other"
```

```
egen  Tot_Equity_NonOwnerOperators =rowtotal(f4_eq_amt_angels  f4_eq_amt_companies  f4_eq_amt_govt  ///
f4_eq_amt_other  f4_eq_amt_parents  f4_eq_amt_spouse   f4_eq_amt_vent_cap )     , missing
egen  Tot_Equity_NonOwnerOp_AllYrs =rowtotal(f4_eq_amt_*_allyrs )  , missing
/* Recode legitimate (hard) missing values */
replace Tot_Equity_NonOwnerOperators  =.a if classf_L <6
replace Tot_Equity_NonOwnerOperators  =.a if  c1z2_legal_status ==1
replace Tot_Equity_NonOwnerOp_AllYrs  =.a if classf_L <6
replace Tot_Equity_NonOwnerOp_AllYrs  =.a if  c1z2_legal_status ==1

egen  Tot_Equity        =rowtotal(Tot_Equity_Owner_Operators  Tot_Equity_NonOwnerOperators ), missing
egen  Tot_Equity_AllYrs =rowtotal(Tot_Equity_OwnerOper_AllYrs  Tot_Equity_NonOwnerOp_AllYrs ), missing
/* Recode legitimate (hard) missing values */
replace Tot_Equity        =.a if classf_L <6
replace Tot_Equity_AllYrs =.a if classf_L <6

egen  Tot_Assets =rowtotal(f29_assetval_* )    , missing
egen  Tot_Liab =rowtotal(f31_value_* ), missing
/* Recode legitimate (hard) missing values */
replace Tot_Assets  =.a if classf_L <6
replace Tot_Liab    =.a if classf_L <6

egen  Tot_Pers_Debt_Resp =rowtotal(f8b_pers_credcard_bal  f8b_bus_credcard_bal  f8c_pers_loan_bank_amt
f8c_pers_loan_fam_amt  f8c_pers_loan_other_amt  f8c_pers_other_amt ),missing
egen  Tot_Pers_Debt_Owed_Resp =rowtotal(f8b_pers_credcard_bal  f8b_bus_credcard_bal  f8d_pers_loan_bank_owed
f8d_pers_loan_fam_owed  f8d_pers_loan_other_owed  f8d_pers_other_owed ),missing
/* Recode legitimate (hard) missing values */
replace Tot_Pers_Debt_Resp       =.a if classf_L <6
replace Tot_Pers_Debt_Owed_Resp  =.a if classf_L <6

egen  Tot_Pers_Debt_Other_Owners =rowtotal(f10b_pers_credcard_bal  f10c_pers_loan_bank_amt  f10b_bus_credcard_bal
f10c_pers_loan_fam_amt  f10c_pers_loan_other_amt   f10c_pers_other_amt ),missing
egen  Tot_Pers_Debt_Owed_OthrOwnrs =rowtotal(f10b_pers_credcard_bal  f10b_bus_credcard_bal  f10d_pers_loan_bank_owed
f10d_pers_loan_fam_owed  f10d_pers_loan_other_owed  f10d_pers_other_owed ),missing
/* Recode legitimate (hard) missing values */
replace Tot_Pers_Debt_Other_Owners   =.a if classf_L <6
replace Tot_Pers_Debt_Owed_OthrOwnrs =.a if classf_L <6
replace Tot_Pers_Debt_Other_Owners   =.a if c4_numowners_confirm <2
replace Tot_Pers_Debt_Owed_OthrOwnrs =.a if c4_numowners_confirm <2

egen  Tot_Debt_Owner_Operators =rowtotal(Tot_Pers_Debt_Resp  Tot_Pers_Debt_Other_Owners ),missing
egen  Tot_Debt_Owed_Owner_Operators =rowtotal(Tot_Pers_Debt_Owed_Resp  Tot_Pers_Debt_Owed_OthrOwnrs ),missing
/* Recode legitimate (hard) missing values */
```

```
replace Tot_Debt_Owner_Operators    =.a if classf_L <6
replace Tot_Debt_Owed_Owner_Operators =.a if classf_L <6

egen Tot_Debt_Bus =rowtotal(f12b_bus_credcard_bal  f12c_bus_loans_bank_amt  f12b_bus_cred_line_bal
f12c_bus_loans_nonbank_amt       f12c_bus_loans_fam_amt  f12c_bus_loans_govt_amt  f12c_bus_loans_emp_amt
f12c_bus_loans_other_ind_amt  f12c_bus_loans_owner_amt  f12c_bus_loans_bus_amt  f12c_bus_other_amt ),missing
egen Tot_Bus_Debt_Owed =rowtotal(f12b_bus_cred_line_bal   f12b_bus_credcard_bal  f12d_bus_loans_bank_owed
f12d_bus_loans_nonbank_owed  f12d_bus_loans_emp_owed   f12d_bus_loans_fam_owed  f12d_bus_loans_govt_owed
f12d_bus_loans_other_ind_owed  f12d_bus_loans_owner_owed   f12d_bus_loans_bus_owed  f12d_bus_other_owed ),missing
/* Recode legitimate (hard) missing values */
replace Tot_Debt_Bus      =.a if classf_L <6
replace Tot_Bus_Debt_Owed =.a if classf_L <6

egen  Tot_Debt =rowtotal(Tot_Debt_Owner_Operators  Tot_Debt_Bus ),missing
egen  Tot_Debt_Owed =rowtotal(Tot_Debt_Owed_Owner_Operators  Tot_Bus_Debt_Owed ),missing
/* Recode legitimate (hard) missing values */
replace Tot_Debt       =.a if classf_L <6
replace Tot_Debt_Owed  =.a if classf_L <6

gen  Net_Profit =f24_profitloss_amt

* Long names create problems in MI data file. Make the names shorter

rename Tot_* *

/*****************************************************************/
/* Merge the file with the primary owner file “primary_owner.dta” */

merge m:1 mprid using "Xxx:\KFS_Manual_and_Data\primary_owner.dta"
drop if _merge!=3
drop _merge

* Primary Owner(PO) Characteristics at the Baseline: PO by Robb's Stata code

gen  PO_emp                    =.
gen  PO_hours                  =.
gen  PO_work_exp           =.
gen  PO_oth_bus_owner       =.
gen  PO_bus_same_ind        =.
gen  PO_age_owner          =.
gen  PO_hisp_origin           =.
gen  PO_race_group            =.
```

```
gen  PO_race_amind_owner    =.
gen  PO_race_asian_owner    =.
gen  PO_race_black_owner    =.
gen  PO_race_nathaw_owner   =.
gen  PO_race_other_owner    =.
gen  PO_race_white_owner    =.
gen  PO_native_born           =.
gen  PO_us_cit                  =.
gen  PO_education           =.
gen  PO_gender                  =.


mi xtset mprid year

forvalues m = 0/5 {
forvalues po = 1/6 {
bysort mprid (year):replace PO_emp          =   g1a_emp_owner_0`po'[1]              if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_hours        =  g1b1_hours_owner_0`po'[1]                   if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_work_exp       =     g2_work_exp_owner_0`po'[1]              if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_oth_bus_owner = g3a_oth_bus_owner_0`po'[1]             if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_bus_same_ind  = g3b_bus_same_ind_owner_0`po'[1]  if  primary_owner==`po' & master==`m'
bysort mprid (year):replace PO_age_owner       =     g4_age_owner_0`po'[1]                    if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_hisp_origin   = g5_hisp_origin_owner_0`po'[1]        if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_race_group    = g6b_race_group_0`po'[1]              if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_race_amind_owner  = g6_race_amind_owner_0`po'[1]     if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_race_asian_owner  = g6_race_asian_owner_0`po'[1]     if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_race_black_owner  = g6_race_black_owner_0`po'[1]     if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_race_nathaw_owner = g6_race_nathaw_owner_0`po'[1]    if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_race_other_owner  = g6_race_other_owner_0`po'[1]     if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_race_white_owner  = g6_race_white_owner_0`po'[1]     if  primary_owner==`po' &
```

```
master==`m'
bysort mprid (year):replace PO_native_born     =      g7_native_born_owner_0`po'[1]    if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_us_cit       =  g8_us_cit_owner_0`po'[1]             if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_education       =      g9_education_owner_0`po'[1]            if  primary_owner==`po' &
master==`m'
bysort mprid (year):replace PO_gender       =  g10_gender_owner_0`po'[1]            if  primary_owner==`po' &
master==`m'
}
}
replace  PO_emp                 = .a if  classf_L   <6
replace  PO_hours            = .a if    classf_L   <6
replace  PO_work_exp            = .a if  classf_L   <6
replace  PO_oth_bus_owner          = .a if     classf_L   <6
replace  PO_bus_same_ind     = .a if    classf_L   <6
replace  PO_age_owner           = .a if  classf_L   <6
replace  PO_hisp_origin     = .a if     classf_L   <6
replace  PO_race_group      = .a if     classf_L   <6
replace  PO_race_amind_owner  = .a if   classf_L   <6
replace  PO_race_asian_owner  = .a if   classf_L   <6
replace  PO_race_black_owner  = .a if   classf_L   <6
replace  PO_race_nathaw_owner = .a if   classf_L   <6
replace  PO_race_other_owner  = .a if   classf_L   <6
replace  PO_race_white_owner  = .a if   classf_L   <6
replace  PO_native_born     = .a if     classf_L   <6
replace  PO_us_cit              = .a if  classf_L   <6
replace  PO_education           = .a if  classf_L   <6
replace  PO_gender              = .a if  classf_L   <6

* Active-Owner-Operators Characteristics (OO)

egen   OO_emp_owner           =   rowmean(g1a_emp_owner_* )
egen   OO_hours_owner         = rowmean(g1b1_hours_owner_* )
egen   OO_work_exp_owner      =   rowmean(g2_work_exp_owner_*       )
egen   OO_oth_bus_owner       =   rowmean(g3a_oth_bus_owner_*       )
egen   OO_bus_same_ind_owner  =   rowmean(g3b_bus_same_ind_owner_* )
egen   OO_age_owner           =   rowmean(g4_age_owner_* )
egen   OO_hisp_origin_owner   =   rowmean(g5_hisp_origin_owner_* )
egen   OO_race_amind_owner    =   rowmean(g6_race_amind_owner_* )
egen   OO_race_asian_owner    =   rowmean(g6_race_asian_owner_* )
egen   OO_race_black_owner    =   rowmean(g6_race_black_owner_* )
```

```
egen   OO_race_nathaw_owner   =   rowmean(g6_race_nathaw_owner_* )
egen   OO_race_other_owner    =   rowmean(g6_race_other_owner_* )
egen   OO_race_white_owner    =   rowmean(g6_race_white_owner_* )
egen   OO_native_born_owner   =   rowmean(g7_native_born_owner_*    )
egen   OO_us_cit_owner        =   rowmean(g8_us_cit_owner_*  )
egen   OO_education_owner     =   rowmean(g9_education_owner_*     )
egen   md_education_owner     =   rowmedian(g9_education_owner_* )
gen    OO_D_education_owner   =(md_education_owner     >6.99)        if      md_education_owner     <11

egen   OO_gender_owner        =   rowmean(g10_gender_owner_*       )


/* Recode legitimate (hard) missing values */
replace     OO_bus_same_ind_owner= .a if   OO_oth_bus_owner==0
replace      OO_emp_owner           = .a if    classf_L   <6
replace      OO_hours_owner         = .a if    classf_L   <6
replace      OO_work_exp_owner      = .a if    classf_L   <6
replace      OO_oth_bus_owner       = .a if    classf_L   <6
replace      OO_bus_same_ind_owner  = .a if    classf_L   <6
replace      OO_age_owner           = .a if    classf_L   <6
replace      OO_hisp_origin_owner   = .a if    classf_L   <6
replace      OO_race_amind_owner    = .a if    classf_L   <6
replace      OO_race_asian_owner    = .a if    classf_L   <6
replace      OO_race_black_owner    = .a if    classf_L   <6
replace      OO_race_nathaw_owner   = .a if    classf_L   <6
replace      OO_race_other_owner    = .a if    classf_L   <6
replace      OO_race_white_owner    = .a if    classf_L   <6
replace      OO_native_born_owner   = .a if    classf_L   <6
replace      OO_us_cit_owner        = .a if    classf_L   <6
replace      OO_education_owner     = .a if    classf_L   <6
replace      md_education_owner     = .a if    classf_L   <6
replace      OO_D_education_owner   = .a if    classf_L   <6

replace      OO_gender_owner        = .a if    classf_L   <6

/*****************************************************************************/
*Diversity / Similarity index
gen xr1 =    OO_race_amind_owner       *      OO_race_amind_owner
gen xr2 =    OO_race_asian_owner       *      OO_race_asian_owner
gen xr3 =    OO_race_black_owner       *      OO_race_black_owner
gen xr4 =    OO_race_nathaw_owner *     OO_race_nathaw_owner
gen xr5 =    OO_race_other_owner  *     OO_race_other_owner
```

```
gen xr6 =    OO_race_white_owner        *      OO_race_white_owner
* Race_similarity
egen Race_similarity =rowtotal(xr1  xr2  xr3  xr4  xr5  xr6 ),missing
drop xr*
* Race_ diversity
gen Race_diversity =1-Race_similarity

/* Recode legitimate (hard) missing values */
replace Race_similarity    =.a if classf_L <6
replace Race_diversity    =.a if Race_similarity ==.a

gen fmal =1-OO_gender_owner
gen xr1 =    OO_gender_owner     *      OO_gender_owner
gen xr2 =    fmal   *      fmal

* Gender_similarity
egen Gender_similarity =rowtotal(xr1  xr2   ),missing
drop xr*  fmal
* Gender_diversity
gen Gender_diversity =1-Gender_similarity

/* Recode legitimate (hard) missing values */
replace Gender_similarity    =.a if classf_L <6
replace Gender_diversity    =.a if Gender_similarity ==.a

*Business level Characteristics
*Home Based Dummy
recode c8_primary_loc  (1=1 "Home Based") (nonmiss=0 "Non Home Based" ) ,into (Home_Based )
*Sole_Proprietorship Dummy
recode c1z2_legal_status  (1=1 "Sole_Proprietorship") (nonmiss=0 "Limited Liability" ) ,into (Sole_Proprietorship )
rename d2_comp_advantage  Comp_advantage
egen Have_IP =anymatch(d3_a_have_patent  d3_b_have_copyright  d3_c_have_trademark ), values(1)
rename c5_num_employees      Full_Part_Time_Employees
rename c6_num_ft_employees   Full_Time_Employees
rename c7_num_pt_employees   Part_Time_Employees
egen   Employee_Owner          =rowtotal(g1a_emp_owner_* )
egen   Total_Employees =rowtotal(Employee_Owner  Full_Part_Time_Employees )

replace Have_IP         =.a if classf_L <6
replace Employee_Owner  =.a if classf_L <6
replace Total_Employees =.a if classf_L <6
```

```
saveold Longitudinal_Long_MI_Long_L1,replace
```

3.3. Comparing the KFS Imputed to Non-Imputed Data

Researchers usually like to compare their research results using imputed data to results using non-imputed data. While this problem can be addressed by using "mi xeq 0:" command, researchers need to exercise caution about which data file to use.

In the data files we created in section 3.2.2.2 (KFS8_CS_w1, KFS8_L7_w1, KFS8_CS_L1, KFS8_L7_L1), we replaced missing continuous values by the midpoints of the class intervals, if the value is reported as a range. Meanwhile, in the KFS multiply imputed data files[4], missing continuous values were not replaced by the midpoints of the class intervals (if the value is reported as a range). This is because those missing continuous values were subject to imputation. Thus, comparing the results using original (m=0) data is not correct.

To perform correct comparison analysis of the imputed versus non-imputed data, we need to have the original (m=0) data where missing continuous values are replaced by the midpoints of the class intervals.

For researchers that would like to compare their research results using the imputed data to the results using non-imputed data, the following Stata code will generate the correct files to use for this purpose. The files names are:

1. Cross_Sectional_wide_MI_Long_w2
2. Longitudinal_wide_MI_Long_w2
3. Cross_Sectional_Long_MI_Long_L2
4. Longitudinal_Long_MI_Long_L2

Make sure not to register any of the variables in those files. We replace the missing continuous values by the midpoints of the class intervals, and all our amount variables are super varying.

```
clear
cd "Xxx:\KFS_Manual_and_Data"
use Cross_Sectional_wide_MI_Long_w1, clear
mi unset
drop mi_*
drop if master==0
tempfile temp1
save "`temp1'"
use KFS8_CS_w1, clear
gen master=0
tempfile temp2
save "`temp2'"
append using "`temp1'"
mi import flong , m(master) id(mprid)  clear
saveold Cross_Sectional_wide_MI_Long_w2,replace

/*******************************************************/
```

[4]MI files : KFS8_Cross_Sectional_wide_MI_Long, KFS8_Cross_Sectional_Long_MI_Long, KFS8_Longitudinal_wide_MI_Long, and KFS8_Longitudinal_Long_MI_Long.

```
clear
cd "Xxx:\KFS_Manual_and_Data"

use Longitudinal_wide_MI_Long_w1, clear
mi unset
drop mi_*
drop if master==0
tempfile temp3
save "`temp3'"
use KFS8_L7_w1, clear
gen master=0
tempfile temp4
save "`temp4'"
append using "`temp3'"
mi import flong , m(master) id(mprid)  clear
saveold Longitudinal_wide_MI_Long_w2,replace

/*******************************************************/

clear
cd "Xxx:\KFS_Manual_and_Data"

use Cross_Sectional_Long_MI_Long_L1, clear
mi unset
drop mi_*
drop if master==0
tempfile temp5
save "`temp5'"
use KFS8_CS_L1, clear
gen master=0
tempfile temp6
save "`temp6'"
append using "`temp5'"
mi import flong , m(master) id(year mprid)  clear
saveold Cross_Sectional_Long_MI_Long_L2,replace

/*******************************************************/

clear
cd "Xxx:\KFS_Manual_and_Data"

use Longitudinal_Long_MI_Long_L1, clear
mi unset
drop mi_*
drop if master==0
tempfile temp7
save "`temp7'"
use KFS8_L7_L1, clear
gen master=0
tempfile temp8
save "`temp8'"
append using "`temp7'"
mi import flong , m(master) id(year mprid)  clear
saveold Longitudinal_Long_MI_Long_L2,replace
```

4.1. Exploratory Data Analysis (EDA)

This chapter will show you how to use Stata version 12.0 to perform exploratory data analysis. Descriptive statistics and exploratory data analysis (EDA) is an approach to analyzing data sets by summarizing and visualizing their main characteristics.

For some KFS users, the selection of the correct weights to use in analysis may seem confusing. The general guideline to determine the correct weights to use is that studies that aim to provide estimates at cross-sectional levels, compare aggregates over time, identify net (macro-level) changes from one follow-up to another, create a snapshot of a population at one point in time, or investigate association should utilize the cross-sectional weights. On the other hand, studies that aim at measuring gross (micro-level) changes over time or investigating causation utilize the longitudinal weights.

If we are only using information collected during the baseline survey, then we would use the baseline cross-section weights (cswgt_final_0). Similarly, if we are only using the fifth survey, then we would use the fifth cross-section weights (cswgt_final_5). If we want to infer how firm characteristics have changed across the four years between baseline and the fourth survey, then we would use the fourth survey longitudinal weights (wgt_4_long). Similarly, if we want to infer how firm characteristics have changed across the eight years between baseline and the seventh survey, then we would use the seventh survey longitudinal weights (wgt_7_long).

4.2. Reading and Declaring Complex Survey Data

The first step after reading the KFS data file is to declare the survey design futures. Next is to define the panel data if we are using the KFS data in the long format. Later, we will show how to declare data to be survival-time data for the purpose of survival analysis.

Also, it is worth noting that we can always extract the original KFS data from the imputed data.

Extract the original KFS data from imputed file	Equivalent to use the following data:
use Cross_Sectional_wide_MI_Long_w2,clear mi extract 0	use KFS8_CS_w1,clear
use Longitudinal_wide_MI_Long_w2,clear mi extract 0	use KFS8_L7_w1,clear
use Cross_Sectional_Long_MI_Long_L2,clear mi extract 0	use KFS8_CS_L1,clear
use Longitudinal_Long_MI_Long_L2,clear mi extract 0	use KFS8_L7_L1,clear

An alternative to the mi extract 0 command will be to use mi xeq 0: Stata command to run the Stata command on the original KFS data. We do not recommend this approach because it is not memory efficient (i.e., very slow)

Running estimation on the original KFS data using imputed file	Equivalent to use the following data:
use Cross_Sectional_wide_MI_Long_w2,clear mi xeq 0: Stata command	use KFS8_CS_w1,clear
use Longitudinal_wide_MI_Long_w2,clear mi xeq 0: Stata command	use KFS8_L7_w1,clear
use Cross_Sectional_Long_MI_Long_L2,clear mi xeq 0: Stata command	use KFS8_CS_L1,clear
use Longitudinal_Long_MI_Long_L2,clear mi xeq 0: Stata command	use KFS8_L7_L1,clear

Example 4.1: KFS in Wide Format

```
*Example 4.1: KFS in wide format
use KFS8_CS_w1,clear
*Declare survey design for dataset (In this example we use the CS seventh survey
data)
svyset  [pweight=cswgt_final_7] , strata(sampleinfo_samplestrata_7)
vce(linearized)  clear
*Describe survey data
svydescribe

use KFS8_L7_w1,clear
*Declare survey design for dataset (In this example we use the eight years panel
data)
svyset  [pweight= wgt_7_long] , strata(sampleinfo_samplestrata_0)
vce(linearized)  clear
*Describe survey data
svydescribe


*Output omitted
```

Example 4.2: KFS MI in Wide Format

```
*Example 4.2: KFS MI in wide format
use Cross_Sectional_wide_MI_Long_w2,clear
*Declare survey design for dataset (In this example we use the CS seventh survey
data)
mi svyset  [pweight=cswgt_final_7] , strata(sampleinfo_samplestrata_7)
vce(linearized)  clear
*Describe survey data
mi xeq: svydescribe

use Longitudinal_wide_MI_Long_w2,clear
*Declare survey design for dataset (In this example we use the eight years panel
data)
mi svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_0)
vce(linearized)  clear
*Describe survey data
mi xeq: svydescribe

*Output omitted
```

Example 4.3: KFS in Long Format

```
*Example 4.3: KFS in long format
use KFS8_CS_L1,clear
drop if cswgt_final==.
*Declare survey design for dataset
svyset  [pweight=cswgt_final] , strata(sampleinfo_samplestrata )
vce(linearized)  clear
*Describe survey data
svydescribe
* Declare data to be panel data
xtset mprid year
*Describe pattern of xt data
xtdescribe

   mprid:  10000016, 10000090, ..., 10324611                n =       4928
    year:  2004, 2005, ..., 2011                            T =          8
           Delta(year) = 1 unit
           Span(year)  = 8 periods
           (mprid*year uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         1       3       7         8         8       8       8

     Freq.  Percent    Cum. |  Pattern
 -------------------------+----------
     3140     63.72   63.72 |  11111111
      138      2.80   66.52 |  1.111111
      124      2.52   69.03 |  1.......
      122      2.48   71.51 |  111.1111
      119      2.41   73.92 |  11.11111
       76      1.54   75.47 |  1111.111
       75      1.52   76.99 |  1..11111
       74      1.50   78.49 |  11......
       71      1.44   79.93 |  111.....
      989     20.07  100.00 | (other patterns)
 -------------------------+----------
     4928    100.00         |  XXXXXXXX
```

```
use KFS8_L7_L1,clear
*Declare survey design for dataset
svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata )
vce(linearized)  clear
*Describe survey data
svydescribe
* Declare data to be panel data
xtset mprid year
*Describe pattern of xt data
xtdescribe
```

```
   mprid:  10000016, 10000090, ..., 10324611                n =       3140
    year:  2004, 2005, ..., 2011                            T =          8
           Delta(year) = 1 unit
           Span(year)  = 8 periods
           (mprid*year uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         1       1       3         8         8       8       8

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+----------
     1630     51.91   51.91 |  11111111
      303      9.65   61.56 |  1.......
      283      9.01   70.57 |  11......
      238      7.58   78.15 |  1111....
      224      7.13   85.29 |  111.....
      164      5.22   90.51 |  11111...
      153      4.87   95.38 |  111111..
      145      4.62  100.00 |  1111111.
 ---------------------------+----------
     3140    100.00         |  XXXXXXXX
```

Example 4.4: KFS MI in Long Format

```
*Example 4.4: KFS MI in long format
use Cross_Sectional_Long_MI_Long_L2,clear
drop if cswgt_final==.
*Declare survey design for dataset
mi svyset  [pweight=cswgt_final] , strata(sampleinfo_samplestrata)
vce(linearized)  clear
*Describe survey data
mi xeq: svydescribe
* Declare data to be panel data
mi xtset mprid year
*Describe pattern of xt data
mi xeq: xtdescribe

use  Longitudinal_Long_MI_Long_L2,clear
*Declare survey design for dataset
mi svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
vce(linearized)  clear
*Describe survey data
mi xeq: svydescribe
*Declare data to be panel data
mi xtset mprid year
*Describe pattern of xt data
mi xeq: xtdescribe

*Output omitted
```

4.3. Tabulate Missing Values

The Stata command "misstable summarize" reports count soft and hard missing values.

Example 4.5: Using KFS in Wide Format

```
*Example 4.5: Using KFS in Wide Format
use KFS8_CS_w1,clear
*Or we can use the Cross_Sectional_wide_MI_Long_w2 file where we can extract the
original KFS data (m=0)
*Extract the original KFS data (m=0)
use Cross_Sectional_wide_MI_Long_w2,clear
mi extract 0
#delimit ;
misstable sum  credrisk_0 c1z2_legal_status_0 c8_primary_loc_0  d3_a_have_patent_0
Equity_Owner_Operators_0
Equity_NonOwnerOperators_0 OO_work_exp_owner_0 PO_work_exp_0 Equity_0   Assets_0
Pers_Debt_Resp_0 Pers_Debt_Other_Owners_0 Debt_Bus_0
e2a_ft_emp_hlth_plan_0 e2b_pt_emp_hlth_plan_0
;
```

```
                                                    Obs<.
                                          +------------------------------
             |                            | Unique
    Variable |     Obs=.     Obs>.     Obs<. | values        Min         Max
-------------+--------------------------------+------------------------------
  credrisk_0 |     1,322                3,606 |      5          1           5
d3_a_have_~0 |        31                4,897 |      2          0           1
Equity_Own.. |       172                4,756 |    281          0    1.01e+08
Equity_Non.. |        71     2,576      2,281 |    125          0    6.00e+07
OO_work_ex~0 |        11                4,917 |    172          0          60
PO_work_ex~0 |         6                4,922 |     55          0          60
    Equity_0 |       208                4,720 |    365          0    1.01e+08
    Assets_0 |       429                4,499 |   >500          0    3.01e+08
Per~t_Resp_0 |       215                4,713 |    423          0     4000000
Pers_D~ers_0 |        83     3,445      1,400 |     99          0      550000
  Debt_Bus_0 |       242                4,686 |    309          0    7.50e+07
e2a~h_plan_0 |        41     1,840      3,047 |      2          0           1
e2b~h_plan_0 |       152     3,842        934 |      2          0           1
-------------------------------------------------------------------------------
```

```
use KFS8_L7_w1,clear
*Or we can use the Longitudinal_wide_MI_Long_w2 where we can extract the original
KFS data (m=0)
*Extract the original KFS data (m=0)
use Longitudinal_wide_MI_Long_w2,clear
mi extract 0
#delimit ;
misstable sum  credrisk_0 c1z2_legal_status_0 c8_primary_loc_0  d3_a_have_patent_0
Equity_Owner_Operators_0
Equity_NonOwnerOperators_0 OO_work_exp_owner_0 PO_work_exp_0 Equity_0   Assets_0
Pers_Debt_Resp_0 Pers_Debt_Other_Owners_0 Debt_Bus_0
e2a_ft_emp_hlth_plan_0 e2b_pt_emp_hlth_plan_0
;
```

```
                                                                  Obs<.
                                                  +------------------------------
             |                                    | Unique
    Variable |     Obs=.     Obs>.     Obs<.     | values        Min          Max
-------------+--------------------------------+------------------------------
  credrisk_0 |       812               2,328    |      5          1           5
d3_a_have_~0 |        20               3,120    |      2          0           1
Equity_Own.. |        84               3,056    |    220          0    1.01e+08
Equity_Non.. |        41     1,766     1,333    |     91          0    1.25e+07
OO_work_ex~0 |         3               3,137    |    142          0          60
PO_work_ex~0 |         2               3,138    |     52          0          60
    Equity_0 |       110               3,030    |    272          0    1.01e+08
    Assets_0 |       219               2,921    |   >500          0    8.00e+07
Per~t_Resp_0 |       127               3,013    |    325          0     3600000
Pers_D~ers_0 |        44     2,240       856    |     73          0      550000
  Debt_Bus_0 |       136               3,004    |    211          0    1.20e+07
e2a~h_plan_0 |        23     1,235     1,882    |      2          0           1
e2b~h_plan_0 |       108     2,452       580    |      2          0           1
-------------------------------------------------------------------------------
```

Example 4.6: Using KFS in Long Format

We can obtain the same output as in example 4.5 using the KFS data in the long format.

```
*Example 4.6: Using KFS in Long Format
use KFS8_CS_L1,clear
*Or we can use the Cross_Sectional_Long_MI_Long_L2 file where we can extract the
original KFS data (m=0)
*Extract the original KFS data (m=0)
use Cross_Sectional_Long_MI_Long_L2,clear
mi extract 0
* Declare data to be panel data
xtset mprid year
*In this case we need to restrict the output to data in year 2004.
#delimit ;
misstable  sum    credrisk  c1z2_legal_status  c8_primary_loc    d3_a_have_patent
Equity_Owner_Operators
Equity_NonOwnerOperators   OO_work_exp_owner   PO_work_exp   Equity      Assets
Pers_Debt_Resp Pers_Debt_Other_Owners Debt_Bus
e2a_ft_emp_hlth_plan e2b_pt_emp_hlth_plan if year==2004
;
                                                                  Obs<.
                                                  +------------------------------
             |                                    | Unique
    Variable |     Obs=.     Obs>.     Obs<.     | values        Min          Max
-------------+--------------------------------+------------------------------
    credrisk |     1,322               3,606    |      5          1           5
d3_a_have_~t |        31               4,897    |      2          0           1
Equity_O~ors |       172               4,756    |    281          0    1.01e+08
Equity_N~ors |        71     2,576     2,281    |    125          0    6.00e+07
OO_work_ex~r |        11               4,917    |    172          0          60
 PO_work_exp |         6               4,922    |     55          0          60
      Equity |       208               4,720    |    365          0    1.01e+08
      Assets |       429               4,499    |   >500          0    3.01e+08
Pers_~t_Resp |       215               4,713    |    423          0     4000000
Pers_Deb~ers |        83     3,445     1,400    |     99          0      550000
    Debt_Bus |       242               4,686    |    309          0    7.50e+07
e2a_f~h_plan |        41     1,840     3,047    |      2          0           1
e2b_p~h_plan |       152     3,842       934    |      2          0           1
-------------------------------------------------------------------------------
```

```
use KFS8_L7_L1,clear
*Or we can use the Longitudinal_Long_MI_Long_L2 file where we can extract the
original KFS data (m=0)
*Extract the original KFS data (m=0)
use Longitudinal_Long_MI_Long_L2,clear
mi extract 0
*Declare data to be panel data
xtset mprid year
*In this case we need to restrict the output to data in year 2004.
#delimit ;
misstable sum   credrisk  c1z2_legal_status  c8_primary_loc   d3_a_have_patent
Equity_Owner_Operators
Equity_NonOwnerOperators   OO_work_exp_owner   PO_work_exp   Equity      Assets
Pers_Debt_Resp Pers_Debt_Other_Owners Debt_Bus
e2a_ft_emp_hlth_plan e2b_pt_emp_hlth_plan if year==2004
;
```

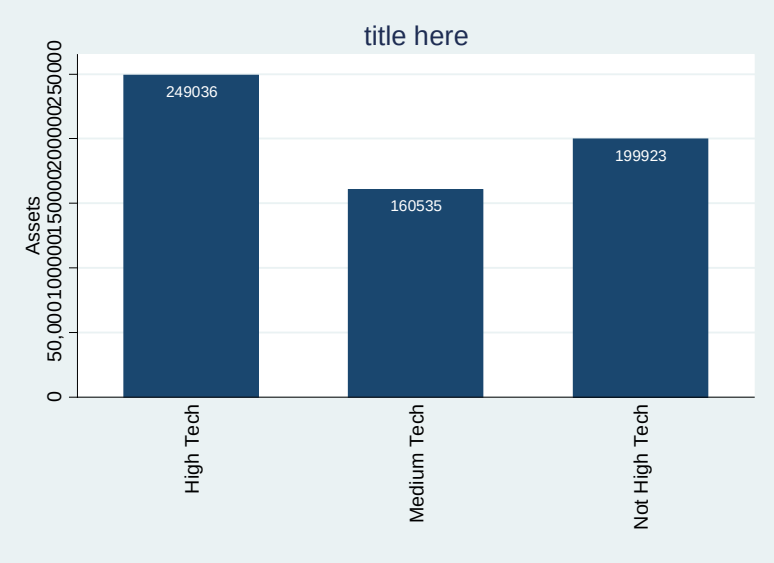
Variable	Obs=.	Obs>.	Obs<.	Obs<. Unique values	Obs<. Min	Obs<. Max
credrisk	812		2,328	5	1	5
d3_a_have_~t	20		3,120	2	0	1
Equity_O~ors	84		3,056	220	0	1.01e+08
Equity_N~ors	41	1,766	1,333	91	0	1.25e+07
OO_work_ex~r	3		3,137	142	0	60
PO_work_exp	2		3,138	52	0	60
Equity	110		3,030	272	0	1.01e+08
Assets	219		2,921	>500	0	8.00e+07
Pers_~t_Resp	127		3,013	325	0	3600000
Pers_Deb~ers	44	2,240	856	73	0	550000
Debt_Bus	136		3,004	211	0	1.20e+07
e2a_f~h_plan	23	1,235	1,882	2	0	1
e2b_p~h_plan	108	2,452	580	2	0	1

4.4. Graphical EDA

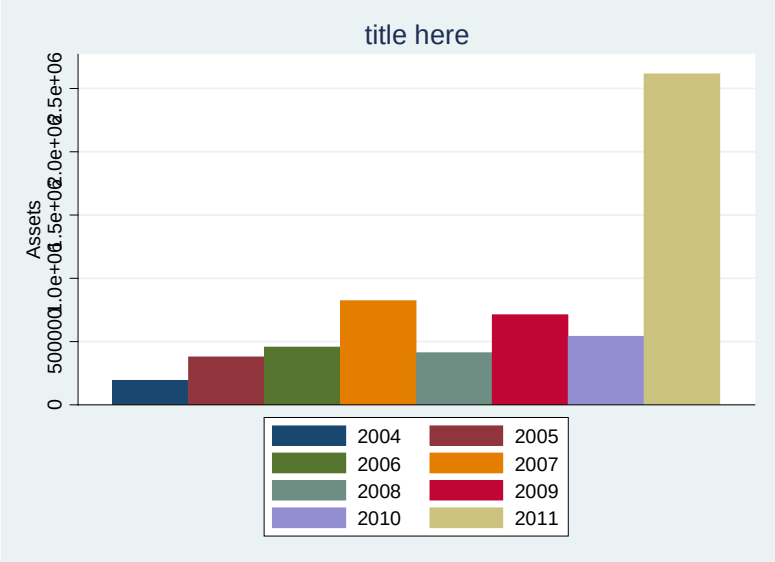
While the graph commands in Stata do not work with the prefix survey (svy), most of Stata's graph commands accept the pweight option. None of the graph commands is suitable to work with MI data, but we will show some examples of how to graph MI data. Given that we are taking an example approach, rather than a syntax approach, we encourage users to read the Stata graphics reference manual.

Example 4.7: Graphs Using KFS in Wide Format

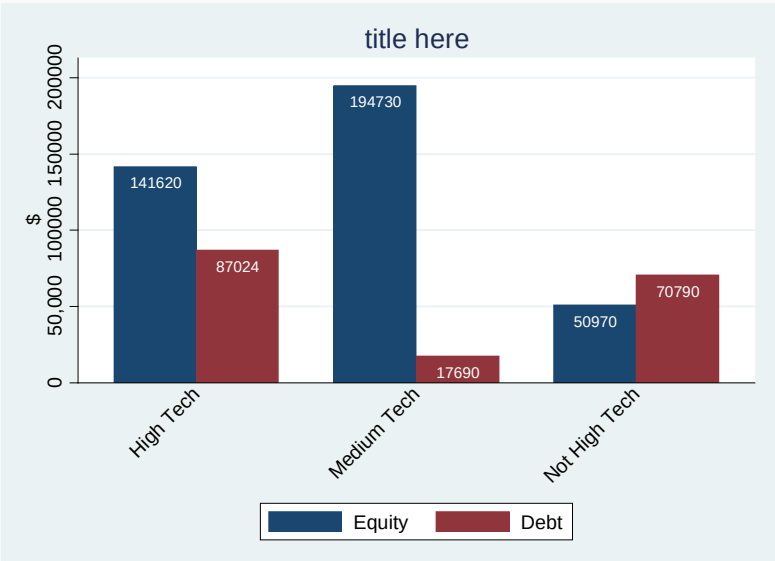
```
*Example 4.7: Using KFS in Wide Format
use KFS8_L7_w1,clear
*Or we can use the Longitudinal_wide_MI_Long_w2 where we can extract the
original KFS data (m=0)
*Extract the original KFS data (m=0)
use Longitudinal_wide_MI_Long_w2,clear
mi extract 0
*Bar chart
graph bar (mean)   Assets_0 [ pw=  wgt_7_long],
over(sampleinfo_strata_0,relabel(1 "High Tech" 2 "Medium Tech"  3 "Not High
Tech") label( angle(90)))  blabel(bar, position(inside) format(%9.0f)
color(white) )  title( title here )  ytitle(Assets)
```



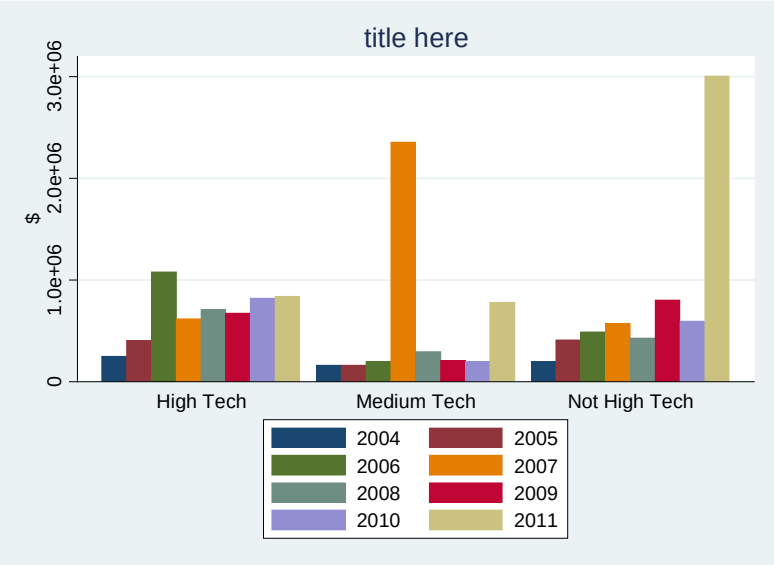
```
graph bar (mean)   Assets_* [ pw=  wgt_7_long], title( title here )
ytitle(Assets) yvaroptions(relabel( 1 "2004" 2 "2005" 3  "2006" 4 "2007" 5
"2008" 6 "2009" 7 "2010" 8 "2011"))
```



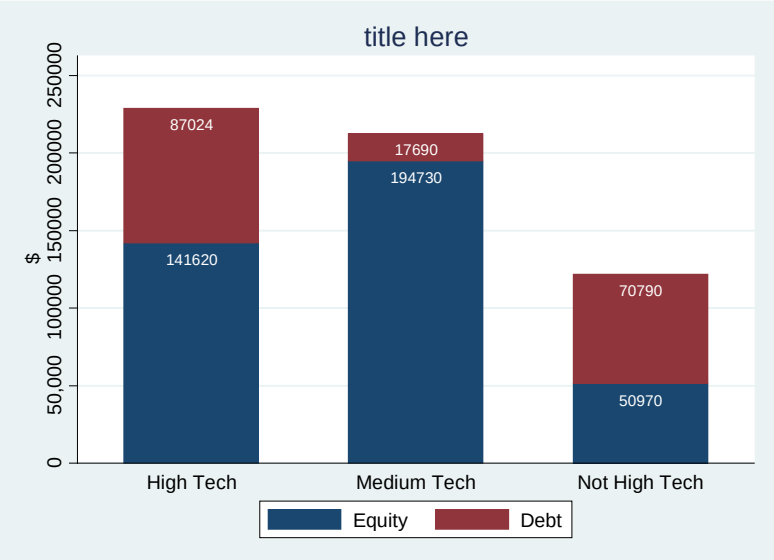

```
*Bar chart with multiple bars graphed over another variable
graph bar (mean)   Equity_0 Debt_0 [ pw=  wgt_7_long],
over(sampleinfo_strata_0,relabel(1 "High Tech" 2 "Medium Tech"  3 "Not High
Tech") label( angle(45))) blabel(bar, position(inside) format(%9.0f)
color(white))  title( title here)  ytitle($) ///
yvaroptions(relabel( 1 "Equity" 2 "Debt" ))
```



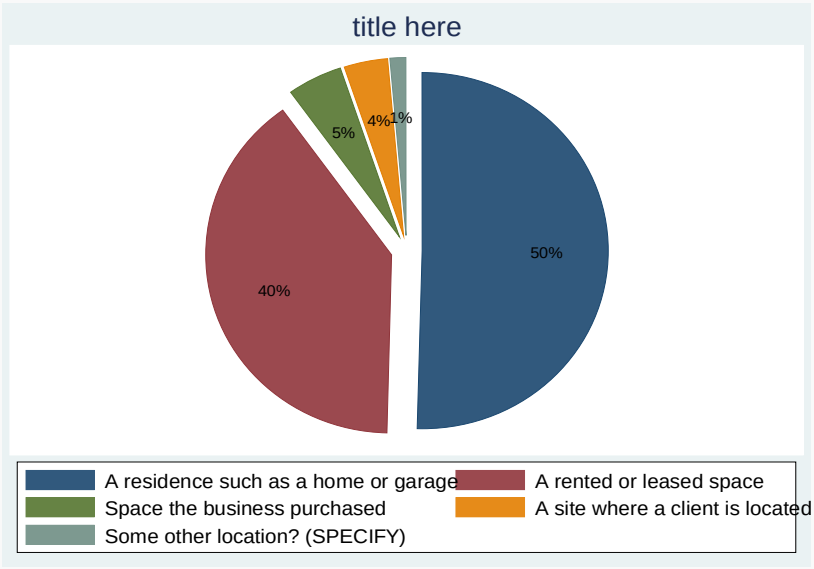

```
*Bar chart with multiple bars graphed over another variable
graph  bar (mean)  Assets_* [ pw=  wgt_7_long], over(sampleinfo_strata_0,relabel(1
"High Tech" 2 "Medium Tech"  3 "Not High Tech") )    title( title here)  ytitle($)
yvaroptions(relabel( 1 "2004" 2 "2005" 3  "2006" 4 "2007" 5 "2008" 6 "2009" 7
"2010" 8 "2011"))
```



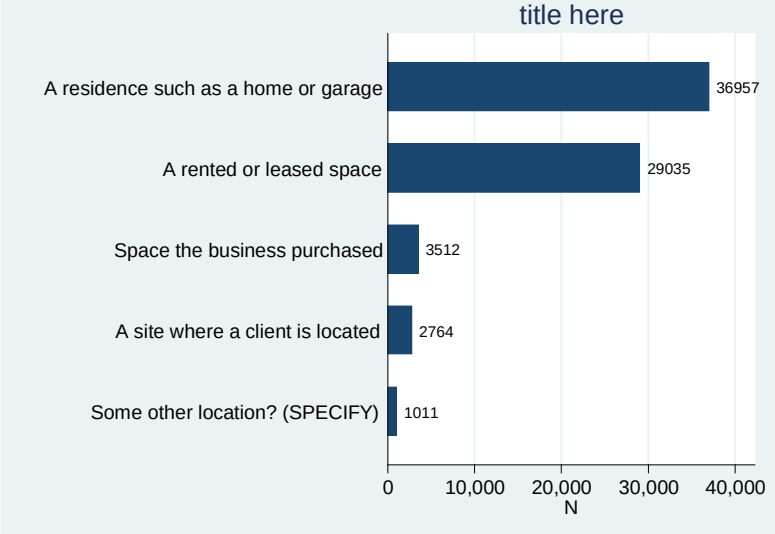

```
*Stacked bar chart
graph bar (mean) Equity_0 Debt_0 [ pw= wgt_7_long],
over(sampleinfo_strata_0,relabel(1 "High Tech" 2 "Medium Tech" 3 "Not High Tech")
) blabel(bar, position(inside) format(%9.0f) color(white)) title( title here)
ytitle($) stack ///
                yvaroptions(relabel( 1 "Equity" 2 "Debt" ))
```



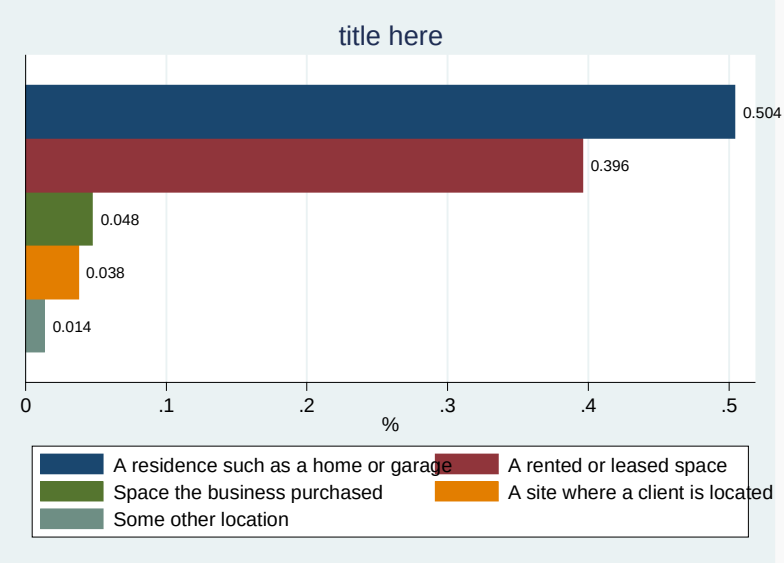

```
*Pie chart - Categorical Variable
graph pie [pweight = wgt_7_long], over(c8_primary_loc_0) pie(_all,explode)
plabel(_all percent, format(%9.0f) ) title(title here)
```




```
*Bar chart - Categorical Variable : frequencies
generate freq = 1
graph hbar (count)   freq [ pw=  wgt_7_long], over(c8_primary_loc_0 ) blabel(bar,
format(%9.0f) color(none))  title(title here)  ytitle(N)
```

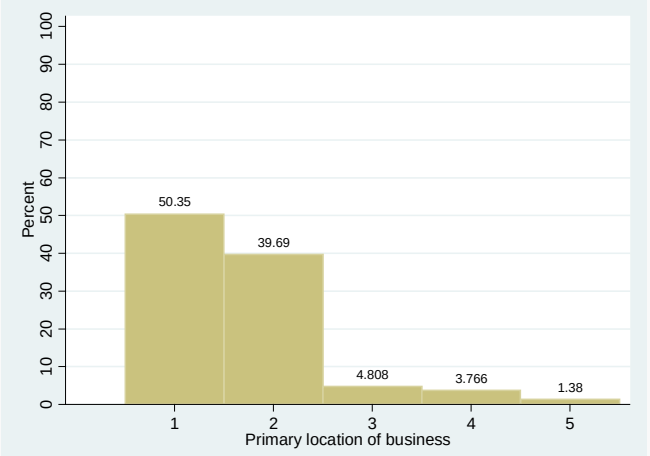


```
*Bar chart - Categorical Variable :  proportions
 tab c8_primary_loc_0, gen(Location)
#delimit ;
graph hbar (mean)  Location* [ pw=  wgt_7_long],  blabel(bar,   format(%9.3f)
color(none))  title(title here)  ytitle(%)
yvaroptions(relabel( 1 "A residence such as a home or garage" 2 "A rented or leased
space" 3  "Space the business purchased" 4    "A site where a client is located"
                          5 "Some other location"));
```

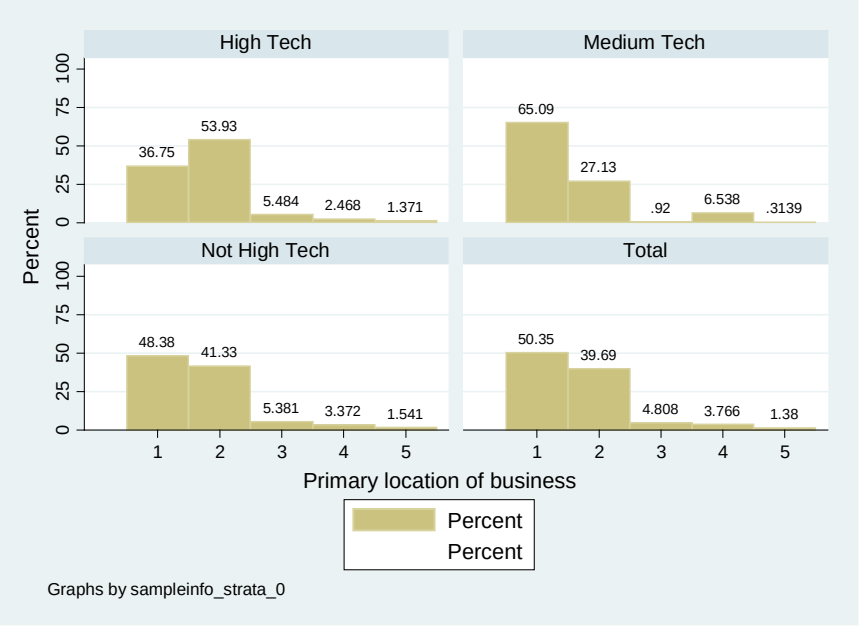


While the histogram command in Stata does not allow for pweight, it allows the use of fweight. The trick is to provide the pweight as an fweight in the histogram command. Because that fweight must be an integer, we need to round the weight values. It is important to emphasize that we can replace fweight by pweight only for the histogram command; thus, we should not do this in any other Stata command.

```
gen fwgt_7_long=int(wgt_7_long)
histogram c8_primary_loc_0 [ fweight= fwgt_7_long], discrete  percent  addlabels
ylabel(0(10)100,grid) xlabel(1(1)5)
```



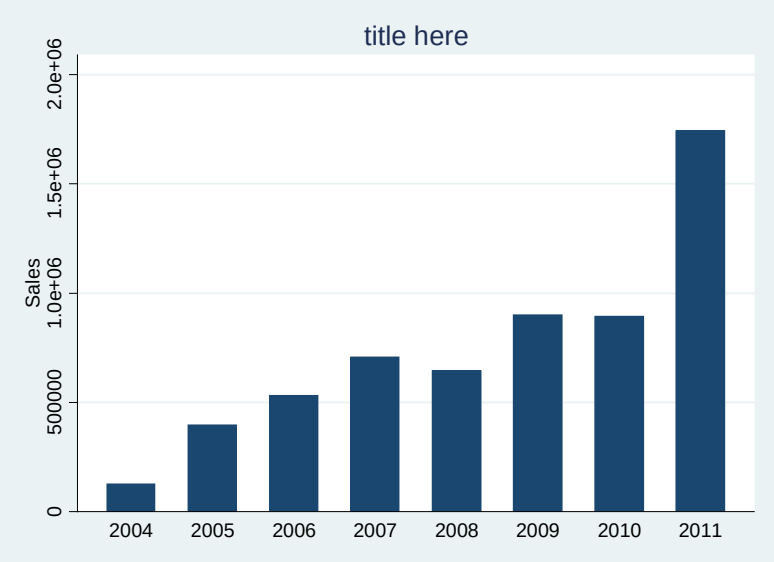
```
label define tech 10 "High Tech" 20 "Medium Tech"  30 "Not High Tech"
label values  sampleinfo_strata_0 tech
histogram c8_primary_loc_0 [ fweight= fwgt_7_long], discrete  percent  addlabels
ylabel(0(25)100,grid) xlabel(1(1)5) by(sampleinfo_strata_0,total)
```



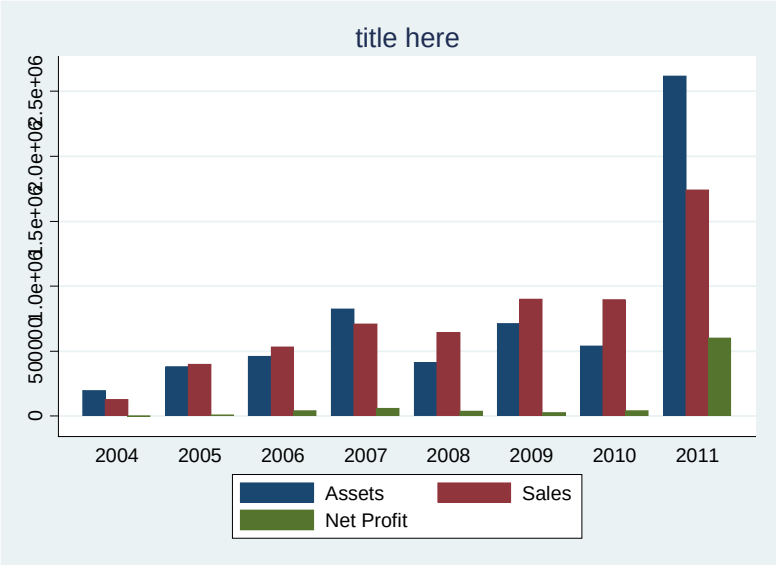
Example 4.8: Graphs Using KFS in Long Format

Given that most users will be using panel data in their analysis, we will focus more on illustrating how to create graphs for panel data. In addition, because we can always extract the original KFS data (m=0) from the KFS MI data, we will be using the KFS MI data from now on. Using the KFS MI files will allow us to compare the KFS MI data to the original data using graphics.

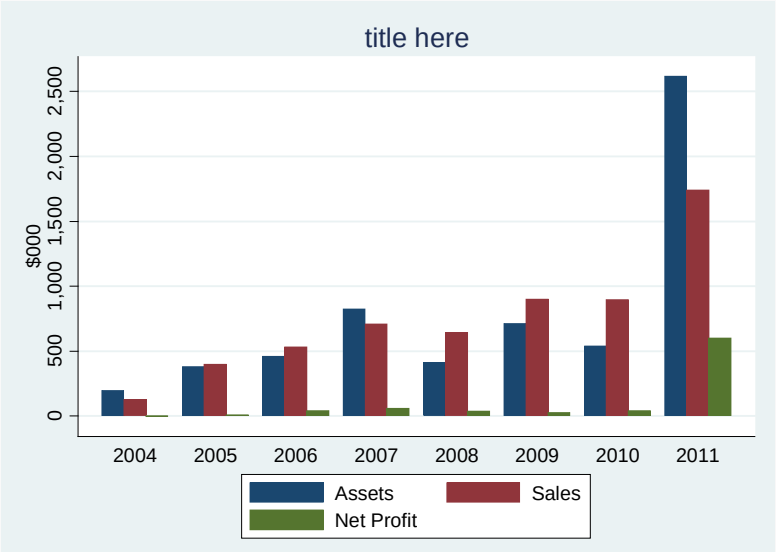
```
*Example 4.8: Using KFS in Long Format
use Longitudinal_Long_MI_Long_L2,clear
*Extract the original KFS data (m=0)
mi extract 0
*The data now the same as the data in the KFS8_L7_L1 file
*Bar chart :original KFS data (m=0)
graph bar (mean)    f16a_rev_amt [ pw=  wgt_7_long], over(year)   title( title
here )   ytitle(Sales)
```



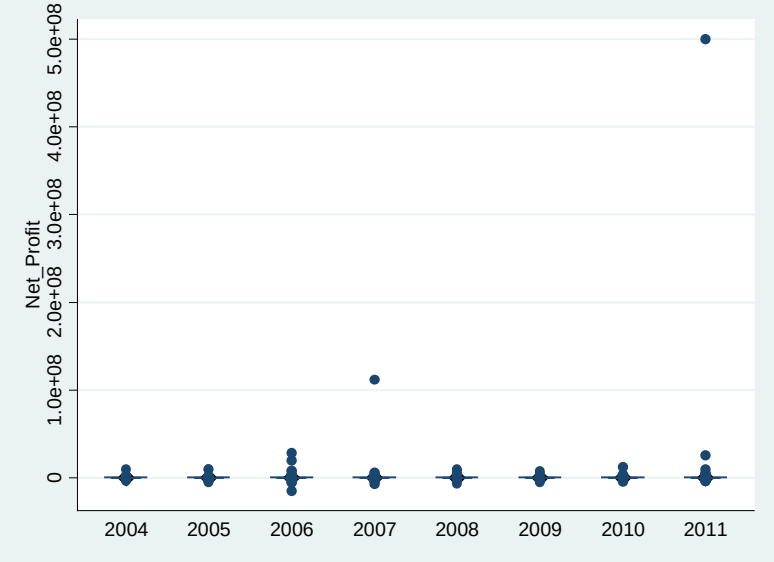
```
graph bar (mean)   Assets f16a_rev_amt Net_Profit [ pw=  wgt_7_long], over(year)
title( title here )  yvaroptions(relabel( 1 "Assets" 2 "Sales" 3  "Net Profit" ))
```




```
*Rescale Y
*Name the temporary variable
tempvar x1 x2 x3
*Generate the temp variable
gen `x1' =  Assets/1000
gen `x2' =  f16a_rev_amt/1000
gen `x3' =  Net_Profit/1000
graph bar (mean)   `x1' `x2' `x3' [ pw=  wgt_7_long], over(year)   title( title
here )  yvaroptions(relabel( 1 "Assets" 2 "Sales" 3  "Net Profit" ))   ytitle($000)
```

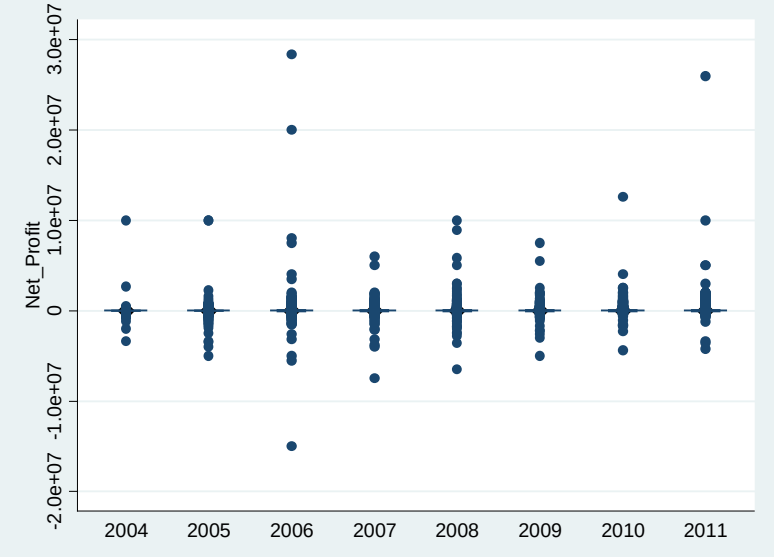


```
*Box plot with symbol as median
graph box Net_Profit [ pw= wgt_7_long] , over(year) medtype(marker)
medmarker(msymbol(diamond))
```



*Note the outlying case in the upper right

```
graph box Net_Profit [ pw= wgt_7_long]if Net_Profit<100000000, over(year)
medtype(marker) medmarker(msymbol(diamond))
```

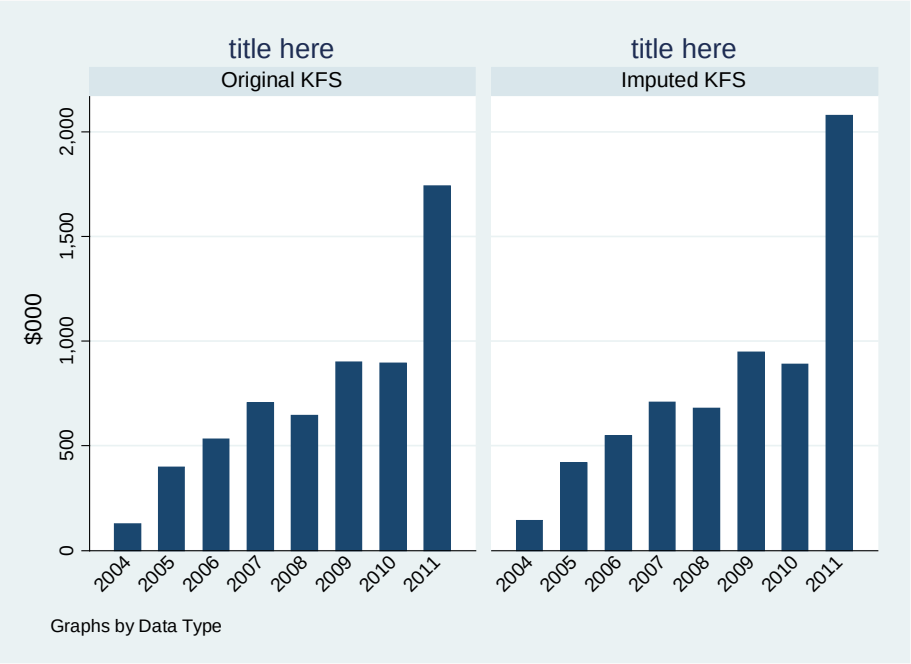


Example 4.9: Graphs Using KFS MI Data

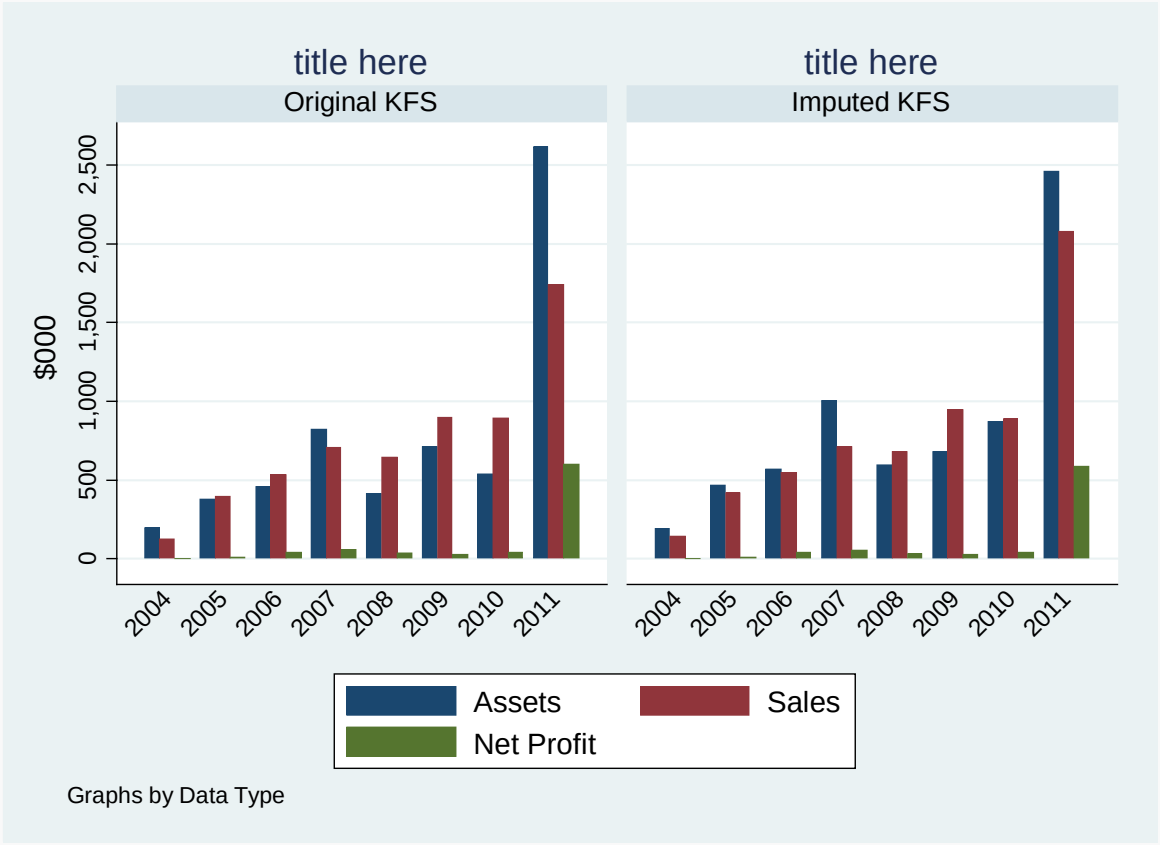
To graph MI data or to compare the KFS MI data to the original data, we need to make a dataset of summary statistics. To create a dataset of summary statistics, we can use the collapse command, which can produce weighted summary statistics. The major two steps to create dataset of summary statistics are:

1. Compute weighted mean for the variable of interest for each m (m=0, 1, 2, 3, 4, 5)
2. Rubin (1987) presented a method for combining results from a data analysis performed m times to obtain a single set of results. The overall mean for the MI data is the average of the individual means among m=1, 2,3,4,5.

```
*Example 4.9: Graphs Using KFS MI Data
use Longitudinal_Long_MI_Long_L2,clear
* To compare the KFS MI data to the original data we need to make dataset of
summary statistics
* Step1 : collapse converts the dataset in memory into a dataset of means, sums,
medians, etc.
collapse (mean) Assets    f16a_rev_amt Net_Profit [ pw=  wgt_7_long], by(master
year)
* Step 2:  create a new var to mark imputed data
recode master  (1/5=1 "Imputed KFS" ) (0=0 "Original KFS" ), gen (data_id)
label variable data_id "Data Type"
* Step 3: calculate mean for MI data using Rubin's rule
collapse (mean) Assets    f16a_rev_amt Net_Profit , by(data_id year)
* Amount in $000
replace Assets =Assets /1000
replace f16a_rev_amt =f16a_rev_amt /1000
replace Net_Profit  =Net_Profit /1000
label variable f16a_rev_amt "Sales"
*Bar chart :original KFS data (m=0) vs. Imputed
graph bar (mean)    f16a_rev_amt , over(year,label( angle(45)))    by(data_id)
title( title here )    ytitle($000)
```



```
graph bar (mean)  Assets f16a_rev_amt Net_Profit , over(year,label( angle(45)))
by(data_id)   title( title here )    ytitle($000)  yvaroptions(relabel( 1 "Assets"
2 "Sales" 3  "Net Profit" ))
```



4.5. Descriptive non-graphical EDA

In this section, we will focus on how to generate descriptive statistics under a complex sample design using Stata.

The KFS is longitudinal data in nature; thus, measurements taken on the same business (owner) tend to be more similar than measurements taken on different businesses (owners). Also, measurements taken close in time on the same business (owner) tend to be more similar than measurements taken far apart in time. Because of the longitudinal data nature, the assumption that observations are independent is not appropriate.

In addition, complex sample design generates sampled observations that are not independent; thus, any inference about differences of descriptive statistics for two subpopulations (or over time) should take into account that the two estimates are correlated.

All those complexities in longitudinal data and complex sample design required a special methods of statistical analysis that accounts for the correlated measurements to draw valid statistical inferences.

4.5.1. Descriptive Statistics: Using KFS Original Data

Stata provides two ways to analyze survey data. The first utilizes commands that begin with "svy:"; svy commands execute while accounting for the survey settings identified by svyset. The second way is to use standard commands that allow the pweight. These commands handle the sampling weights properly, but they will not account for stratification and finite population corrections, and they do not allow for sub population analysis. As we explained in chapter one, stratification and finite population corrections do not affect the parameter estimates, but they affect the variance, standard error, and confidence intervals. Stratification usually makes standard errors smaller; thus, ignoring stratification gives us a more conservative estimate of standard errors.
The following table shows descriptive statistics commands that support survey command.

Survey commands	Descriptive statistics
svy: mean	Estimate means
svy: total	Estimate totals
svy: proportion	Estimate proportions
svy: ratio	Estimate ratios
svy: tabulate oneway	One-way tables for survey data
svy: tabulate twoway	Two-way tables for survey data

The following table shows descriptive statistics standard commands that allow the pweight.

Standard commands with pweight	Descriptive statistics
mean	Estimate means
proportion	Estimate proportions
ratio	Estimate ratios
total	Estimate totals
tabulate oneway	One-way tables
tabulate twoway	Two-way
table	Tables of summary statistics
collapse	Make dataset of summary statistics
_pctile	Compute percentiles and store them in r()
pctile	Create variable containing percentiles
xtile	Create variable containing quantile categories

In addition to the standard commands that allow the pweight, we have few other standard commands that accept aweight—but not pweight—that can be used with KFS. These standard commands are:

Standard commands with aweight	Descriptive statistics
summarize	Summary statistics
correlate	Correlations of variables or coefficients
tabstat	Display table of summary statistics

Stata uses casewise (listwise) deletion in all its commands; if an observation has a missing value in any of the variables, it will be excluded from all the calculations; thus, we need be careful not to include any variable that has a missing value due to skip logic (hard missing) in the variable list.

Example 4.10: Estimating the Mean Value

```
*KFS format : Wide
*Example 4.10: Estimating the Mean Value
use Longitudinal_wide_MI_Long_w2,clear
/*Or We can use Cross_Sectional_wide_MI_Long_w2 with cswgt_final_x as weight and
varlist_x*/
*Extract the original KFS data (m=0)
mi extract 0
*Declare survey design for dataset (In this example we use the eight years panel
data)
svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_0)
vce(linearized)  clear
* Stata use casewise (listwise) deletion in all its commands: if an observation
has a missing value in any of the variables, it is to be excluded from all the
calculations.
svy: mean  Assets_0 Total_Employees_0 Net_Profit_0  PO_gender_0 OO_gender_owner_0
```

```
Number of strata =       6          Number of obs    =    2560
Number of PSUs   =    2560          Population size  = 59428.1
                                    Design df        =    2554

------------------------------------------------------------------
                  |             Linearized
                  |       Mean   Std. Err.     [95% Conf. Interval]
------------------+-----------------------------------------------
         Assets_0 |   204375.2   46232.29      113718.6    295031.8
Total_Employees_0 |   2.376467   .1263324      2.128743    2.624192
     Net_Profit_0 |  -4657.031   4056.582     -12611.55    3297.493
      PO_gender_0 |   .7002383   .0081666      .6842244    .7162521
OO_gender_owner_0 |   .6848716    .007471      .6702217    .6995215
------------------------------------------------------------------



*Postestimation statistics for survey data
*Design and misspecification effects for point estimates
estat effects, deff deft meff meft

------------------------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.       DEFF      DEFT      MEFF      MEFT
-------------+----------------------------------------------------------------
    Assets_0 |   204375.2   46232.29     1.50644   1.22737   1.42339   1.19306
Total_Empl~0 |   2.376467   .1263324     1.47251   1.21347   1.71722   1.31043
Net_Profit_0 |  -4657.031   4056.582     1.13196   1.06394   .810368   .900205
 PO_gender_0 |   .7002383   .0081666      .81308   .901709   .881584   .938927
OO_gender_~0 |   .6848716    .007471     .800028   .894443   .856072   .925242
------------------------------------------------------------------------------


*KFS format : Long
use Longitudinal_Long_MI_Long_L2,clear
*Or We can use Cross_Sectional_Long_MI_Long_w2 with cswgt_final
*Extract the original KFS data (m=0)
mi extract 0
*Declare survey design for dataset (In this example we use the eight years panel
data)
svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata )
vce(linearized)  clear
*Declare data to be panel data
xtset mprid year
svy:  mean    Assets  Total_Employees  Net_Profit    PO_gender  OO_gender_owner  if
year==2004

Number of strata =       6          Number of obs    =    2560
Number of PSUs   =    2560          Population size  = 59428.1
                                    Design df        =    2554
------------------------------------------------------------------
                |             Linearized
                |       Mean   Std. Err.     [95% Conf. Interval]
----------------+-------------------------------------------------
         Assets |   204375.2   46232.29      113718.6    295031.8
Total_Employees |   2.376467   .1263324      2.128743    2.624192
     Net_Profit |  -4657.031   4056.582     -12611.55    3297.493
      PO_gender |   .7002383   .0081666      .6842244    .7162521
OO_gender_owner |   .6848716    .007471      .6702217    .6995215
------------------------------------------------------------------
```

```
* Comparing “svy: mean” vs.  “mean [pweight]” vs. "sum [aweight]" vs "tabstat [aweight]" vs. "table [pweight]" vs. “collapse [pweight]”

use Longitudinal_wide_MI_Long_w2,clear
/*Or We can use Cross_Sectional_wide_MI_Long_w2 with cswgt_final_x as weight and varlist_x*/
*Extract the original KFS data (m=0)
mi extract 0
*Declare survey design for dataset (In this example we use the eight years panel data)
svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_0)
vce(linearized)  clear

svy: mean  Assets_0
estat sd

Number of strata =       6          Number of obs    =    2921
Number of PSUs   =    2921          Population size  = 68043.6
                                    Design df        =    2915

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    Assets_0 |   195562.7   40631.34      115893.7    275231.7
--------------------------------------------------------------

. estat sd

------------------------------------
             |       Mean   Std. Dev.
-------------+----------------------
    Assets_0 |   195562.7     1793371
------------------------------------

svyset,clear
quietly mean  Assets_0  [pweight=wgt_7_long]
matrix b= r(table)
matrix list b

b[9,1]
         Assets_0
     b  195562.69
    se  40645.102
     t    4.81147
pvalue  1.574e-06
    ll  115866.72
    ul  275258.66
    df       2920
  crit  1.9607767
 eform          0

 estat sd

------------------------------------
             |       Mean   Std. Dev.
-------------+----------------------
    Assets_0 |   195562.7     1793371
------------------------------------
```

```
quietly sum  Assets_0  [aweight=wgt_7_long],d
return list
scalars:
                  r(N) =  2921
              r(sum_w) =  68043.63169658184
               r(mean) =  195562.6882448013
                r(Var) =  3216180071363.471
                 r(sd) =  1793371.147131422
           r(skewness) =  26.20285127050286
           r(kurtosis) =  866.9437035331393
                r(sum) =  13306795532.52272
                r(min) =  0
                r(max) =  80000600
                 r(p1) =  0
                 r(p5) =  0
                r(p10) =  2
                r(p25) =  3080
                r(p50) =  19000
                r(p75) =  70500
                r(p90) =  257500
                r(p95) =  550000
                r(p99) =  2010000
```

```
tabstat  Assets_0  [aweight=wgt_7_long],  statistics(       mean n p1     p5   p10
       p25    p50    p75    p90    p95    p99    sd     semean   )
columns(variables)   format (%20.4f)
```

```
   stats |  Assets_0
---------+----------
    mean |         195562.6882
       N |           2921.0000
      p1 |              0.0000
      p5 |              0.0000
     p10 |              2.0000
     p25 |           3080.0000
     p50 |          19000.0000
     p75 |          70500.0000
     p90 |         257500.0000
     p95 |         550000.0000
     p99 |        2010000.0000
      sd |        1793371.1471
se(mean) |          33182.1406
--------------------
```

```
local p " 1 5 10 25 50 75 90 95 99"
foreach x of local p {
_pctile Assets_0 [pweight=wgt_7_long], p(`x')
return list
}

scalars:
                  r(r1) =  0
scalars:
                  r(r1) =  0
scalars:
                  r(r1) =  2
scalars:
                  r(r1) =  3080
scalars:
                  r(r1) =  19000
scalars:
                  r(r1) =  70500
scalars:
                  r(r1) =  257500
scalars:
                  r(r1) =  550000
scalars:
                  r(r1) =  2010000


gen All_Obs=1
*select up to five statistics
table All_Obs[pweight=wgt_7_long], contents(mean  Assets_0 n Assets_0    p1
Assets_0 p5 Assets_0       p10 Assets_0  )

-------------------------------------------------------------------------------
  All_Obs | mean(Asse~0)   N(Assets_0)  p1(Assets_0)  p5(Assets_0)  p10(Asset~0)
----------+--------------------------------------------------------------------
        1 |     195562.7      68,043.6             0             0             2
-------------------------------------------------------------------------------

table All_Obs[pweight=wgt_7_long], contents(p25 Assets_0    p50 Assets_0  p75
Assets_0     p90 Assets_0 p95 Assets_0)
-------------------------------------------------------------------------------
  All_Obs | p25(Asset~0)  med(Asset~0)  p75(Asset~0)  p90(Asset~0)  p95(Asset~0)
----------+--------------------------------------------------------------------
        1 |         3080         19000         70500        257500        550000
-------------------------------------------------------------------------------

table All_Obs[pweight=wgt_7_long], contents(   p95 Assets_0  p99 Assets_0    )
----------------------------------------
  All_Obs | p95(Assets_0)  p99(Assets_0)
----------+-----------------------------
        1 |        550000        2010000
----------------------------------------

collapse (mean )Assets_0 (p1)    p1=Assets_0 (p5)    p5=Assets_0           (p10)
       p10=Assets_0 (p25)  p25=Assets_0 (p50)  p50=Assets_0 (p75)  p75=Assets_0
(p90)  p90=Assets_0 (p95)  p95=Assets_0 (p99)  p99=Assets_0   [ pw=  wgt_7_long]

list
     +-------------------------------------------------------------------------+
     | Assets_0   p1   p5   p10    p25     p50     p75      p90      p95       p99 |
     |-------------------------------------------------------------------------|
  1. | 195562.7    0    0     2   3080   19000   70500   257500   550000   2010000 |
     +-------------------------------------------------------------------------+
```

As we can see, all standard commands with pweight option (aweight) that we used report the correct mean and percentiles, but not the correct standard error (mean).

Example 4.11: Estimating the Mean Value of Subpopulation

```
*KFS format : Wide
*Example 4.11: Estimating the Mean Value for subpopulation
use Longitudinal_wide_MI_Long_w2,clear
 /*Or We can use Cross_Sectional_wide_MI_Long_w2 with cswgt_final_x as weight and
varlist_x*/
*Extract the original KFS data (m=0)
mi extract 0
*Declare survey design for dataset (In this example we use the eight years panel
data)
svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_0)
vce(linearized)  clear

label define gender 0 "Female" 1 "Male"
label values  PO_gender_* gender
*subpop(subpop) specifies that estimates be computed for the single subpopulation
identified by subpop.
*subpopulation is defined by the observations for which varname!=0 that also meet
the if conditions.
*Typically,  varname=1  defines  the  subpopulation,  and  varname=0  indicates
observations not belonging to the subpopulation
svy: mean Assets_0 Total_Employees_0 Net_Profit_0 ,subpop (if PO_gender_0==1  )



 Survey: Mean estimation

Number of strata =       6          Number of obs    =    2722
Number of PSUs   =    2722          Population size  =   63820
                                    Subpop. no. obs  =    1888
                                    Subpop. size     = 41613.9
                                    Design df        =    2716

--------------------------------------------------------------------------
                  |             Linearized
                  |       Mean   Std. Err.     [95% Conf. Interval]
------------------+-------------------------------------------------------
         Assets_0 |   265127.4   65867.71      135971.5    394283.3
Total_Employees_0 |   2.682674   .1662145      2.356754    3.008594
     Net_Profit_0 |   -5675.02   5759.809     -16969.07    5619.032
--------------------------------------------------------------------------
```

```
*Estimates for multiple subpopulations
svy: mean Assets_0 Total_Employees_0 Net_Profit_0, over(PO_gender_0)

Number of strata =       6          Number of obs    =    2560
Number of PSUs   =    2560          Population size  = 59428.1
                                    Design df        =    2554

       Female: PO_gender_0 = Female
         Male: PO_gender_0 = Male

--------------------------------------------------------------------
                  |             Linearized
             Over |       Mean   Std. Err.     [95% Conf. Interval]
------------------+-------------------------------------------------
Assets_0          |
           Female |   62459.08   8733.229      45334.15    79584.01
             Male |   265127.4    65868.6      135966.1    394288.7
------------------+-------------------------------------------------
Total_Employees_0 |
           Female |   1.661174   .1615848      1.344323    1.978024
             Male |   2.682674   .1662169       2.35674    3.008608
------------------+-------------------------------------------------
Net_Profit_0      |
           Female |  -2279.026   1435.326     -5093.546    535.4949
             Male |   -5675.02   5759.864      -16969.5    5619.459
--------------------------------------------------------------------


*KFS format : Long
use Longitudinal_Long_MI_Long_L2,clear
*Or We can use Cross_Sectional_Long_MI_Long_w2 with cswgt_final
*Extract the original KFS data (m=0)
mi extract 0
*Declare survey design for dataset (In this example we use the eight years panel
data)
svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
vce(linearized)  clear

label define gender 0 "Female" 1 "Male"
label values  PO_gender  gender


*subpop(subpop) specifies that estimates be computed for the single subpopulation
identified by subpop.
svy: mean Assets Total_Employees Net_Profit if year==2004,subpop (if PO_gender==1
)

Number of strata =       6          Number of obs    =    2722
Number of PSUs   =    2722          Population size  =   63820
                                    Subpop. no. obs  =    1888
                                    Subpop. size     = 41613.9
                                    Design df        =    2716

--------------------------------------------------------------------
                |            Linearized
                |       Mean   Std. Err.     [95% Conf. Interval]
----------------+---------------------------------------------------
         Assets |   265127.4   65867.71      135971.5    394283.3
Total_Employees |   2.682674   .1662145      2.356754    3.008594
     Net_Profit |   -5675.02   5759.809     -16969.07    5619.032
--------------------------------------------------------------------
```

```
*Estimates for multiple subpopulations
svy: mean Assets Total_Employees Net_Profit if year==2004, over(PO_gender)


Number of strata =       6          Number of obs    =    2560
Number of PSUs   =    2560          Population size  = 59428.1
                                   Design df        =    2554

       Female: PO_gender = Female
         Male: PO_gender = Male

--------------------------------------------------------------
                |             Linearized
           Over |       Mean   Std. Err.     [95% Conf. Interval]
----------------+---------------------------------------------
Assets          |
         Female |   62459.08   8733.229      45334.15    79584.01
           Male |   265127.4    65868.6      135966.1    394288.7
----------------+---------------------------------------------
Total_Employees |
         Female |   1.661174   .1615848      1.344323    1.978024
           Male |   2.682674   .1662169       2.35674    3.008608
----------------+---------------------------------------------
Net_Profit      |
         Female |  -2279.026   1435.326     -5093.546    535.4949
           Male |   -5675.02   5759.864      -16969.5    5619.459
--------------------------------------------------------------
```

Example 4.12: Estimating the Population Totals

```
*KFS format : Wide
*Example 4.12: Estimating the Population Totals
use Longitudinal_wide_MI_Long_w2,clear
/*Or We can use Cross_Sectional_wide_MI_Long_w2 with cswgt_final_x as weight and
varlist_x*/
*Extract the original KFS data (m=0)
mi extract 0
*Declare survey design for dataset (In this example we use the eight years panel
data)
svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_7)
vce(linearized)  clear
svy:   total     d3_a_num_patent_6   Total_Employees_6     c4_numowners_confirm_6
PO_gender_6

Survey: Total estimation

Number of strata =       6          Number of obs    =    1707
Number of PSUs   =    1707          Population size  = 36776.7
                                   Design df        =    1701

------------------------------------------------------------------------------
                       |             Linearized
                       |      Total   Std. Err.     [95% Conf. Interval]
-----------------------+------------------------------------------------------
     d3_a_num_patent_6 |   4998.258   2394.225      302.3225    9694.193
     Total_Employees_6 |   164427.4   11995.11      140900.7    187954.1
c4_numowners_confirm_6 |    50683.2   995.2998      48731.06    52635.34
           PO_gender_6 |   26292.71   427.5696      25454.09    27131.33
------------------------------------------------------------------------------
```

```
*Estimating the Population Totals of Subpopulation
svy:  total      d3_a_num_patent_6  Total_Employees_6    c4_numowners_confirm_6
PO_gender_6 , over(Home_Based_6)


Number of strata =       6          Number of obs    =    1707
Number of PSUs   =    1707          Population size  = 36776.7
                                    Design df        =    1701

    _subpop_1: Home_Based_6 = Non Home Based
    _subpop_2: Home_Based_6 = Home Based
------------------------------------------------------------------------------
                             |             Linearized
                        Over |      Total   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
d3_a_num_patent_6            |
                    _subpop_1 |   1995.434   494.0567      1026.411    2964.456
                    _subpop_2 |   3002.824   2343.345     -1593.317    7598.966
-----------------------------+------------------------------------------------
Total_Employees_6            |
                    _subpop_1 |   135059.7   11884.02      111750.9    158368.5
                    _subpop_2 |   29367.69    2880.68      23717.64    35017.74
-----------------------------+------------------------------------------------
c4_numowners_confirm_6       |
                    _subpop_1 |   28986.73    1206.08      26621.18    31352.29
                    _subpop_2 |   21696.47   714.2633      20295.54    23097.39
-----------------------------+------------------------------------------------
PO_gender_6                  |
                    _subpop_1 |   14046.78   515.2875      13036.12    15057.45
                    _subpop_2 |   12245.93   454.9526       11353.6    13138.26
------------------------------------------------------------------------------

*KFS format : Long
use Longitudinal_Long_MI_Long_L2,clear
*Or We can use Cross_Sectional_Long_MI_Long_w2 with cswgt_final
*Extract the original KFS data (m=0)
mi extract 0
*Declare survey design for dataset (In this example we use the eight years panel
data)
svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
vce(linearized)  clear
svy: total  d3_a_num_patent  Total_Employees   c4_numowners_confirm     PO_gender
if year==2010

Number of strata =       6          Number of obs    =    1707
Number of PSUs   =    1707          Population size  = 36776.7
                                    Design df        =    1701

------------------------------------------------------------------------------
                     |             Linearized
                     |      Total   Std. Err.     [95% Conf. Interval]
---------------------+--------------------------------------------------------
     d3_a_num_patent |   4998.258   2394.225      302.3225    9694.193
     Total_Employees |   164427.4   11995.11      140900.7    187954.1
c4_numowners_confirm |    50683.2   995.2998      48731.06    52635.34
           PO_gender |   26292.71   427.5696      25454.09    27131.33
------------------------------------------------------------------------------
```

```
*Estimating the Population Totals of Subpopulation
svy: total  d3_a_num_patent  Total_Employees   c4_numowners_confirm     PO_gender
if year==2010, over(Home_Based)

Number of strata =       6          Number of obs    =    1707
Number of PSUs   =    1707          Population size  = 36776.7
                                    Design df        =    1701

    _subpop_1: Home_Based = Non Home Based
    _subpop_2: Home_Based = Home Based

------------------------------------------------------------------------------
                     |             Linearized
                Over |      Total   Std. Err.     [95% Conf. Interval]
---------------------+--------------------------------------------------------
d3_a_num_patent      |
           _subpop_1 |   1995.434   494.0567      1026.411    2964.456
           _subpop_2 |   3002.824   2343.345     -1593.317    7598.966
---------------------+--------------------------------------------------------
Total_Employees      |
           _subpop_1 |   135059.7   11884.02      111750.9    158368.5
           _subpop_2 |   29367.69    2880.68      23717.64    35017.74
---------------------+--------------------------------------------------------
c4_numowners_confirm |
           _subpop_1 |   28986.73    1206.08      26621.18    31352.29
           _subpop_2 |   21696.47   714.2633      20295.54    23097.39
---------------------+--------------------------------------------------------
PO_gender            |
           _subpop_1 |   14046.78   515.2875      13036.12    15057.45
           _subpop_2 |   12245.93   454.9526       11353.6    13138.26
------------------------------------------------------------------------------
```

Example 4.13: Estimating the Proportions for Binary and Categorical Variables

```
*KFS format : Wide
*Example 4.13: Estimating the Proportion for Binary and Categorical Variables
use Cross_Sectional_wide_MI_Long_w2,clear
/*Or We can use Longitudinal_wide_MI_Long_w2  */
*Extract the original KFS data (m=0)
mi extract 0
*Declare survey design for dataset : In this example we study the baseline survey
svyset         [pweight=cswgt_final_0]     ,     strata(sampleinfo_samplestrata_0)
vce(linearized)  clear
* Binary Variables
svy:mean     PO_race_amind_owner_0   PO_race_asian_owner_0   PO_race_black_owner_0
PO_race_nathaw_owner_0 PO_race_other_owner_0 PO_race_white_owner_0

Number of strata =       6          Number of obs    =    4915
Number of PSUs   =    4915          Population size  =   73055
                                    Design df        =    4909

------------------------------------------------------------------------------
                     |             Linearized
                     |       Mean   Std. Err.     [95% Conf. Interval]
---------------------+--------------------------------------------------------
 PO_race_amind_owner_0 |   .0118253   .0018296      .0082385    .0154121
 PO_race_asian_owner_0 |   .0372986     .00315      .0311231     .043474
 PO_race_black_owner_0 |   .0860292   .0047308      .0767548    .0953036
PO_race_nathaw_owner_0 |   .0056917   .0013593      .0030269    .0083565
 PO_race_other_owner_0 |   .0513871    .003746      .0440433     .058731
 PO_race_white_owner_0 |   .8077681    .006627      .7947762      .82076
------------------------------------------------------------------------------
```

```
svy:      proportion            PO_race_amind_owner_0      PO_race_asian_owner_0
PO_race_black_owner_0        PO_race_nathaw_owner_0        PO_race_other_owner_0
PO_race_white_owner_0

Number of strata =       6          Number of obs    =    4915
Number of PSUs   =    4915          Population size  =   73055
                                    Design df        =    4909

------------------------------------------------------------------------------
                        |             Linearized
                        | Proportion   Std. Err.     [95% Conf. Interval]
------------------------+-----------------------------------------------------
PO_race_amind_owner_0   |
                      0 |   .9881747   .0018296      .9845879    .9917615
                      1 |   .0118253   .0018296      .0082385    .0154121
------------------------+-----------------------------------------------------
PO_race_asian_owner_0   |
                      0 |   .9627014     .00315       .956526    .9688769
                      1 |   .0372986     .00315      .0311231     .043474
------------------------+-----------------------------------------------------
PO_race_black_owner_0   |
                      0 |   .9139708   .0047308      .9046964    .9232452
                      1 |   .0860292   .0047308      .0767548    .0953036
------------------------+-----------------------------------------------------
PO_race_nathaw_owner_0  |
                      0 |   .9943083   .0013593      .9916435    .9969731
                      1 |   .0056917   .0013593      .0030269    .0083565
------------------------+-----------------------------------------------------
PO_race_other_owner_0   |
                      0 |   .9486129    .003746       .941269    .9559567
                      1 |   .0513871    .003746      .0440433     .058731
------------------------+-----------------------------------------------------
PO_race_white_owner_0   |
                      0 |   .1922319    .006627        .17924    .2052238
                      1 |   .8077681    .006627      .7947762      .82076
------------------------------------------------------------------------------
```

```
svy: tab PO_race_amind_owner_0, se ci col

Number of strata   =         6                  Number of obs      =      4915
Number of PSUs     =      4915                  Population size    = 73055.022
                                                 Design df          =      4909

------------------------------------------------------------
PO_race_a |
mind_owne |
r_0       |     column          se          lb          ub
----------+-------------------------------------------------
        0 |      .9882       .0018        .984       .9913
        1 |      .0118       .0018       .0087        .016
          |
    Total |          1
------------------------------------------------------------
  Key:  column    =  column proportions
        se        =  linearized standard errors of column proportions
        lb        =  lower 95% confidence bounds for column proportions
        ub        =  upper 95% confidence bounds for column proportions
```

```
/* The CI using tab is not the same as the ones from mean and proportion commands,
tab use Logit transformation technique while mean and proportion use symmetric
interval  technique*/


*Categorical Variables
svy: proportion  PO_race_group_0

Number of strata =       6          Number of obs    =    4915
Number of PSUs   =    4915          Population size  =   73055
                                    Design df        =    4909

--------------------------------------------------------------
                |             Linearized
                | Proportion   Std. Err.     [95% Conf. Interval]
----------------+---------------------------------------------
PO_race_group_0 |
              1 |   .0118253   .0018296      .0082385    .0154121
              2 |   .0056917   .0013593      .0030269    .0083565
              3 |   .0372986     .00315      .0311231     .043474
              4 |   .0860292   .0047308      .0767548    .0953036
              5 |   .8077681    .006627      .7947762      .82076
              6 |   .0513871    .003746      .0440433     .058731
--------------------------------------------------------------

svy: tab  PO_race_group_0 , se ci col

Number of strata   =         6                  Number of obs      =      4915
Number of PSUs     =      4915                  Population size    = 73055.022
                                                 Design df          =      4909

-----------------------------------------------------------
PO_race_g |
roup_0    |     column          se          lb          ub
----------+------------------------------------------------
        1 |      .0118       .0018       .0087        .016
        2 |      .0057       .0014       .0036       .0091
        3 |      .0373       .0032       .0316        .044
        4 |       .086       .0047       .0772       .0958
        5 |      .8078       .0066       .7944       .8204
        6 |      .0514       .0037       .0445       .0592
          |
    Total |          1
-----------------------------------------------------------
  Key:  column    =  column proportions
        se        =  linearized standard errors of column proportions
        lb        =  lower 95% confidence bounds for column proportions
        ub        =  upper 95% confidence bounds for column proportions
```

```
*KFS format : Long
use Cross_Sectional_Long_MI_Long_L2 ,clear
*Or We can use Longitudinal_Long_MI_Long_L2 with wgt_7_long
*Extract the original KFS data (m=0)
mi extract 0
*Declare survey design for dataset (In this example we use the eight years panel
data)
svyset          [pweight=cswgt_final]     ,     strata(sampleinfo_samplestrata)
      vce(linearized)  clear
* Binary Variables
svy:mean        PO_race_amind_owner    PO_race_asian_owner    PO_race_black_owner
PO_race_nathaw_owner PO_race_other_owner PO_race_white_owner if year==2004

Number of strata =       6          Number of obs    =    4915
Number of PSUs   =    4915          Population size  =   73055
                                    Design df        =    4909
------------------------------------------------------------------------------
                     |             Linearized
                     |       Mean   Std. Err.     [95% Conf. Interval]
---------------------+--------------------------------------------------------
 PO_race_amind_owner |   .0118253   .0018296      .0082385    .0154121
 PO_race_asian_owner |   .0372986     .00315      .0311231     .043474
 PO_race_black_owner |   .0860292   .0047308      .0767548    .0953036
PO_race_nathaw_owner |   .0056917   .0013593      .0030269    .0083565
 PO_race_other_owner |   .0513871    .003746      .0440433     .058731
 PO_race_white_owner |   .8077681    .006627      .7947762      .82076
------------------------------------------------------------------------------

svy:  proportion    PO_race_amind_owner  PO_race_asian_owner  PO_race_black_owner
PO_race_nathaw_owner PO_race_other_owner PO_race_white_owner  if year==2004

Number of strata =       6          Number of obs    =    4915
Number of PSUs   =    4915          Population size  =   73055
                                    Design df        =    4909
------------------------------------------------------------------------------
                     |             Linearized
                     | Proportion   Std. Err.     [95% Conf. Interval]
---------------------+--------------------------------------------------------
PO_race_amind_owner  |
                   0 |   .9881747   .0018296      .9845879    .9917615
                   1 |   .0118253   .0018296      .0082385    .0154121
---------------------+--------------------------------------------------------
PO_race_asian_owner  |
                   0 |   .9627014     .00315       .956526    .9688769
                   1 |   .0372986     .00315      .0311231     .043474
---------------------+--------------------------------------------------------
PO_race_black_owner  |
                   0 |   .9139708   .0047308      .9046964    .9232452
                   1 |   .0860292   .0047308      .0767548    .0953036
---------------------+--------------------------------------------------------
PO_race_nathaw_owner |
                   0 |   .9943083   .0013593      .9916435    .9969731
                   1 |   .0056917   .0013593      .0030269    .0083565
---------------------+--------------------------------------------------------
PO_race_other_owner  |
                   0 |   .9486129    .003746       .941269    .9559567
                   1 |   .0513871    .003746      .0440433     .058731
---------------------+--------------------------------------------------------
PO_race_white_owner  |
                   0 |   .1922319    .006627        .17924    .2052238
                   1 |   .8077681    .006627      .7947762      .82076
```

```
svy: tab PO_race_amind_owner if year==2004 , se ci col


Number of strata   =         6                  Number of obs      =      4915
Number of PSUs     =      4915                  Population size    = 73055.022
                                                 Design df          =      4909

-------------------------------------------------------------
PO_race_a |
mind_owne |
r         |     column          se          lb          ub
----------+--------------------------------------------------
        0 |      .9882       .0018        .984       .9913
        1 |      .0118       .0018       .0087        .016
          |
    Total |          1
-------------------------------------------------------------
  Key:  column    =  column proportions
        se        =  linearized standard errors of column proportions
        lb        =  lower 95% confidence bounds for column proportions
        ub        =  upper 95% confidence bounds for column proportions

*Categorical Variables
svy: proportion  PO_race_group if year==2004

Number of strata =       6          Number of obs    =    4915
Number of PSUs   =    4915          Population size  =   73055
                                    Design df        =    4909

--------------------------------------------------------------------
              |             Linearized
              | Proportion   Std. Err.     [95% Conf. Interval]
--------------+-----------------------------------------------------
PO_race_group |
            1 |   .0118253   .0018296      .0082385    .0154121
            2 |   .0056917   .0013593      .0030269    .0083565
            3 |   .0372986     .00315      .0311231     .043474
            4 |   .0860292   .0047308      .0767548    .0953036
            5 |   .8077681    .006627      .7947762      .82076
            6 |   .0513871    .003746      .0440433     .058731
--------------------------------------------------------------------
```

```
svy: tab  PO_race_group if year==2004, se ci col

Number of strata   =         6                  Number of obs      =      4915
Number of PSUs     =      4915                  Population size    = 73055.022
                                                 Design df          =      4909

------------------------------------------------------------
PO_race_g |
roup      |     column          se          lb          ub
----------+-------------------------------------------------
        1 |      .0118       .0018       .0087        .016
        2 |      .0057       .0014       .0036       .0091
        3 |      .0373       .0032       .0316        .044
        4 |       .086       .0047       .0772       .0958
        5 |      .8078       .0066       .7944       .8204
        6 |      .0514       .0037       .0445       .0592
          |
    Total |          1
------------------------------------------------------------
  Key:  column    =  column proportions
        se        =  linearized standard errors of column proportions
        lb        =  lower 95% confidence bounds for column proportions
        ub        =  upper 95% confidence bounds for column proportions
```

Example 4.14: Estimating Ratios

```
*KFS format : Wide
*Example 4.14: Estimating the Ratio for Continuous Variables
use Longitudinal_wide_MI_Long_w2,clear
/*Or We can use Cross_Sectional_wide_MI_Long_w2 with cswgt_final_x as weight and
varlist_x*/
*Extract the original KFS data (m=0)
mi extract 0
*Declare survey design for dataset (In this example we use the eight years panel
data)
svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_0)
vce(linearized)  clear
svy:ratio  (Full_Time_Ratio_3:  Full_Time_Employees_3/Full_Part_Time_Employees_3)
(Part_Time_Ratio_3: Part_Time_Employees_3/Full_Part_Time_Employees_3)

Number of strata =       6          Number of obs    =    2001
Number of PSUs   =    2001          Population size  = 44303.7
                                    Design df        =    1995

 Full_Time_~3: Full_Time_Employees_3/Full_Part_Time_Employ~3
 Part_Time_~3: Part_Time_Employees_3/Full_Part_Time_Employ~3

------------------------------------------------------------------------------
                  |            Linearized
                  |      Ratio   Std. Err.     [95% Conf. Interval]
------------------+-----------------------------------------------------------
Full_Time_Ratio_3 |   .6566937   .0254331      .6068154     .706572
Part_Time_Ratio_3 |   .3433063   .0254331       .293428    .3931846
------------------------------------------------------------------------------
```

```
*KFS format : Long
*Example 4.14: Estimating the Ratio for Continuous Variables
use Longitudinal_Long_MI_Long_L2,clear
*Or We can use Cross_Sectional_Long_MI_Long_w2 with cswgt_final
*Extract the original KFS data (m=0)
mi extract 0
*Declare survey design for dataset (In this example we use the eight years panel
data)
svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
vce(linearized)  clear
svy:     ratio  (Full_Time_Ratio:  Full_Time_Employees/Full_Part_Time_Employees)
(Part_Time_Ratio: Part_Time_Employees/Full_Part_Time_Employees) if year==2007

Number of strata =       6          Number of obs    =    2001
Number of PSUs   =    2001          Population size  = 44303.7
                                    Design df        =    1995

 Full_Time_~o: Full_Time_Employees/Full_Part_Time_Employ~s
 Part_Time_~o: Part_Time_Employees/Full_Part_Time_Employ~s
------------------------------------------------------------------
                |             Linearized
                |      Ratio   Std. Err.     [95% Conf. Interval]
----------------+-------------------------------------------------
Full_Time_Ratio |   .6566937   .0254331      .6068154     .706572
Part_Time_Ratio |   .3433063   .0254331       .293428    .3931846
------------------------------------------------------------------
```

Example 4.15: One-Way Tables for Survey Data

```
*KFS format : Wide
use Longitudinal_wide_MI_Long_w2,clear
/*Or We can use Cross_Sectional_wide_MI_Long_w2 with cswgt_final_x as weight and
varlist_x*/
*Extract the original KFS data (m=0)
mi extract 0
*Declare survey design for dataset
svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_0)
vce(linearized)  clear
svy: tab  d9a_perc_internet_sales_3     ,  ci  obs   percent  format(%16.4f)

Number of strata   =         6                  Number of obs      =       517
Number of PSUs     =       517                  Population size    = 12050.846
                                                Design df          =       511

------------------------------------------------------------
d9a_perc_ |
internet_ |
sales     | percentages           lb           ub          obs
----------+-------------------------------------------------
 Less tha |     32.4956      27.8776      37.4811     165.0000
 5% - 25% |     30.9659      26.4464      35.8809     162.0000
 26% - 50 |     10.2036       7.4507      13.8216      46.0000
 51% - 75 |      9.2944       6.6483      12.8489      41.0000
 76% - 10 |     17.0405      13.6534      21.0630     103.0000
          |
    Total |    100.0000                               517.0000
------------------------------------------------------------
  Key:  percentages  =  cell percentages
        lb           =  lower 95% confidence bounds for cell percentages
        ub           =  upper 95% confidence bounds for cell percentages
        obs          =  number of observations
```

```
*KFS format : Long
*Example 4.15: One-Way Tables for Survey Data
use Longitudinal_Long_MI_Long_L2,clear
*Or We can use Cross_Sectional_Long_MI_Long_w2 with cswgt_final
*Extract the original KFS data (m=0)
mi extract 0
*Declare survey design for dataset (In this example we use the eight years panel
data)
svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
vce(linearized)  clear
svy:  tab   d9a_perc_internet_sales  if  year==2007       ,   ci   obs     percent
format(%16.4f)

Number of strata   =         6                  Number of obs      =       517
Number of PSUs     =       517                  Population size    = 12050.846
                                                Design df          =       511

------------------------------------------------------------
d9a_perc_ |
internet_ |
sales     | percentages           lb           ub          obs
----------+-------------------------------------------------
 Less tha |     32.4956      27.8776      37.4811     165.0000
 5% - 25% |     30.9659      26.4464      35.8809     162.0000
 26% - 50 |     10.2036       7.4507      13.8216      46.0000
 51% - 75 |      9.2944       6.6483      12.8489      41.0000
 76% - 10 |     17.0405      13.6534      21.0630     103.0000
          |
    Total |    100.0000                               517.0000
------------------------------------------------------------
  Key:  percentages  =  cell percentages
        lb           =  lower 95% confidence bounds for cell percentages
        ub           =  upper 95% confidence bounds for cell percentages
        obs          =  number of observations
```

Example 4.16: Two-Way Tables for Survey Data

```
*KFS format : Wide
use Longitudinal_wide_MI_Long_w2,clear
/*Or We can use Cross_Sectional_wide_MI_Long_w2 with cswgt_final_x as weight and
varlist_x*/
*Extract the original KFS data (m=0)
mi extract 0
*Declare survey design for dataset
svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_0)
vce(linearized)  clear
label define gender_0 0 "Female" 1 "Male"
label values  PO_gender_0  gender
*svy: tabulate produces two-way tabulations with tests of independence for complex
survey data
svy: tab  c8_primary_loc_0 PO_gender_0 ,  se ci

Number of strata   =         6                  Number of obs      =      3140
Number of PSUs     =      3140                  Population size    = 73278.441
                                                 Design df          =      3134

------------------------------------------------------------
Primary   |
location  |
of        |                    PO_gender_0
business  |               0                1            Total
----------+-------------------------------------------------
 A reside |           .1583            .3461            .5043
          |          (.007)          (.0093)          (.0106)
          |    [.145,.1725]    [.3281,.3645]    [.4836,.5251]
          |
 A rented |            .122            .2742            .3962
          |         (.0069)          (.0092)          (.0105)
          |   [.1091,.1363]    [.2565,.2926]     [.3758,.417]
          |
 Space th |           .0102            .0377            .0479
          |         (.0022)          (.0041)          (.0046)
          |   [.0066,.0157]    [.0305,.0466]    [.0397,.0578]
          |
 A site w |           .0109            .0268            .0377
          |         (.0021)          (.0032)          (.0039)
          |   [.0074,.0159]    [.0212,.0339]     [.0308,.046]
          |
 Some oth |           .0017            .0121            .0138
          |       (8.8e-04)          (.0023)          (.0025)
          | [5.9e-04,.0047]    [.0083,.0176]    [.0097,.0196]
          |
    Total |            .303             .697                1
          |         (.0075)          (.0075)
          |    [.2885,.318]     [.682,.7115]
------------------------------------------------------------
  Key:  cell proportions
        (linearized standard errors of cell proportions)
        [95% confidence intervals for cell proportions]

  Pearson:
    Uncorrected   chi2(4)         =   13.7122
    Design-based  F(3.99, 12494.09)=    2.3516    P = 0.0520
```

```
*KFS format : Long
* Example 4.16: One-Way Tables for Survey Data
use Longitudinal_Long_MI_Long_L2,clear
*Or We can use Cross_Sectional_Long_MI_Long_w2 with cswgt_final
*Extract the original KFS data (m=0)
mi extract 0
*Declare survey design for dataset (In this example we use the eight years panel
data)
svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
vce(linearized)  clear
label define gender 0 "Female" 1 "Male"
label values  PO_gender  gender
*svy: tabulate produces two-way tabulations with tests of independence for complex
survey data
svy: tab  c8_primary_loc PO_gender if year==2004     ,  se ci

Number of strata   =         6                  Number of obs      =      3140
Number of PSUs     =      3140                  Population size    = 73278.441
                                                 Design df          =      3134

------------------------------------------------------------
c8_primar |                     PO_gender
y_loc     |          Female              Male            Total
----------+-------------------------------------------------
 A reside |           .1583            .3461            .5043
          |          (.007)          (.0093)          (.0106)
          |    [.145,.1725]    [.3281,.3645]    [.4836,.5251]
          |
 A rented |            .122            .2742            .3962
          |         (.0069)          (.0092)          (.0105)
          |   [.1091,.1363]    [.2565,.2926]     [.3758,.417]
          |
 Space th |           .0102            .0377            .0479
          |         (.0022)          (.0041)          (.0046)
          |   [.0066,.0157]    [.0305,.0466]    [.0397,.0578]
          |
 A site w |           .0109            .0268            .0377
          |         (.0021)          (.0032)          (.0039)
          |   [.0074,.0159]    [.0212,.0339]     [.0308,.046]
          |
 Some oth |           .0017            .0121            .0138
          |       (8.8e-04)          (.0023)          (.0025)
          | [5.9e-04,.0047]    [.0083,.0176]    [.0097,.0196]
          |
    Total |            .303             .697                1
          |         (.0075)          (.0075)
          |    [.2885,.318]     [.682,.7115]
------------------------------------------------------------
  Key:  cell proportions
        (linearized standard errors of cell proportions)
        [95% confidence intervals for cell proportions]

  Pearson:
    Uncorrected   chi2(4)         =   13.7122
    Design-based  F(3.99, 12494.09)=    2.3516    P = 0.0520
```

Example 4.17: Correlations

```
*KFS format : Wide
*Example 4.17: Correlations of Variables
use Cross_Sectional_wide_MI_Long_w2,clear
*Extract the original KFS data (m=0)
mi extract 0

*Continuous variables
*See  :  http://www.stata.com/support/faqs/statistics/estimate-correlations-with-survey-data/


correlate Assets_7   Debt_7 Total_Employees_7  credrisk_7 [aweight=cswgt_final_7]

             | Assets_7   Debt_7 Total_~7 credri~7
-------------+------------------------------------
    Assets_7 |   1.0000
      Debt_7 |   0.2025   1.0000
Total_Empl~7 |   0.1361   0.2488   1.0000
  credrisk_7 |   0.0382  -0.0224   0.0400   1.0000


pwcorr         Assets_7       Debt_7  Total_Employees_7    credrisk_7  [aweight=
cswgt_final_7]


             | Assets_7   Debt_7 Total_~7 credri~7
-------------+------------------------------------
    Assets_7 |   1.0000
      Debt_7 |   0.2030   1.0000
Total_Empl~7 |   0.1345   0.6448   1.0000
  credrisk_7 |   0.0062  -0.0435  -0.0098   1.0000


/*listwise handles missing values through listwise deletion, meaning that the
entire observation is
omitted from the estimation sample if any of the variables in varlist is missing
for that observation.
By default, pwcorr handles missing values by pairwise deletion; all available
observations are used to
calculate each pairwise correlation without regard to whether variables outside
that pair are missing.*/


pwcorr         Assets_7       Debt_7  Total_Employees_7    credrisk_7  [aweight=
cswgt_final_7],listwise

             | Assets_7   Debt_7 Total_~7 credri~7
-------------+------------------------------------
    Assets_7 |   1.0000
      Debt_7 |   0.2025   1.0000
Total_Empl~7 |   0.1361   0.2488   1.0000
  credrisk_7 |   0.0382  -0.0224   0.0400   1.0000
```

```
* Tetrachoric correlations for binary variables
svyset  [pweight=cswgt_final_7] , strata(sampleinfo_samplestrata_7)
*Tetrachoric correlation coefficient ("rho")
svy:biprobit Home_Based_7 PO_gender_7

Survey: Bivariate probit regression

Number of strata   =         6                  Number of obs      =      2007
Number of PSUs     =      2007                  Population size    =  32681.32
                                                 Design df          =      2001
                                                 F(   0,   2001)    =         .
                                                 Prob > F           =         .

------------------------------------------------------------------------------
             |             Linearized
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Home_Based_7 |
       _cons |  -.0387146   .0330115    -1.17   0.241    -.1034551     .026026
-------------+----------------------------------------------------------------
PO_gender_7  |
       _cons |   .5330558   .0278404    19.15   0.000     .4784566     .587655
-------------+----------------------------------------------------------------
     /athrho |  -.1080754   .0458323    -2.36   0.018    -.1979594   -.0181915
-------------+----------------------------------------------------------------
         rho |  -.1076566   .0453011                     -.1954134   -.0181895
------------------------------------------------------------------------------

*KFS format : Long
use Cross_Sectional_Long_MI_Long_L2,clear
*Extract the original KFS data (m=0)
mi extract 0
*Continuous variables
*See  :  http://www.stata.com/support/faqs/statistics/estimate-correlations-with-survey-data/

correlate  Assets     Debt  Total_Employees    credrisk  [aweight=cswgt_final]  if
year==2011

             |   Assets     Debt Total_~s credrisk
-------------+------------------------------------
      Assets |   1.0000
        Debt |   0.2025   1.0000
Total_Empl~s |   0.1361   0.2488   1.0000
    credrisk |   0.0382  -0.0224   0.0400   1.0000
```

```
/*listwise handles missing values through listwise deletion, meaning that the
entire observation is
omitted from the estimation sample if any of the variables in varlist is missing
for that observation.
By default, pwcorr handles missing values by pairwise deletion; all available
observations are used to
calculate each pairwise correlation without regard to whether variables outside
that pair are missing.*/

pwcorr   Assets  Debt Total_Employees  credrisk [aweight=         cswgt_final]if
year==2011


             |   Assets     Debt Total_~s credrisk
-------------+------------------------------------
      Assets |   1.0000
        Debt |   0.2030   1.0000
Total_Empl~s |   0.1345   0.6448   1.0000
    credrisk |   0.0062  -0.0435  -0.0098   1.0000

pwcorr    Assets   Debt Total_Employees  credrisk [aweight=          cswgt_final]
if year==2011,listwise


             |   Assets     Debt Total_~s credrisk
-------------+------------------------------------
      Assets |   1.0000
        Debt |   0.2025   1.0000
Total_Empl~s |   0.1361   0.2488   1.0000
    credrisk |   0.0382  -0.0224   0.0400   1.0000


* Tetrachoric correlations for binary variables
svyset  [pweight=cswgt_final] , strata(sampleinfo_samplestrata)
*Tetrachoric correlation coefficient ("rho")
svy:biprobit Home_Based PO_gender  if year==2011

Number of strata   =         6                  Number of obs      =      2007
Number of PSUs     =      2007                  Population size    =  32681.32
                                                 Design df          =      2001
                                                 F(   0,   2001)    =         .
                                                 Prob > F           =         .

------------------------------------------------------------------------------
             |             Linearized
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Home_Based   |
       _cons |  -.0387146   .0330115    -1.17   0.241    -.1034551     .026026
-------------+----------------------------------------------------------------
PO_gender    |
       _cons |   .5330558   .0278404    19.15   0.000     .4784566     .587655
-------------+----------------------------------------------------------------
     /athrho |  -.1080754   .0458323    -2.36   0.018    -.1979594   -.0181915
-------------+----------------------------------------------------------------
         rho |  -.1076566   .0453011                     -.1954134   -.0181895
------------------------------------------------------------------------------
```

Example 4.18: Differences of Means for Two Subpopulations

Since complex sample design generates sampled observations that are not independent, thus any inference about differences of descriptive statistics for two subpopulations should take into account that the two estimates are correlated. Stata has many commands that account for this covariance when computing the variance of the difference. For example formulas for t test in complex sample design verses simple random sample (SRS) are:

Complex sample design	Simple random sample (SRS)
Subpopulations i and j	Subpopulations i and j
$t = \frac{\bar{x}_i - \bar{x}_j}{se(\bar{x}_i - \bar{x}_j)}$	$t = \frac{\bar{x}_i - \bar{x}_j}{se(\bar{x}_i - \bar{x}_j)}$
$se(\bar{x}_i - \bar{x}_j) = \sqrt{Var(\bar{x}_i) + Var(\bar{x}_j) - 2Cov(\bar{x}_i, \bar{x}_j)}$	$se(\bar{x}_i - \bar{x}_j) = \sqrt{Var(\bar{x}_i) + Var(\bar{x}_j)}$

```
*KFS format : Wide
*Example 4.18: Differences of Means for Two Subpopulations
use Longitudinal_wide_MI_Long_w2,clear
*Extract the original KFS data (m=0)
mi extract 0
*Declare survey design for dataset
svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_3)
*Differences of Means for Two Subpopulations (White vs. Black in 2007)
svy: mean    Debt_Bus_3 , over(PO_race_group_3)

Number of strata =       6          Number of obs    =    1892
Number of PSUs   =    1892          Population size  = 41996.7
                                    Design df        =    1886

            1: PO_race_group_3 = 1
            2: PO_race_group_3 = 2
            3: PO_race_group_3 = 3
            4: PO_race_group_3 = 4
            5: PO_race_group_3 = 5
            6: PO_race_group_3 = 6

--------------------------------------------------------------
             |             Linearized
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
Debt_Bus_3   |
           1 |   2886.553   1484.727     -25.32613    5798.432
           2 |   1542.883   636.8408      293.8968     2791.87
           3 |   17042.12   8570.926      232.6282    33851.61
           4 |   10049.69   3870.223      2459.324    17640.06
           5 |      61045   20153.56      21519.38    100570.6
           6 |   9490.999   4208.718      1236.767    17745.23
--------------------------------------------------------------
```

```
vce
```

e(V)	Debt_Bus_3 1	2	3	4	5	6
Debt_Bus_3						
1	2204412.9					
2	1240.2526	405566.25				
3	707.68351	1460.1242	73460768			
4	11022.785	-3972.3375	2304.3986	14978624		
5	46216.01	-1661.6005	-51449.429	-31860.012	4.062e+08	
6	3112.3847	276.38624	-9230.7897	-1503.5248	-46501.297	17713304

```
qui mean    Debt_Bus_3 , over(PO_race_group_3)
vce

Covariance matrix of coefficients of mean model
```

e(V)	Debt_Bus_3 1	2	3	4	5	6
Debt_Bus_3						
1	1512391.5					
2	0	1150000				
3	0	0	63134401			
4	0	0	0	6.914e+08		
5	0	0	0	0	2.153e+08	
6	0	0	0	0	0	11019538

```
*Using the lincom command
lincom [Debt_Bus_3]5 - [Debt_Bus_3]4

( 1)  - [Debt_Bus_3]4 + [Debt_Bus_3]5 = 0

------------------------------------------------------------------------------
        Mean |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   50995.31   20523.36     2.48   0.013     10744.43    91246.18
------------------------------------------------------------------------------


*Using the test command
test  [Debt_Bus_3]5 = [Debt_Bus_3]4

Adjusted Wald test

 ( 1)  - [Debt_Bus_3]4 + [Debt_Bus_3]5 = 0

       F(  1,  1886) =    6.17
            Prob > F =    0.0131
```

```
*Using the regress command
svy , subpop(if PO_race_group_3==4 | PO_race_group_3==5): reg Debt_Bus_3
i.PO_race_group_3


Number of strata   =         6                  Number of obs      =      2849
Number of PSUs     =      2849                  Population size    = 67102.547
                                                 Subpop. no. of obs =      1749
                                                 Subpop. size       = 38742.358
                                                 Design df          =      2843
                                                 F(   1,   2843)    =      6.17
                                                 Prob > F           =    0.0130
                                                 R-squared          =    0.0005

------------------------------------------------------------------------------
-
             |             Linearized
  Debt_Bus_3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
-
5.PO_race_gro~3 |   50995.31   20523.58     2.48   0.013      10752.7    91237.92
       _cons |   10049.69   3873.739     2.59   0.010     2454.069    17645.32
------------------------------------------------------------------------------

test 5.PO_race_group_3

Adjusted Wald test

 ( 1)  5.PO_race_group_3 = 0

       F(  1,  2843) =    6.17
            Prob > F =    0.0130

*The t-test simply a special case of the F-test where only two groups are being
compared. The result of either will be exactly the same in terms of the p-value
```

```
*KFS format : Long
use Longitudinal_Long_MI_Long_L2,clear
*Extract the original KFS data (m=0)
mi extract 0
*Declare survey design for dataset
svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
*Differences of Means for Two Subpopulations (White vs. Black in 2007)
svy: mean    Debt_Bus if year==2007, over(PO_race_group)

Number of strata =       6          Number of obs    =    1892
Number of PSUs   =    1892          Population size  = 41996.7
                                    Design df        =    1886

           1: PO_race_group = 1
           2: PO_race_group = 2
           3: PO_race_group = 3
           4: PO_race_group = 4
           5: PO_race_group = 5
           6: PO_race_group = 6

--------------------------------------------------------------
             |             Linearized
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
Debt_Bus     |
           1 |   2886.553   1484.727     -25.32613    5798.432
           2 |   1542.883   636.8408      293.8968     2791.87
           3 |   17042.12   8570.926      232.6282    33851.61
           4 |   10049.69   3870.223      2459.324    17640.06
           5 |      61045   20153.56      21519.38    100570.6
           6 |   9490.999   4208.718      1236.767    17745.23
--------------------------------------------------------------

*Using the lincom command
lincom [Debt_Bus]5 - [Debt_Bus]4

( 1)  - [Debt_Bus]4 + [Debt_Bus]5 = 0

------------------------------------------------------------------------------
        Mean |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   50995.31   20523.36     2.48   0.013     10744.43    91246.18
------------------------------------------------------------------------------

*Using the test command
test  [Debt_Bus]5 = [Debt_Bus]4

Adjusted Wald test

 ( 1)  - [Debt_Bus]4 + [Debt_Bus]5 = 0

       F(  1,  1886) =    6.17
            Prob > F =    0.0131
```

```
*Using the regress command
svy , subpop(if PO_race_group==4 | PO_race_group==5): reg Debt_Bus i.PO_race_group
if year==2007


Number of strata   =         6                  Number of obs      =      2039
Number of PSUs     =      2039                  Population size    = 45489.683
                                                 Subpop. no. of obs =      1749
                                                 Subpop. size       = 38742.358
                                                 Design df          =      2033
                                                 F(   1,   2033)    =      6.17
                                                 Prob > F           =    0.0130
                                                 R-squared          =    0.0005

------------------------------------------------------------------------------
             |             Linearized
     Debt_Bus |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
5.PO_race_g~p |   50995.31   20523.47     2.48   0.013     10746.08    91244.54
        _cons |   10049.69   3870.878     2.60   0.009     2458.391    17640.99
------------------------------------------------------------------------------

test 5.PO_race_group

Adjusted Wald test

 ( 1)  5.PO_race_group = 0

       F(  1,  2033) =    6.17
            Prob > F =    0.0130

*The t-test simply a special case of the F-test where only two groups are being
compared. The result of either will be exactly the same in terms of the p-value
```

Example 4.19: Differences of Means over Time

```
*KFS format : Wide
*Example 4.19: Differences of Means over Time
use Longitudinal_wide_MI_Long_w2,clear
*Extract the original KFS data (m=0)
mi extract 0
*Declare survey design for dataset
svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_0)
* Using svy:reg command
svy: reg Debt_Bus_0
estimates store s0


Survey: Linear regression

Number of strata   =         6                  Number of obs      =      3004
Number of PSUs     =      3004                  Population size    = 70178.998
                                                 Design df          =      2998
                                                 F(   0,   2998)    =         .
                                                 Prob > F           =         .
                                                 R-squared          =    0.0000

------------------------------------------------------------------------------
             |             Linearized
  Debt_Bus_0 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   38703.36   8778.405     4.41   0.000     21491.05    55915.66
------------------------------------------------------------------------------

svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_7)
svy: reg Debt_Bus_7
estimates store s7


Survey: Linear regression

Number of strata   =         6                  Number of obs      =      1415
Number of PSUs     =      1415                  Population size    = 30151.826
                                                 Design df          =      1409
                                                 F(   0,   1409)    =         .
                                                 Prob > F           =         .
                                                 R-squared          =    0.0000

------------------------------------------------------------------------------
             |             Linearized
  Debt_Bus_7 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   22727.15   4566.987     4.98   0.000     13768.32    31685.97
------------------------------------------------------------------------------
```

```
suest s0 s7, svy


Simultaneous survey results for s0, s7

Number of strata   =         6                  Number of obs      =      3067
Number of PSUs     =      3067                  Population size    = 71351.244
                                                 Design df          =      3061

------------------------------------------------------------------------------
             |             Linearized
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
s0           |
       _cons |   38703.36   8778.377     4.41   0.000     21491.25    55915.47
-------------+----------------------------------------------------------------
s7           |
       _cons |   22727.15    4570.79     4.97   0.000     13765.02    31689.28
------------------------------------------------------------------------------

*Using the lincom command
lincom  [s0]_cons  -  [s7]_cons

( 1)  [s0]_cons - [s7]_cons = 0

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   15976.21   9817.513     1.63   0.104    -3273.375    35225.79
------------------------------------------------------------------------------

*Using the test command
test  [s0]_cons  =  [s7]_cons

Adjusted Wald test

 ( 1)  [s0]_cons - [s7]_cons = 0

       F(  1,  3061) =    2.65
            Prob > F =    0.1038
```

```
*KFS format : Long
*Example 4.19: Differences of Means over Time
use Longitudinal_Long_MI_Long_L2,clear
*Extract the original KFS data (m=0)
mi extract 0
*Declare survey design for dataset
svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
*Differences of Means  (2004 vs. 2011)
svy: mean    Debt_Bus if year==2004 | year==2011 , over(year)

Number of strata =       6          Number of obs    =    4419
Number of PSUs   =    4419          Population size  =  100331
                                    Design df        =    4413

         2004: year = 2004
         2011: year = 2011

--------------------------------------------------------------
             |             Linearized
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
Debt_Bus     |
        2004 |   38703.36   8777.597      21494.86    55911.85
        2011 |   22727.15   4571.552      13764.61    31689.68
--------------------------------------------------------------

*Using the lincom command
lincom [Debt_Bus]2011 - [Debt_Bus]2004

( 1)  - [Debt_Bus]2004 + [Debt_Bus]2011 = 0

------------------------------------------------------------------------------
        Mean |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -15976.21   9896.068    -1.61   0.107    -35377.47    3425.049
------------------------------------------------------------------------------

*Using the test command
test  [Debt_Bus]2011 = [Debt_Bus]2004

Adjusted Wald test

 ( 1)  - [Debt_Bus]2004 + [Debt_Bus]2011 = 0

       F(  1,  4413) =    2.61
            Prob > F =    0.1065
```

```
*Using the regress command
svy: reg Debt_Bus i.year if year==2004 | year==2011

Number of strata   =         6                  Number of obs      =      4419
Number of PSUs     =      4419                  Population size    = 100330.82
                                                 Design df          =      4413
                                                 F(   1,   4413)    =      2.61
                                                 Prob > F           =    0.1065
                                                 R-squared          =    0.0005

------------------------------------------------------------------------------
             |             Linearized
    Debt_Bus |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |
       2011  |  -15976.21   9896.068    -1.61   0.107    -35377.47    3425.049
             |
       _cons |   38703.36   8777.597     4.41   0.000     21494.86    55911.85
------------------------------------------------------------------------------

*Using the test command
test  2011.year

Adjusted Wald test

 ( 1)  2011.year = 0

       F(  1,  4413) =    2.61
            Prob > F =    0.1065

*Comparing the Mean of Differences for Paired Data
*Businesses that they responded to the survey in both years
*This test is used when the samples are dependent; that is, when there is only one
sample that has been tested twice (repeated measures)
*KFS format : Wide
use Longitudinal_wide_MI_Long_w2,clear
*Extract the original KFS data (m=0)
mi extract 0
*Declare survey design for dataset
svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_0)
* Using svy:reg command
svy,subpop(if Debt_Bus_7<.): reg Debt_Bus_0
estimates store s0
svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_7)
svy, subpop(if Debt_Bus_0<.) : reg Debt_Bus_7
estimates store s7
suest s0 s7, svy
```

```
Number of strata   =         6                  Number of obs      =      3140
Number of PSUs     =      3140                  Population size    = 73278.441
                                                 Design df          =      3134

------------------------------------------------------------------------------
             |             Linearized
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
s0           |
       _cons |   31903.96    8380.19     3.81   0.000     15472.75    48335.18
-------------+----------------------------------------------------------------
s7           |
       _cons |   22202.83   4706.816     4.72   0.000     12974.08    31431.59
------------------------------------------------------------------------------
*Differences of Means  (2004 vs. 2011)
*Using the lincom command
lincom  [s0]_cons  -  [s7]_cons

( 1)  [s0]_cons - [s7]_cons = 0

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    9701.13   9406.269     1.03   0.302    -8741.941     28144.2
------------------------------------------------------------------------------

*Using the test command
test  [s0]_cons  =  [s7]_cons

Adjusted Wald test

 ( 1)  [s0]_cons - [s7]_cons = 0

       F(  1,  3134) =    1.06
            Prob > F =    0.3025

* Another way is to test Difference=0
gen diff=Debt_Bus_0-Debt_Bus_7
svy:mean diff
Number of strata =       6          Number of obs    =    1352
Number of PSUs   =    1352          Population size  = 28979.6
                                    Design df        =    1346

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        diff |    9701.13   9406.705    -8752.268    28154.53
--------------------------------------------------------------

*Using the lincom command
lincom  diff

( 1)  diff = 0

------------------------------------------------------------------------------
        Mean |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    9701.13   9406.705     1.03   0.303    -8752.268    28154.53
------------------------------------------------------------------------------
```

```
*Using the test command
test  diff=0

Adjusted Wald test

 ( 1)  diff = 0

       F(  1,  1346) =    1.06
            Prob > F =    0.3026

* Using seemingly unrelated regression via Structural Equation Modeling
svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_0)
svy:sem     (Debt_Bus_0     <-     _cons)     (Debt_Bus_7     <-     _cons)     ,
cov(e.Debt_Bus_0*e.Debt_Bus_7)  nocapslatent

Number of strata   =         6                  Number of obs      =      1352
Number of PSUs     =      1352                  Population size    = 28979.579
                                                 Design df          =      1346

------------------------------------------------------------------------------
                 |             Linearized
                 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
Structural       |
  Debt_Bus_0 <-  |
           _cons |   31903.96   8386.012     3.80   0.000     15452.89    48355.04
  ---------------+------------------------------------------------------------
--
  Debt_Bus_7 <-  |
           _cons |   22202.83   4702.994     4.72   0.000     12976.84    31428.83
-----------------+------------------------------------------------------------
Variance         |
    e.Debt_Bus_0 |   4.65e+10   2.07e+10                      1.94e+10    1.11e+11
    e.Debt_Bus_7 |   3.80e+10   1.69e+10                      1.59e+10    9.09e+10
-----------------+------------------------------------------------------------
Covariance       |
  e.Debt_Bus_0   |
    e.Debt_Bus_7 |   2.54e+09   1.50e+09     1.70   0.090    -3.97e+08    5.47e+09
------------------------------------------------------------------------------

*Using the lincom command
lincom [Debt_Bus_0]_cons - [Debt_Bus_7]_cons

( 1)  [Debt_Bus_0]_cons - [Debt_Bus_7]_cons = 0

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    9701.13   9406.705     1.03   0.303    -8752.268    28154.53
------------------------------------------------------------------------------
```

```
*KFS format : Long
*Example 4.19: Differences of Means over Time
use Longitudinal_Long_MI_Long_L2,clear
*Extract the original KFS data (m=0)
mi extract 0
*Declare survey design for dataset
svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
*Differences of Means  (2004 vs. 2011)
xtset mprid year
*Mark businesses that they responded to the survey in both years
bysort mprid (year): gen diff=Debt_Bus[1]-Debt_Bus[8]  if year==2004 | year==2011
svy:mean diff if year==2004

Number of strata =       6          Number of obs    =    1352
Number of PSUs   =    1352          Population size  = 28979.6
                                    Design df        =    1346

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        diff |    9701.13   9406.705     -8752.268    28154.53
--------------------------------------------------------------

*Using the lincom command
lincom  diff

( 1)  diff = 0

------------------------------------------------------------------------------
        Mean |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    9701.13   9406.705     1.03   0.303    -8752.268    28154.53
------------------------------------------------------------------------------
```

Example 4.20: Estimating Percentiles

```
*KFS format : Wide
*Example 4.20: Estimating Percentiles
use Longitudinal_wide_MI_Long_w2,clear
*Extract the original KFS data (m=0)
mi extract 0
gen All_Obs=1
*select up to five statistics
table All_Obs[pweight=wgt_7_long], contents(p25 Assets_0    p50 Assets_0 p75
Assets_0 )

----------------------------------------------------
  All_Obs | p25(Assets_0)  med(Assets_0)  p75(Assets_0)
----------+-----------------------------------------
        1 |          3080          19000          70500
----------------------------------------------------


*KFS format : Long
*Example 4.20: Estimating Percentiles
use Longitudinal_Long_MI_Long_L2,clear
*Extract the original KFS data (m=0)
mi extract 0
*select up to five statistics
table year[pweight=wgt_7_long] if year==2004, contents(p25 Assets        p50
Assets       p75 Assets)

-------------------------------------------------
     year | p25(Assets)  med(Assets)  p75(Assets)
----------+--------------------------------------
     2004 |        3080        19000        70500
-------------------------------------------------




table year[pweight=wgt_7_long]  , contents(p25 Assets       p50 Assets    p75
Assets)

-------------------------------------------------
     year | p25(Assets)  med(Assets)  p75(Assets)
----------+--------------------------------------
     2004 |        3080        19000        70500
     2005 |        7000        31800       114000
     2006 |        7717        37000       141000
     2007 |        7950        42000       150000
     2008 |        6050        37000       160000
     2009 |        6700        35500       150000
     2010 |        6000        40000       175000
     2011 |        7148        40000       195000
-------------------------------------------------
```

4.5.2. Descriptive: Using KFS Imputed Data

For multiply imputed data, the descriptive statistics commands that work with "mi estimate:" command and support survey commands are the following:

Survey commands	Descriptive statistics
mi estimate: svy: mean	Estimate means
mi estimate: svy: proportion	Estimate proportions
mi estimate: svy: ratio	Estimate ratios
mi estimate: svy: total	Estimate totals

Example 4.21: Estimating the Mean Value

```
*Using The KFS multiply imputed data
*KFS format : Wide
use Longitudinal_wide_MI_Long_w2,clear
*Declare multiply imputed data & survey design for dataset (In this example we use
the eight years panel data)
mi svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_0)
vce(linearized)  clear
*Use post option to post estimated coefficients and VCE to e(b) and e(V)
mi estimate, post :svy: mean  Assets_0 Total_Employees_0 Net_Profit_0  PO_gender_0
OO_gender_owner_0

Multiple-imputation estimates      Imputations     =         5
Survey: Mean estimation            Number of obs   =      3140

Number of strata  =         6      Population size = 73278.441
Number of PSUs    =      3140
                                   Average RVI     =    0.0109
                                   Largest FMI     =    0.0182
                                   Complete DF     =      3134
DF adjustment:   Small sample      DF:     min     =   2462.87
                                           avg     =   2939.90
Within VCE type:   Linearized              max     =   3132.00

------------------------------------------------------------------------------
                  |       Mean   Std. Err.     [95% Conf. Interval]
------------------+-----------------------------------------------------------
         Assets_0 |   191409.9   37843.09      117208.7    265611.1
Total_Employees_0 |   2.372919   .1132702      2.150824    2.595015
     Net_Profit_0 |  -4457.505   3409.301      -11142.9    2227.888
      PO_gender_0 |   .6969625   .0075423      .6821743    .7117508
OO_gender_owner_0 |   .6816901   .0068766       .668207    .6951731
------------------------------------------------------------------------------


*Comparing Imputed vs. Original
*Drop all stored estimation results.
estimates clear
estimates store Imputed
* Run estimate on m=0
mi  xeq  0:  svy:  mean   Assets_0  Total_Employees_0  Net_Profit_0   PO_gender_0
OO_gender_owner_0
estimates store  Original
estimates table Imputed Original,  b(%16.4f) se(%16.4f)  stats(N)
```

```
------------------------------------------------
    Variable |     Imputed           Original
-------------+----------------------------------
    Assets_0 |      191409.9214        204375.1688
             |       37843.0902         46232.2939
Total_Empl~0 |          2.3729             2.3765
             |          0.1133             0.1263
Net_Profit_0 |       -4457.5046         -4657.0307
             |        3409.3009          4056.5819
 PO_gender_0 |          0.6970             0.7002
             |          0.0075             0.0082
OO_gender_~0 |          0.6817             0.6849
             |          0.0069             0.0075
-------------+----------------------------------
           N |            3140               2560
------------------------------------------------
                                  legend: b/se



*Using The KFS multiply imputed data
*KFS format : Long
use Longitudinal_Long_MI_Long_L2,clear
*Or We can use Cross_Sectional_Long_MI_Long_w2 with cswgt_final
*Declare survey design for dataset (In this example we use the eight years panel
data)
mi svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata )
vce(linearized)  clear
*Declare data to be panel data
mi xtset mprid year
mi   estimate:svy:   mean     Assets   Total_Employees   Net_Profit     PO_gender
OO_gender_owner if year==2004

Multiple-imputation estimates      Imputations     =         5
Survey: Mean estimation            Number of obs   =      3140

Number of strata  =         6      Population size = 73278.441
Number of PSUs    =      3140
                                   Average RVI     =    0.0109
                                   Largest FMI     =    0.0182
                                   Complete DF     =      3134
DF adjustment:   Small sample      DF:     min     =   2462.87
                                           avg     =   2939.90
Within VCE type:   Linearized              max     =   3132.00

------------------------------------------------------------------------------
                |       Mean   Std. Err.     [95% Conf. Interval]
----------------+-------------------------------------------------------------
         Assets |   191409.9   37843.09      117208.7    265611.1
Total_Employees |   2.372919   .1132702      2.150824    2.595015
     Net_Profit |  -4457.505   3409.301      -11142.9    2227.888
      PO_gender |   .6969625   .0075423      .6821743    .7117508
OO_gender_owner |   .6816901   .0068766       .668207    .6951731
------------------------------------------------------------------------------
```

Example 4.22: Estimating the Mean Value of Subpopulation

```
*Using The KFS multiply imputed data
*KFS format : Wide
use Longitudinal_wide_MI_Long_w2,clear
 /*Or We can use Cross_Sectional_wide_MI_Long_w2 with cswgt_final_x as weight and
varlist_x*/
*Declare survey design for dataset (In this example we use the eight years panel
data)
mi svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_0)
vce(linearized)  clear

label define gender 0 "Female" 1 "Male"
label values  PO_gender_* gender
*subpop(subpop) specifies that estimates be computed for the single subpopulation
identified by subpop.
*subpopulation is defined by the observations for which varname!=0 that also meet
the if conditions.
*Typically,  varname=1  defines  the  subpopulation,  and  varname=0  indicates
observations not belonging to the subpopulation
mi  estimate:svy:  mean  Assets_0  Total_Employees_0  Net_Profit_0  ,subpop  (if
PO_gender_0==1  )

Multiple-imputation estimates      Imputations     =         5
Survey: Mean estimation            Number of obs   =      3140

Number of strata  =         6      Population size = 73278.441
Number of PSUs    =      3140      Subpop. no. obs =      2306
                                   Subpop. size    = 51072.328
                                   Average RVI     =    0.0158
                                   Largest FMI     =    0.0108
                                   Complete DF     =      3134
DF adjustment:   Small sample      DF:     min     =   2846.58
                                           avg     =   2956.69
Within VCE type:   Linearized              max     =   3109.50

------------------------------------------------------------------------------
                   |       Mean   Std. Err.     [95% Conf. Interval]
-------------------+----------------------------------------------------------
         Assets_0 |   238929.5   53917.81      133208.6    344650.3
Total_Employees_0 |   2.643505   .1480409      2.353237    2.933773
     Net_Profit_0 |  -5574.044   4821.869     -15028.75    3880.665
------------------------------------------------------------------------------
```

```
*Estimates for multiple subpopulations
mi estimate:svy: mean Assets_0 Total_Employees_0 Net_Profit_0, over(PO_gender_0)


Multiple-imputation estimates      Imputations     =         5
Survey: Mean estimation            Number of obs   =      3140

Number of strata  =         6      Population size = 73278.441
Number of PSUs    =      3140
                                   Average RVI     =    0.0396
                                   Largest FMI     =    0.1151
                                   Complete DF     =      3134
DF adjustment:   Small sample      DF:     min     =    296.84
                                           avg     =   2118.11
Within VCE type:   Linearized              max     =   3109.50

       Female: PO_gender_0 = Female
         Male: PO_gender_0 = Male

------------------------------------------------------------------------------
             Over |       Mean   Std. Err.     [95% Conf. Interval]
------------------+-----------------------------------------------------------
Assets_0          |
           Female |   82118.65   15182.78      52339.16    111898.1
             Male |   238929.5   53917.81      133208.6    344650.3
------------------+-----------------------------------------------------------
Total_Employees_0 |
           Female |   1.750592    .152563      1.451382    2.049802
             Male |   2.643505   .1480409      2.353237    2.933773
------------------+-----------------------------------------------------------
Net_Profit_0      |
           Female |  -1889.552   1710.491     -5255.777    1476.674
             Male |  -5574.044   4821.869     -15028.75    3880.665
------------------------------------------------------------------------------



* Using The KFS multiply imputed data
*KFS format : Long
use Longitudinal_Long_MI_Long_L2,clear
*Or We can use Cross_Sectional_Long_MI_Long_w2 with cswgt_final
*Extract the original KFS data (m=0)
*Declare survey design for dataset (In this example we use the eight years panel
data)
mi svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
vce(linearized)  clear
label define gender 0 "Female" 1 "Male"
label values  PO_gender  gender
*subpop(subpop) specifies that estimates be computed for the single subpopulation
identified by subpop.
*subpopulation is defined by the observations for which varname!=0 that also meet
the if conditions.
*Typically,  varname=1  defines  the  subpopulation,  and  varname=0  indicates
observations not belonging to the subpopulation
mi estimate:svy: mean Assets Total_Employees Net_Profit if year==2004,subpop (if
PO_gender==1  )
```

```
Multiple-imputation estimates      Imputations     =         5
Survey: Mean estimation            Number of obs   =      3140

Number of strata  =         6      Population size = 73278.441
Number of PSUs    =      3140      Subpop. no. obs =      2306
                                   Subpop. size    = 51072.328
                                   Average RVI     =    0.0158
                                   Largest FMI     =    0.0108
                                   Complete DF     =      3134
DF adjustment:   Small sample      DF:     min     =   2846.58
                                           avg     =   2956.69
Within VCE type:   Linearized              max     =   3109.50

------------------------------------------------------------------
                |       Mean   Std. Err.     [95% Conf. Interval]
----------------+-------------------------------------------------
         Assets |   238929.5   53917.81      133208.6    344650.3
Total_Employees |   2.643505   .1480409      2.353237    2.933773
     Net_Profit |  -5574.044   4821.869     -15028.75    3880.665
------------------------------------------------------------------

*Estimates for multiple subpopulations
mi   estimate:svy:   mean   Assets   Total_Employees   Net_Profit   if   year==2004,
over(PO_gender)

Multiple-imputation estimates      Imputations     =         5
Survey: Mean estimation            Number of obs   =      3140

Number of strata  =         6      Population size = 73278.441
Number of PSUs    =      3140
                                   Average RVI     =    0.0396
                                   Largest FMI     =    0.1151
                                   Complete DF     =      3134
DF adjustment:   Small sample      DF:     min     =    296.84
                                           avg     =   2118.11
Within VCE type:   Linearized              max     =   3109.50

       Female: PO_gender = Female
         Male: PO_gender = Male

------------------------------------------------------------------
           Over |       Mean   Std. Err.     [95% Conf. Interval]
----------------+-------------------------------------------------
Assets          |
         Female |   82118.65   15182.78      52339.16    111898.1
           Male |   238929.5   53917.81      133208.6    344650.3
----------------+-------------------------------------------------
Total_Employees |
         Female |   1.750592    .152563      1.451382    2.049802
           Male |   2.643505   .1480409      2.353237    2.933773
----------------+-------------------------------------------------
Net_Profit      |
         Female |  -1889.552   1710.491     -5255.777    1476.674
           Male |  -5574.044   4821.869     -15028.75    3880.665
------------------------------------------------------------------
```

Example 4.23: Estimating the Population Totals

```
*Using The KFS multiply imputed data
*KFS format : Wide
use Longitudinal_wide_MI_Long_w2,clear
/*Or We can use Cross_Sectional_wide_MI_Long_w2 with cswgt_final_x as weight and
varlist_x*/
*Declare survey design for dataset (In this example we use the eight years panel
data)
mi svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_7)
vce(linearized)  clear
mi     estimate:svy:     total          d3_a_num_patent_6     Total_Employees_6
c4_numowners_confirm_6    PO_gender_6


Multiple-imputation estimates      Imputations     =         5
Survey: Total estimation           Number of obs   =      1644

Number of strata  =         6      Population size = 35209.339
Number of PSUs    =      1644
                                   Average RVI     =    0.0000
                                   Largest FMI     =    0.0001
                                   Complete DF     =      1638
DF adjustment:   Small sample      DF:     min     =   1635.84
                                           avg     =   1635.96
Within VCE type:   Linearized              max     =   1636.00

------------------------------------------------------------------------------
                        |      Total   Std. Err.     [95% Conf. Interval]
------------------------+-----------------------------------------------------
     d3_a_num_patent_6 |   5140.176   2395.016      442.5552    9837.796
     Total_Employees_6 |   156513.8   11261.63        134425    178602.5
c4_numowners_confirm_6 |   48180.07   848.2294       46516.34     49843.8
           PO_gender_6 |   25216.38   419.8443        24392.9    26039.87
------------------------------------------------------------------------------
```

```
*Estimating the Population Totals of Subpopulation
mi    estimate:svy:    total             d3_a_num_patent_6    Total_Employees_6
c4_numowners_confirm_6    PO_gender_6  , over(Home_Based_6)

Number of strata  =         6      Population size = 35209.339
Number of PSUs    =      1644
                                 Average RVI     =    0.0004
                                 Largest FMI     =    0.0029
                                 Complete DF     =      1638
DF adjustment:   Small sample    DF:     min     =   1625.51
                                         avg     =   1634.68
Within VCE type:   Linearized            max     =   1636.00

    _subpop_1: Home_Based_6 = Non Home Based
    _subpop_2: Home_Based_6 = Home Based

------------------------------------------------------------------------------
                  Over |      Total   Std. Err.     [95% Conf. Interval]
-----------------------+------------------------------------------------------
d3_a_num_patent_6      |
             _subpop_1 |   2057.262   496.6237      1083.172    3031.352
             _subpop_2 |   3082.914   2343.733     -1514.119    7679.947
-----------------------+------------------------------------------------------
Total_Employees_6      |
             _subpop_1 |   128296.1   11136.07      106453.6    150138.5
             _subpop_2 |   28217.69   2846.972      22633.59    33801.78
-----------------------+------------------------------------------------------
c4_numowners_confirm_6 |
             _subpop_1 |   27413.01    1069.93      25314.43    29511.58
             _subpop_2 |   20767.06   697.4575      19399.06    22135.07
-----------------------+------------------------------------------------------
PO_gender_6            |
             _subpop_1 |   13659.85   507.8643      12663.71    14655.98
             _subpop_2 |   11556.54   441.4079      10690.75    12422.32
------------------------------------------------------------------------------

*Using The KFS multiply imputed data
*KFS format : Long
use Longitudinal_Long_MI_Long_L2,clear
*Or We can use Cross_Sectional_Long_MI_Long_w2 with cswgt_final
*Declare survey design for dataset (In this example we use the eight years panel
data)
mi svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
vce(linearized)  clear

mi estimate:svy: total  d3_a_num_patent  Total_Employees   c4_numowners_confirm
PO_gender if year==2010
```

```
Multiple-imputation estimates      Imputations     =         5
Survey: Total estimation           Number of obs   =      1775

Number of strata  =         6      Population size = 38271.277
Number of PSUs    =      1775
                                   Average RVI     =    0.0000
                                   Largest FMI     =    0.0001
                                   Complete DF     =      1769
DF adjustment:   Small sample      DF:     min     =   1766.89
                                           avg     =   1766.97
Within VCE type:   Linearized              max     =   1767.00

------------------------------------------------------------------------------
                     |      Total   Std. Err.     [95% Conf. Interval]
---------------------+--------------------------------------------------------
     d3_a_num_patent |   5207.797   2395.398      509.6845     9905.91
     Total_Employees |   167731.2    12031.5      144133.7    191328.7
c4_numowners_confirm |   52584.77   1006.111      50611.47    54558.06
           PO_gender |   27258.47    437.057      26401.27    28115.67
------------------------------------------------------------------------------

*Estimating the Population Totals of Subpopulation
mi estimate:svy: total   d3_a_num_patent  Total_Employees   c4_numowners_confirm
PO_gender  if year==2010 , over(Home_Based )


Multiple-imputation estimates      Imputations     =         5
Survey: Total estimation           Number of obs   =      1775

Number of strata  =         6      Population size = 38271.277
Number of PSUs    =      1775
                                   Average RVI     =    0.0003
                                   Largest FMI     =    0.0021
                                   Complete DF     =      1769
DF adjustment:   Small sample      DF:     min     =   1759.93
                                           avg     =   1766.10
Within VCE type:   Linearized              max     =   1767.00

    _subpop_1: Home_Based = Non Home Based
    _subpop_2: Home_Based = Home Based

------------------------------------------------------------------------------
                Over |      Total   Std. Err.     [95% Conf. Interval]
---------------------+--------------------------------------------------------
d3_a_num_patent      |
           _subpop_1 |   2124.884   498.0722      1148.008    3101.759
           _subpop_2 |   3082.914    2343.76     -1513.921    7679.748
---------------------+--------------------------------------------------------
Total_Employees      |
           _subpop_1 |   137268.9   11920.33      113889.4    160648.3
           _subpop_2 |   30462.34    2897.81      24778.84    36145.83
---------------------+--------------------------------------------------------
c4_numowners_confirm |
           _subpop_1 |   29790.22   1222.233      27393.04    32187.39
           _subpop_2 |   22794.55   725.7221      21371.18    24217.91
---------------------+--------------------------------------------------------
PO_gender            |
           _subpop_1 |   14358.82   524.0877      13330.93    15386.72
           _subpop_2 |   12899.65   465.6983      11986.27    13813.03
------------------------------------------------------------------------------
```

Example 4.24: Estimating the Proportions for Binary and Categorical Variables

```
*Using The KFS multiply imputed data
*KFS format : Wide
*Example 4.24: Estimating the Proportions for Binary and Categorical Variables
use Cross_Sectional_wide_MI_Long_w2,clear
/*Or We can use Longitudinal_wide_MI_Long_w2  */
*Declare survey design for dataset : In this example we study the baseline survey
mi svyset  [pweight=cswgt_final_0] , strata(sampleinfo_samplestrata_0)
vce(linearized)  clear

* Binary Variables
mi   estimate:svy:   proportion     PO_race_amind_owner_0   PO_race_asian_owner_0
PO_race_black_owner_0        PO_race_nathaw_owner_0        PO_race_other_owner_0
PO_race_white_owner_0

Multiple-imputation estimates      Imputations     =         5
Survey: Proportion estimation      Number of obs   =      4928

Number of strata  =         6      Population size = 73278.441
Number of PSUs    =      4928
                                   Average RVI     =    0.0086
                                   Largest FMI     =    0.0194
                                   Complete DF     =      4922
DF adjustment:   Small sample      DF:     min     =   3338.36
                                           avg     =   4406.27
Within VCE type:   Linearized              max     =   4920.00

------------------------------------------------------------------------------
                      | Proportion   Std. Err.     [95% Conf. Interval]
----------------------+-------------------------------------------------------
PO_race_amind_owner_0  |
                    0 |   .9880734   .0018487      .9844489     .991698
                    1 |   .0119266   .0018487       .008302    .0155511
----------------------+-------------------------------------------------------
PO_race_asian_owner_0  |
                    0 |   .9626239   .0031555      .9564377    .9688101
                    1 |   .0373761   .0031555      .0311899    .0435623
----------------------+-------------------------------------------------------
PO_race_black_owner_0  |
                    0 |   .9140265   .0047351      .9047435    .9233094
                    1 |   .0859735   .0047351      .0766906    .0952565
----------------------+-------------------------------------------------------
PO_race_nathaw_owner_0 |
                    0 |   .9943257   .0013552      .9916689    .9969824
                    1 |   .0056743   .0013552      .0030176    .0083311
----------------------+-------------------------------------------------------
PO_race_other_owner_0  |
                    0 |   .9483999   .0037835      .9409816    .9558181
                    1 |   .0516001   .0037835      .0441819    .0590184
----------------------+-------------------------------------------------------
PO_race_white_owner_0  |
                    0 |   .1925507   .0066411      .1795309    .2055705
                    1 |   .8074493   .0066411      .7944295    .8204691
------------------------------------------------------------------------------
```

```
* Using The KFS multiply imputed data
mi estimate:svy: proportion  PO_race_group_0

Multiple-imputation estimates      Imputations     =         5
Survey: Proportion estimation      Number of obs   =      4928

Number of strata  =         6      Population size = 73278.441
Number of PSUs    =      4928
                                   Average RVI     =    0.0086
                                   Largest FMI     =    0.0194
                                   Complete DF     =      4922
DF adjustment:   Small sample      DF:     min     =   3338.36
                                           avg     =   4406.27
Within VCE type:   Linearized              max     =   4920.00

------------------------------------------------------------------
                | Proportion   Std. Err.     [95% Conf. Interval]
----------------+-------------------------------------------------
              1 |   .0119266   .0018487       .008302    .0155511
              2 |   .0056743   .0013552      .0030176    .0083311
              3 |   .0373761   .0031555      .0311899    .0435623
              4 |   .0859735   .0047351      .0766906    .0952565
              5 |   .8074493   .0066411      .7944295    .8204691
              6 |   .0516001   .0037835      .0441819    .0590184
------------------------------------------------------------------

*Using The KFS multiply imputed data
*KFS format : Long
use Cross_Sectional_Long_MI_Long_L2,clear
*Or We can use Cross_Sectional_Long_MI_Long_w2 with cswgt_final
*Declare survey design for dataset
mi svyset  [pweight=cswgt_final] , strata(sampleinfo_samplestrata)
vce(linearized)  clear

* Binary Variables
mi   estimate:svy:   proportion       PO_race_amind_owner   PO_race_asian_owner
PO_race_black_owner PO_race_nathaw_owner PO_race_other_owner PO_race_white_owner
if year==2004
```

```
Multiple-imputation estimates      Imputations     =         5
Survey: Proportion estimation      Number of obs   =      4928

Number of strata  =         6      Population size = 73278.441
Number of PSUs    =      4928
                                   Average RVI     =    0.0000
                                   Largest FMI     =    0.0000
                                   Complete DF     =      4922
DF adjustment:   Small sample      DF:     min     =   4920.00
                                           avg     =   4920.00
Within VCE type:   Linearized              max     =   4920.00

------------------------------------------------------------------------------
                     | Proportion   Std. Err.     [95% Conf. Interval]
---------------------+--------------------------------------------------------
PO_race_amind_owner  |
                   0 |   .9882108    .001824      .9846349    .9917867
                   1 |   .0117892    .001824      .0082133    .0153651
---------------------+--------------------------------------------------------
PO_race_asian_owner  |
                   0 |   .9626913   .0031428        .95653    .9688526
                   1 |   .0373087   .0031428      .0311474      .04347
---------------------+--------------------------------------------------------
PO_race_black_owner  |
                   0 |   .9142331   .0047171      .9049854    .9234808
                   1 |   .0857669   .0047171      .0765192    .0950146
---------------------+--------------------------------------------------------
PO_race_nathaw_owner |
                   0 |   .9943257   .0013552      .9916689    .9969824
                   1 |   .0056743   .0013552      .0030176    .0083311
---------------------+--------------------------------------------------------
PO_race_other_owner  |
                   0 |   .9486778    .003736      .9413536     .956002
                   1 |   .0513222    .003736       .043998    .0586464
---------------------+--------------------------------------------------------
PO_race_white_owner  |
                   0 |   .1918614   .0066107      .1789015    .2048213
                   1 |   .8081386   .0066107      .7951787    .8210985
------------------------------------------------------------------------------
```

```
*Categorical Variables
mi estimate:svy: proportion  PO_race_group if year==2004

Multiple-imputation estimates      Imputations     =         5
Survey: Proportion estimation      Number of obs   =      4928

Number of strata  =         6      Population size = 73278.441
Number of PSUs    =      4928
                                   Average RVI     =    0.0000
                                   Largest FMI     =    0.0000
                                   Complete DF     =      4922
DF adjustment:   Small sample      DF:     min     =   4920.00
                                           avg     =   4920.00
Within VCE type:   Linearized              max     =   4920.00

------------------------------------------------------------------
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
           1 |   .0117892    .001824      .0082133    .0153651
           2 |   .0056743   .0013552      .0030176    .0083311
           3 |   .0373087   .0031428      .0311474      .04347
           4 |   .0857669   .0047171      .0765192    .0950146
           5 |   .8081386   .0066107      .7951787    .8210985
           6 |   .0513222    .003736       .043998    .0586464
------------------------------------------------------------------
```

Example 4.25: Estimating Ratios

```
*Using The KFS multiply imputed data
*KFS format : Wide
use Longitudinal_wide_MI_Long_w2,clear
/*Or We can use Cross_Sectional_wide_MI_Long_w2 with cswgt_final_x as weight and
varlist_x*/
*Declare survey design for dataset (In this example we use the eight years panel
data)
mi svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_0)
vce(linearized)  clear
mi              estimate:              svy:ratio              (Full_Time_Ratio_3:
Full_Time_Employees_3/Full_Part_Time_Employees_3)              (Part_Time_Ratio_3:
Part_Time_Employees_3/Full_Part_Time_Employees_3)

Multiple-imputation estimates      Imputations     =         5
Survey: Ratio estimation           Number of obs   =      2330

Number of strata  =         6      Population size = 51665.577
Number of PSUs    =      2330
                                   Average RVI     =    0.3850
                                   Largest FMI     =    0.0042
                                   Complete DF     =      2324
DF adjustment:   Small sample      DF:     min     =   2289.19
                                           avg     =   2305.20
Within VCE type:   Linearized              max     =   2321.22

 Full_Time_~3: Full_Time_Employees_3/Full_Part_Time_Employ~3
 Part_Time_~3: Part_Time_Employees_3/Full_Part_Time_Employ~3

------------------------------------------------------------------------------
                  |      Ratio   Std. Err.     [95% Conf. Interval]
------------------+-----------------------------------------------------------
Full_Time_Ratio_3 |   .6721141   .0231846      .6266492    .7175791
Part_Time_Ratio_3 |   .3312539   .0231923      .2857741    .3767338
------------------------------------------------------------------------------
```

```
*KFS format : Long
use Longitudinal_Long_MI_Long_L2,clear
*Or We can use Cross_Sectional_Long_MI_Long_w2 with cswgt_final
*Declare survey design for dataset (In this example we use the eight years panel
data)
mi svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
vce(linearized)  clear

mi                      estimate:svy:ratio                      (Full_Time_Ratio:
Full_Time_Employees/Full_Part_Time_Employees)                   (Part_Time_Ratio:
Part_Time_Employees/Full_Part_Time_Employees) if year==2007

Multiple-imputation estimates      Imputations     =         5
Survey: Ratio estimation           Number of obs   =      2330

Number of strata  =         6      Population size = 51665.577
Number of PSUs    =      2330
                                   Average RVI     =    0.3850
                                   Largest FMI     =    0.0042
                                   Complete DF     =      2324
DF adjustment:   Small sample      DF:     min     =   2289.19
                                           avg     =   2305.20
Within VCE type:   Linearized              max     =   2321.22

 Full_Time_~o: Full_Time_Employees/Full_Part_Time_Employ~s
 Part_Time_~o: Part_Time_Employees/Full_Part_Time_Employ~s

------------------------------------------------------------------
                |      Ratio   Std. Err.     [95% Conf. Interval]
----------------+-------------------------------------------------
Full_Time_Ratio |   .6721141   .0231846      .6266492    .7175791
Part_Time_Ratio |   .3312539   .0231923      .2857741    .3767338
------------------------------------------------------------------
```

Example 4.26: One-Way Tables for Survey Data

```
*Using The KFS multiply imputed data
*KFS format : Wide
/*mi estimate: svy:tabulate command not supported by mi estimate:
You can use option cmdok to allow estimation anyway. *The cmdok option tell mi
estimate to apply Rubin's rules to a non-mi command. However, it is your
responsibility to ensure that the results will be valid. */

use Longitudinal_wide_MI_Long_w2,clear
/*Or We can use Cross_Sectional_wide_MI_Long_w2 with cswgt_final_x as weight and
varlist_x*/
*Declare survey design for dataset (In this example we use the eight years panel
data)
mi svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_0)
vce(linearized)  clear

*estimation sample varies between m
*Allow estimation when estimation sample varies across imputations due to skip
logic
mi estimate,cmdok  esampvaryok :svy: tab  d9a_perc_internet_sales_3    ,ci    obs
percent  format(%16.4f)


Multiple-imputation estimates                   Imputations        =          5
                                                Number of obs      =        522
Number of strata  =         6
Number of PSUs    =       522                   Population size    = 12184.159

                                                Average RVI        =     0.0061
                                                Largest FMI        =     0.0110
                                                Complete DF        =        516
DF adjustment:   Small sample                   DF:     min        =     500.90
                                                        avg        =     507.52
Within VCE type:   Linearized                           max        =     513.57

------------------------------------------------------------------------------
        Mean |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         p11 |   .3239847   .0244099    13.27   0.000     .2760275     .371942
         p21 |   .3093666   .0240365    12.87   0.000     .2621419    .3565914
         p31 |   .1022253   .0160251     6.38   0.000     .0707416     .133709
         p41 |   .0927528   .0155104     5.98   0.000     .0622807     .123225
         p51 |   .1716705   .0187555     9.15   0.000     .1348236    .2085175
------------------------------------------------------------------------------
Warning: estimation sample varies across imputations; results may be biased.
         Sample sizes vary between 522 and 525.
Note: number of primary clusters varies among imputations.
Note: population size varies among imputations.
```

```
*Verify that tab apply Rubin's rules to a non-mi command correctly
mi estimate, esampvaryok :svy: proportion  d9a_perc_internet_sales_3

Multiple-imputation estimates      Imputations     =         5
Survey: Proportion estimation      Number of obs   =       522

Number of strata  =         6      Population size = 12184.159
Number of PSUs    =       522
                                   Average RVI     =    0.0061
                                   Largest FMI     =    0.0110
                                   Complete DF     =       516
DF adjustment:   Small sample      DF:     min     =    500.90
                                           avg     =    507.52
Within VCE type:   Linearized              max     =    513.57

      _prop_1: d9a_perc_internet_sales_3 = Less than 5%
      _prop_2: d9a_perc_internet_sales_3 = 5% - 25%
      _prop_3: d9a_perc_internet_sales_3 = 26% - 50%
      _prop_4: d9a_perc_internet_sales_3 = 51% - 75%
      _prop_5: d9a_perc_internet_sales_3 = 76% - 100%

------------------------------------------------------------------------------
                           | Proportion   Std. Err.     [95% Conf. Interval]
---------------------------+--------------------------------------------------
                  _prop_1 |   .3239847   .0244099      .2760275     .371942
                  _prop_2 |   .3093666   .0240365      .2621419    .3565914
                  _prop_3 |   .1022253   .0160251      .0707416     .133709
                  _prop_4 |   .0927528   .0155104      .0622807     .123225
                  _prop_5 |   .1716705   .0187555      .1348236    .2085175
------------------------------------------------------------------------------
Warning: estimation sample varies across imputations; results may be biased.
         Sample sizes vary between 522 and 525.
Note: numbers of observations in e(_N) vary among imputations.
Note: number of primary clusters varies among imputations.
Note: population size varies among imputations.
```

```
*NOTE
*subpopulation analysis is more appropriate here , d9a_perc_internet_sales has a skip logic
*subpop(subpop) specifies that estimates be computed for the single subpopulation identified by subpop.
*subpopulation is defined by the observations for which varname!=0 that also meet the if conditions.
*Typically, varname=1 defines the subpopulation, and varname=0 indicates observations not belonging to the subpopulation
mi estimate, esampvaryok :svy: proportion  d9a_perc_internet_sales_3   , subpop(if d9_internet_sales_3==1)


Multiple-imputation estimates      Imputations     =         5
Survey: Proportion estimation      Number of obs   =      3140

Number of strata  =         6      Population size = 73278.441
Number of PSUs    =      3140      Subpop. no. obs =       524
                                   Subpop. size    =  12230.85
                                   Average RVI     =    0.0049
                                   Largest FMI     =    0.0109
                                   Complete DF     =      3134
DF adjustment:   Small sample      DF:     min     =   2839.21
                                           avg     =   2997.40
Within VCE type:   Linearized              max     =   3128.08

      _prop_1: d9a_perc_internet_sales_3 = Less than 5%
      _prop_2: d9a_perc_internet_sales_3 = 5% - 25%
      _prop_3: d9a_perc_internet_sales_3 = 26% - 50%
      _prop_4: d9a_perc_internet_sales_3 = 51% - 75%
      _prop_5: d9a_perc_internet_sales_3 = 76% - 100%

------------------------------------------------------------------------------
                             | Proportion   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
                   _prop_1 |   .3239847   .0243638       .276213    .3717564
                   _prop_2 |   .3093666   .0240628      .2621844    .3565489
                   _prop_3 |   .1022253   .0160047       .070844    .1336065
                   _prop_4 |   .0927528   .0156129        .06214    .1233657
                   _prop_5 |   .1716705   .0187723      .1348633    .2084778
------------------------------------------------------------------------------
Note: numbers of observations in e(_N) vary among imputations.
Note: subpopulation size varies among imputations.
Note: number of observations in a subpopulation varies among imputations.
```

```
*KFS format : Long
use Longitudinal_Long_MI_Long_L2,clear
*Or We can use Cross_Sectional_Long_MI_Long_w2 with cswgt_final
*Declare survey design for dataset (In this example we use the eight years panel
data)
mi svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
vce(linearized)  clear
/*mi estimate: svy:tabulate command not supported by mi estimate:
You can use option cmdok to allow estimation anyway. *The cmdok option tell mi
estimate to apply Rubin's rules to a non-mi command. However, it is your
responsibility to ensure that the results will be valid. */
mi estimate,cmdok  esampvaryok :svy: tab  d9a_perc_internet_sales if year==2007
,ci    obs   percent  format(%16.4f)

Multiple-imputation estimates                   Imputations        =         5
                                                Number of obs      =       522
Number of strata  =         6
Number of PSUs    =       522                   Population size    = 12184.159

                                                Average RVI        =    0.0061
                                                Largest FMI        =    0.0110
                                                Complete DF        =       516
DF adjustment:   Small sample                   DF:     min        =    500.90
                                                        avg        =    507.52
Within VCE type:   Linearized                           max        =    513.57

------------------------------------------------------------------------------
        Mean |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         p11 |   .3239847   .0244099    13.27   0.000     .2760275     .371942
         p21 |   .3093666   .0240365    12.87   0.000     .2621419    .3565914
         p31 |   .1022253   .0160251     6.38   0.000     .0707416     .133709
         p41 |   .0927528   .0155104     5.98   0.000     .0622807     .123225
         p51 |   .1716705   .0187555     9.15   0.000     .1348236    .2085175
------------------------------------------------------------------------------
Warning: estimation sample varies across imputations; results may be biased.
         Sample sizes vary between 522 and 525.
Note: number of primary clusters varies among imputations.
Note: population size varies among imputations.
```

```
*Verify that tab apply Rubin's rules to a non-mi command correctly
mi estimate, esampvaryok :svy: proportion  d9a_perc_internet_sales if year==2007

Multiple-imputation estimates      Imputations     =         5
Survey: Proportion estimation      Number of obs   =       522

Number of strata  =         6      Population size = 12184.159
Number of PSUs    =       522
                                   Average RVI     =    0.0061
                                   Largest FMI     =    0.0110
                                   Complete DF     =       516
DF adjustment:   Small sample      DF:     min     =    500.90
                                           avg     =    507.52
Within VCE type:   Linearized              max     =    513.57

      _prop_1: d9a_perc_internet_sales = Less than 5%
      _prop_2: d9a_perc_internet_sales = 5% - 25%
      _prop_3: d9a_perc_internet_sales = 26% - 50%
      _prop_4: d9a_perc_internet_sales = 51% - 75%
      _prop_5: d9a_perc_internet_sales = 76% - 100%

------------------------------------------------------------------------------
                        | Proportion   Std. Err.     [95% Conf. Interval]
------------------------+-----------------------------------------------------
                _prop_1 |   .3239847   .0244099      .2760275     .371942
                _prop_2 |   .3093666   .0240365      .2621419    .3565914
                _prop_3 |   .1022253   .0160251      .0707416     .133709
                _prop_4 |   .0927528   .0155104      .0622807     .123225
                _prop_5 |   .1716705   .0187555      .1348236    .2085175
------------------------------------------------------------------------------
Warning: estimation sample varies across imputations; results may be
         biased.  Sample sizes vary between 522 and 525.
Note: numbers of observations in e(_N) vary among imputations.
Note: number of primary clusters varies among imputations.
Note: population size varies among imputations.
```

```
*subpopulation analysis is more appropriate here , d9a_perc_internet_sales has a
skip logic
*subpop(subpop) specifies that estimates be computed for the single subpopulation
identified by subpop.
*subpopulation is defined by the observations for which varname!=0 that also meet
the if conditions.
*Typically, varname=1 defines the subpopulation, and varname=0 indicates
observations not belonging to the subpopulation
mi estimate :svy: proportion  d9a_perc_internet_sales if year==2007   , subpop(if
d9_internet_sales==1)

Multiple-imputation estimates      Imputations     =         5
Survey: Proportion estimation      Number of obs   =      2330

Number of strata  =         6      Population size = 51665.577
Number of PSUs    =      2330      Subpop. no. obs =       524
                                   Subpop. size    =  12230.85
                                   Average RVI     =    0.0061
                                   Largest FMI     =    0.0109
                                   Complete DF     =      2324
DF adjustment:   Small sample      DF:     min     =   2151.37
                                           avg     =   2243.27
Within VCE type:   Linearized              max     =   2319.38

      _prop_1: d9a_perc_internet_sales = Less than 5%
      _prop_2: d9a_perc_internet_sales = 5% - 25%
      _prop_3: d9a_perc_internet_sales = 26% - 50%
      _prop_4: d9a_perc_internet_sales = 51% - 75%
      _prop_5: d9a_perc_internet_sales = 76% - 100%

------------------------------------------------------------------------------
                        | Proportion   Std. Err.     [95% Conf. Interval]
------------------------+-----------------------------------------------------
                _prop_1 |   .3239847   .0243677       .276199    .3717705
                _prop_2 |   .3093666   .0240609      .2621817    .3565516
                _prop_3 |   .1022253   .0160065      .0708362    .1336144
                _prop_4 |   .0927528   .0156043      .0621528    .1233529
                _prop_5 |   .1716705   .0187708      .1348613    .2084798
------------------------------------------------------------------------------
Note: numbers of observations in e(_N) vary among imputations.
Note: subpopulation size varies among imputations.
Note: number of observations in a subpopulation varies among imputations.
```

Example 4.27: Two-Way Tables for Survey Data

```
*KFS format : Wide
use Longitudinal_wide_MI_Long_w2,clear
/*Or We can use Cross_Sectional_wide_MI_Long_w2 with cswgt_final_x as weight and
varlist_x*/
*Declare survey design for dataset
mi svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_0)
vce(linearized)  clear
label define gender_0 0 "Female" 1 "Male"
label values  PO_gender_0  gender
*two-way tabulations is not officially supported by mi estimate
*The cmdok option tell mi estimate to apply Rubin's rules to a non-mi command.
However, it is your responsibility to ensure that the results will be valid.
mi estimate, cmdok  :svy:tab  c8_primary_loc_0 PO_gender_0 ,  se ci


Multiple-imputation estimates                   Imputations        =         5
                                                Number of obs      =      3140
Number of strata  =         6
Number of PSUs    =      3140                   Population size    = 73278.441

                                                Average RVI        =    0.0000
                                                Largest FMI        =    0.0000
                                                Complete DF        =      3134
DF adjustment:   Small sample                   DF:     min        =   3132.00
                                                        avg        =   3132.00
Within VCE type:   Linearized                           max        =   3132.00

------------------------------------------------------------------------------
        Mean |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         p11 |   .1582581   .0070007    22.61   0.000     .1445317    .1719845
         p12 |   .3460793   .0093016    37.21   0.000     .3278414    .3643173
         p21 |   .1220259   .0069395    17.58   0.000     .1084194    .1356323
         p22 |   .2741976   .0092211    29.74   0.000     .2561175    .2922777
         p31 |    .010198   .0022418     4.55   0.000     .0058025    .0145934
         p32 |   .0377286   .0040645     9.28   0.000     .0297592     .045698
         p41 |   .0108836   .0021244     5.12   0.000     .0067183    .0150489
         p42 |   .0268359   .0032287     8.31   0.000     .0205053    .0331665
         p51 |   .0016719   .0008849     1.89   0.059    -.0000631    .0034069
         p52 |   .0121211   .0023237     5.22   0.000     .0075651    .0166772
------------------------------------------------------------------------------
```

```
*KFS format : Long
use Longitudinal_Long_MI_Long_L2,clear
*Or We can use Cross_Sectional_Long_MI_Long_w2 with cswgt_final
*Declare survey design for dataset (In this example we use the eight years panel
data)
mi svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
vce(linearized)  clear
label define gender 0 "Female" 1 "Male"
label values  PO_gender  gender
*svy: tabulate produces two-way tabulations with tests of independence for complex
survey data
mi estimate, cmdok  :svy: tab  c8_primary_loc PO_gender if year==2004     ,  se
ci



Multiple-imputation estimates                   Imputations        =         5
                                                Number of obs      =      3140
Number of strata  =         6
Number of PSUs    =      3140                   Population size    = 73278.441

                                                Average RVI        =    0.0000
                                                Largest FMI        =    0.0000
                                                Complete DF        =      3134
DF adjustment:   Small sample                   DF:     min        =   3132.00
                                                        avg        =   3132.00
Within VCE type:   Linearized                           max        =   3132.00

------------------------------------------------------------------------------
        Mean |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         p11 |   .1582581   .0070007    22.61   0.000     .1445317    .1719845
         p12 |   .3460793   .0093016    37.21   0.000     .3278414    .3643173
         p21 |   .1220259   .0069395    17.58   0.000     .1084194    .1356323
         p22 |   .2741976   .0092211    29.74   0.000     .2561175    .2922777
         p31 |    .010198   .0022418     4.55   0.000     .0058025    .0145934
         p32 |   .0377286   .0040645     9.28   0.000     .0297592     .045698
         p41 |   .0108836   .0021244     5.12   0.000     .0067183    .0150489
         p42 |   .0268359   .0032287     8.31   0.000     .0205053    .0331665
         p51 |   .0016719   .0008849     1.89   0.059    -.0000631    .0034069
         p52 |   .0121211   .0023237     5.22   0.000     .0075651    .0166772
------------------------------------------------------------------------------
```

Example 4.28: Correlations

```
*KFS format : Wide
* Using The KFS multiply imputed data
use Cross_Sectional_wide_MI_Long_w2,clear
mi svyset  [pweight=cswgt_final_7] , strata(sampleinfo_samplestrata_7)
*Continuous variables
*See  :  http://www.stata.com/support/faqs/statistics/estimate-correlations-with-survey-data/
*correlate is not officially supported by mi estimate
*by averaging the Correlations over imputed data (by applying Rubin's rules to the estimates in the original metric)

gen Correlation = .
forvalues m = 1/5 {
local mid = "_mi_m"
quietly: correlate Assets_7   Debt_7   [aweight=cswgt_final_7] if(`mid'==`m')
replace Correlation  = r(rho) in `m'
}

cap quietly:summarize  Correlation
di "Correlation  = " as result r(mean)
cap drop Correlation

Correlation  = .10773858

* Tetrachoric correlations for binary variables
*biprobit is not officially supported by mi estimate
mi estimate,cmdok:svy:biprobit Home_Based_7 PO_gender_7

Number of strata  =         6                  Population size    =  32681.32
Number of PSUs    =      2007
                                                 Average RVI        =    0.0000
                                                 Largest FMI        =    0.0000
                                                 Complete DF        =      2001
                                                 DF:     min        =   1999.00
                                                         avg        =   1999.00
DF adjustment:   Small sample                            max        =   1999.00
                                                 F(   0,      .)    =         .
Within VCE type:   Linearized                    Prob > F           =         .

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Home_Based_7 |
       _cons |  -.0387146   .0330115    -1.17   0.241    -.1034551     .026026
-------------+----------------------------------------------------------------
PO_gender_7  |
       _cons |   .5330558   .0278404    19.15   0.000     .4784566     .587655
-------------+----------------------------------------------------------------
     /athrho |  -.1080754   .0458323    -2.36   0.018    -.1979594   -.0181914
-------------+----------------------------------------------------------------
         rho |  -.1076566   .0453011                     -.1954134   -.0181894
------------------------------------------------------------------------------
```

```
*KFS format : Long

use Cross_Sectional_Long_MI_Long_L2,clear
mi svyset  [pweight=cswgt_final ] , strata(sampleinfo_samplestrata )
*Continuous variables
*See  :  http://www.stata.com/support/faqs/statistics/estimate-correlations-with-survey-data/
 gen Correlation = .
forvalues m = 1/5 {
local mid = "_mi_m"
cap correlate Assets    Debt    [aweight=cswgt_final] if(`mid'==`m') & year==2011
replace Correlation  = r(rho) in `m'
}

cap quietly:summarize  Correlation
di "Correlation  = " as result r(mean)
cap drop Correlation

Correlation  = .10773858

* Tetrachoric correlations for binary variables
*biprobit is not officially supported by mi estimate
mi estimate,cmdok:svy:biprobit Home_Based  PO_gender if year==2011

Multiple-imputation estimates                   Imputations        =         5
Survey: Bivariate probit regression             Number of obs      =      2007

Number of strata  =         6                   Population size    =  32681.32
Number of PSUs    =      2007
                                                Average RVI        =    0.0000
                                                Largest FMI        =    0.0000
                                                Complete DF        =      2001
                                                DF:     min        =   1999.00
                                                        avg        =   1999.00
DF adjustment:   Small sample                           max        =   1999.00
                                                F(   0,      .)    =         .
Within VCE type:   Linearized                   Prob > F           =         .

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Home_Based   |
       _cons |  -.0387146   .0330115    -1.17   0.241    -.1034551     .026026
-------------+----------------------------------------------------------------
PO_gender    |
       _cons |   .5330558   .0278404    19.15   0.000     .4784566     .587655
-------------+----------------------------------------------------------------
     /athrho |  -.1080754   .0458323    -2.36   0.018    -.1979594   -.0181914
-------------+----------------------------------------------------------------
         rho |  -.1076566   .0453011                     -.1954134   -.0181894
------------------------------------------------------------------------------
```

Example 4.29: Differences of Means for Two Subpopulations

```
*KFS format : Wide
* Using The KFS multiply imputed data
use Longitudinal_wide_MI_Long_w2,clear
mi svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_3)
*Differences of Means for Two Subpopulations (Home vs. Non Home based in 2007)
cap mi estimate:svy: mean    Debt_Bus_3 , over(Home_Based_3)
*Using the mi testtransform command
mi  estimate  (diff:  [Debt_Bus_3]_subpop_1-[Debt_Bus_3]_subpop_2),  saving(miest,
replace):svy: mean    Debt_Bus_3 , over(Home_Based_3)

Number of strata  =         6      Population size = 51665.577
Number of PSUs    =      2330
                                   Average RVI     =    0.0018
                                   Largest FMI     =    0.0036
                                   Complete DF     =      2324
DF adjustment:   Small sample      DF:     min     =   2296.25
                                           avg     =   2309.12
Within VCE type:   Linearized              max     =   2321.99

    _subpop_1: Home_Based_3 = Non Home Based
    _subpop_2: Home_Based_3 = Home Based

--------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
   _subpop_1 |   59937.62    18216.1      24215.89    95659.35
   _subpop_2 |   31842.08   21144.92     -9622.807    73306.97
--------------------------------------------------------------

Transformations                    Average RVI     =    0.0016
                                   Largest FMI     =    0.0016
                                   Complete DF     =      2324
DF adjustment:   Small sample      DF:     min     =   2315.12
                                           avg     =   2315.12
Within VCE type:   Linearized              max     =   2315.12

         diff: [Debt_Bus_3]_subpop_1-[Debt_Bus_3]_subpop_2

--------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        diff |   28095.54   27914.87     -26645.21    82836.29
--------------------------------------------------------------

mi testtransform diff

( 1)  diff = 0
       F(  1,2315.1) =    1.01
            Prob > F =    0.3143

*Using the regress command
cap mi estimate:svy: reg Debt_Bus_3 i.Home_Based_3
mi test 1.Home_Based_3

note: assuming equal fractions of missing information

 ( 1)  1.Home_Based_3 = 0

       F(  1,2315.1) =    1.01
            Prob > F =    0.3143
```

```
*Differences of Means for Two Subpopulations (White vs. Black in 2007)
*Using the mi testtransform command
mi estimate (diff: [Debt_Bus_3]5 - [Debt_Bus_3]4 ), saving(miest, replace):svy:
mean    Debt_Bus_3 , over(PO_race_group_3)

Multiple-imputation estimates      Imputations     =         5
Survey: Mean estimation            Number of obs   =      2330

Number of strata  =         6      Population size = 51665.577
Number of PSUs    =      2330
                                   Average RVI     =    0.1143
                                   Largest FMI     =    0.4358
                                   Complete DF     =      2324
DF adjustment:   Small sample      DF:     min     =     25.36
                                           avg     =   1794.59
Within VCE type:   Linearized              max     =   2322.00

            1: PO_race_group_3 = 1
            2: PO_race_group_3 = 2
            3: PO_race_group_3 = 3
            4: PO_race_group_3 = 4
            5: PO_race_group_3 = 5
            6: PO_race_group_3 = 6

------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------
           1 |   3299.178     1431.9      491.2417    6107.115
           2 |   1561.998   1098.006     -697.7628    3821.758
           3 |   15637.59   7349.678      1224.846    30050.34
           4 |   9131.632   3160.482       2933.97    15329.29
           5 |   52452.14   16674.07      19754.47     85149.8
           6 |   8344.614   3314.038      1843.974    14845.25
------------------------------------------------------------

Transformations                    Average RVI     =    0.0015
                                   Largest FMI     =    0.0015
                                   Complete DF     =      2324
DF adjustment:   Small sample      DF:     min     =   2315.78
                                           avg     =   2315.78
Within VCE type:   Linearized              max     =   2315.78

         diff: [Debt_Bus_3]5 - [Debt_Bus_3]4

------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------
        diff |    43320.5   16972.44      10037.74    76603.27
------------------------------------------------------------

mi testtransform diff

         diff: [Debt_Bus_3]5 - [Debt_Bus_3]4

 ( 1)  diff = 0

       F(  1,2315.8) =    6.51
            Prob > F =    0.0108
```

```
*Using the regress command
cap mi estimate:svy , subpop(if PO_race_group_3==4 | PO_race_group_3==5): reg
Debt_Bus_3 i.PO_race_group_3
mi test 5.PO_race_group_3

( 1)  5.PO_race_group_3 = 0

       F(  1,2315.8) =    6.51
            Prob > F =    0.0108

*KFS format : Long
* Using The KFS multiply imputed data
use Longitudinal_Long_MI_Long_L2,clear
mi svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
*Differences of Means for Two Subpopulations (Home vs. Non Home based in 2007)
mi  estimate:svy: mean Debt_Bus if year==2007, over(Home_Based)
mi    estimate  (diff:  [Debt_Bus]_subpop_1-[Debt_Bus]_subpop_2),  saving(miest,
replace) :svy: mean Debt_Bus if year==2007, over(Home_Based)

Multiple-imputation estimates      Imputations     =         5
Survey: Mean estimation            Number of obs   =      2330

Number of strata  =         6      Population size = 51665.577
Number of PSUs    =      2330
                                   Average RVI     =    0.0018
                                   Largest FMI     =    0.0036
                                   Complete DF     =      2324
DF adjustment:   Small sample      DF:     min     =   2296.25
                                           avg     =   2309.12
Within VCE type:   Linearized              max     =   2321.99

    _subpop_1: Home_Based = Non Home Based
    _subpop_2: Home_Based = Home Based

------------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
   _subpop_1 |   59937.62    18216.1      24215.89    95659.35
   _subpop_2 |   31842.08   21144.92     -9622.807    73306.97
------------------------------------------------------------------

Transformations                    Average RVI     =    0.0016
                                   Largest FMI     =    0.0016
                                   Complete DF     =      2324
DF adjustment:   Small sample      DF:     min     =   2315.12
                                           avg     =   2315.12
Within VCE type:   Linearized              max     =   2315.12

         diff: [Debt_Bus]_subpop_1-[Debt_Bus]_subpop_2

------------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
        diff |   28095.54   27914.87     -26645.21    82836.29
------------------------------------------------------------------

mi testtransform diff
( 1)  diff = 0

       F(  1,2315.1) =    1.01
            Prob > F =    0.3143
```

```
*We can also get the same results using regression
mi  estimate:svy: reg Debt_Bus Home_Based if year==2007

Multiple-imputation estimates                   Imputations        =         5
Survey: Linear regression                       Number of obs      =      2330

Number of strata  =         6                   Population size    = 51665.577
Number of PSUs    =      2330
                                                Average RVI        =    0.0018
                                                Largest FMI        =    0.0036
                                                Complete DF        =      2324
DF adjustment:   Small sample                   DF:     min        =   2296.25
                                                        avg        =   2305.69
                                                        max        =   2315.12
Model F test:       Equal FMI                   F(   1, 2315.1)    =      1.01
Within VCE type:   Linearized                   Prob > F           =    0.3143

------------------------------------------------------------------------------
    Debt_Bus |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Home_Based |  -28095.54   27914.87    -1.01   0.314    -82836.29    26645.21
       _cons |   59937.62    18216.1     3.29   0.001     24215.89    95659.35
------------------------------------------------------------------------------

mi test  Home_Based

( 1)  Home_Based = 0

       F(  1,2315.1) =    1.01
            Prob > F =    0.3143
```

```
*Differences of Means for Two Subpopulations (White vs. Black in 2007)
mi  estimate:svy: mean    Debt_Bus if year==2007 , over(PO_race_group )
mi estimate (diff: [Debt_Bus]5 - [Debt_Bus]4 ), saving(miest, replace):svy: mean
Debt_Bus if year==2007, over(PO_race_group)


Multiple-imputation estimates      Imputations     =         5
Survey: Mean estimation            Number of obs   =      2330

Number of strata  =         6      Population size = 51665.577
Number of PSUs    =      2330
                                   Average RVI     =    0.1143
                                   Largest FMI     =    0.4358
                                   Complete DF     =      2324
DF adjustment:   Small sample      DF:     min     =     25.36
                                           avg     =   1794.59
Within VCE type:   Linearized              max     =   2322.00

            1: PO_race_group = 1
            2: PO_race_group = 2
            3: PO_race_group = 3
            4: PO_race_group = 4
            5: PO_race_group = 5
            6: PO_race_group = 6

------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------
           1 |   3299.178     1431.9      491.2417    6107.115
           2 |   1561.998   1098.006     -697.7628    3821.758
           3 |   15637.59   7349.678      1224.846    30050.34
           4 |   9131.632   3160.482       2933.97    15329.29
           5 |   52452.14   16674.07      19754.47     85149.8
           6 |   8344.614   3314.038      1843.974    14845.25
------------------------------------------------------------

Transformations                    Average RVI     =    0.0015
                                   Largest FMI     =    0.0015
                                   Complete DF     =      2324
DF adjustment:   Small sample      DF:     min     =   2315.78
                                           avg     =   2315.78
Within VCE type:   Linearized              max     =   2315.78

         diff: [Debt_Bus]5 - [Debt_Bus]4

------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------
        diff |    43320.5   16972.44      10037.74    76603.27
------------------------------------------------------------

mi testtransform diff

( 1)  diff = 0

       F(  1,2315.8) =    6.51
            Prob > F =    0.0108
```

```
*We can also get the same results using regression
mi   estimate:svy, subpop(if PO_race_group==4 | PO_race_group==5): reg Debt_Bus
i.PO_race_group if year==2007

Multiple-imputation estimates                   Imputations        =         5
Survey: Linear regression                       Number of obs      =      2330

Number of strata  =         6                   Population size    = 51665.577
Number of PSUs    =      2330                   Subpop. no. of obs =      2147
                                                Subpop. size       = 47389.648
                                                Average RVI        =    0.0008
                                                Largest FMI        =    0.0015
                                                Complete DF        =      2324
DF adjustment:   Small sample                   DF:     min        =   2315.78
                                                        avg        =   2318.89
                                                        max        =   2322.00
Model F test:       Equal FMI                   F(   1, 2315.8)    =      6.51
Within VCE type:   Linearized                   Prob > F           =    0.0108

------------------------------------------------------------------------------
     Debt_Bus |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
5.PO_race_g~p |    43320.5   16972.44     2.55   0.011     10037.74    76603.27
        _cons |   9131.632   3160.482     2.89   0.004      2933.97    15329.29
------------------------------------------------------------------------------

mi test 5.PO_race_group

( 1)  5.PO_race_group = 0

       F(  1,2315.8) =    6.51
            Prob > F =    0.0108
```

Example 4.30: Differences of Means over Time

```
 *KFS format : Long
*Example 4.30: Differences of Means over Time
use Longitudinal_Long_MI_Long_L2,clear
*Declare survey design for dataset
mi svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
*Differences of Means  (2004 vs. 2011)
mi  estimate (diff: [Debt_Bus]2004-[Debt_Bus]2011), saving(miest, replace) :svy:
mean Debt_Bus if year==2004 | year==2011 , over(year)

Multiple-imputation estimates      Imputations     =         5
Survey: Mean estimation            Number of obs   =      4770

Number of strata  =         6      Population size = 108170.48
Number of PSUs    =      4770
                                   Average RVI     =    0.0027
                                   Largest FMI     =    0.0051
                                   Complete DF     =      4764
DF adjustment:   Small sample      DF:     min     =   4595.12
                                           avg     =   4677.39
Within VCE type:   Linearized              max     =   4759.67

         2004: year = 2004
         2011: year = 2011

------------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
        2004 |   38121.59   8410.318      21633.48     54609.7
        2011 |   25566.51   4224.924      17283.63    33849.39
------------------------------------------------------------------

Transformations                    Average RVI     =    0.0019
                                   Largest FMI     =    0.0019
                                   Complete DF     =      4764
DF adjustment:   Small sample      DF:     min     =   4732.79
                                           avg     =   4732.79
Within VCE type:   Linearized              max     =   4732.79

         diff: [Debt_Bus]2004-[Debt_Bus]2011

------------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
        diff |   12555.08   9414.037     -5900.811    31010.98
------------------------------------------------------------------

mi testtransform diff

( 1)  diff = 0

       F(  1,4732.8) =    1.78
            Prob > F =    0.1824
```

```
*Using the regress command
mi  estimate:svy: reg Debt_Bus i.year if year==2004 | year==2011

Multiple-imputation estimates                   Imputations        =         5
Survey: Linear regression                       Number of obs      =      4770

Number of strata  =         6                   Population size    = 108170.48
Number of PSUs    =      4770
                                                Average RVI        =    0.0027
                                                Largest FMI        =    0.0019
                                                Complete DF        =      4764
DF adjustment:   Small sample                   DF:     min        =   4732.79
                                                        avg        =   4746.23
                                                        max        =   4759.67
Model F test:       Equal FMI                   F(   1, 4732.8)    =      1.78
Within VCE type:   Linearized                   Prob > F           =    0.1824

------------------------------------------------------------------------------
    Debt_Bus |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |
       2011  |  -12555.08   9414.037    -1.33   0.182    -31010.98    5900.811
             |
       _cons |   38121.59   8410.318     4.53   0.000     21633.48     54609.7
------------------------------------------------------------------------------

*Using the test command
mi test  2011.year

1)  2011.year = 0

       F(  1,4732.8) =    1.78
            Prob > F =    0.1824
```

```
*Comparing the Mean of Differences for Paired Data
*Businesses that they responded to the survey in both years
*This test is used when the samples are dependent; that is, when there is only one
sample that has been tested twice (repeated measures)


*KFS format : Long
use Longitudinal_Long_MI_Long_L2,clear
*Declare survey design for dataset
mi svyset  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
*Differences of Means  (2004 vs. 2011)
mi xtset mprid year
*Mark businesses that they responded to the survey in both years
mi xeq: bysort mprid (year): gen diff=Debt_Bus[1]-Debt_Bus[8]  if year==2004 |
year==2011
mi  estimate:svy:mean diff if year==2004

Multiple-imputation estimates      Imputations     =         5
Survey: Mean estimation            Number of obs   =      1630

Number of strata  =         6      Population size = 34892.039
Number of PSUs    =      1630
                                   Average RVI     =    0.0014
                                   Largest FMI     =    0.0014
                                   Complete DF     =      1624
DF adjustment:   Small sample      DF:     min     =   1618.52
                                           avg     =   1618.52
Within VCE type:   Linearized              max     =   1618.52

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        diff |    16700.3   10411.71     -3721.552    37122.15
--------------------------------------------------------------

mi  test diff

( 1)  diff = 0

       F(  1,1618.5) =    2.57
            Prob > F =    0.1089
```

```
*Using the regress command
mi  estimate:svy: reg Debt_Bus i.year if   diff<.

Multiple-imputation estimates                   Imputations        =         5
Survey: Linear regression                       Number of obs      =      3260

Number of strata  =         6                   Population size    = 69784.078
Number of PSUs    =      3260
                                                Average RVI        =    0.0028
                                                Largest FMI        =    0.0014
                                                Complete DF        =      3254
DF adjustment:   Small sample                   DF:     min        =   3242.79
                                                        avg        =   3246.60
                                                        max        =   3250.42
Model F test:       Equal FMI                   F(   1, 3242.8)    =      2.52
Within VCE type:   Linearized                   Prob > F           =    0.1122

------------------------------------------------------------------------------
    Debt_Bus |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |
       2011  |   -16700.3   10510.33    -1.59   0.112    -37307.86     3907.26
             |
       _cons |   42266.81   9626.712     4.39   0.000     23391.77    61141.84
------------------------------------------------------------------------------

*Using the test command
mi test  2011.year

( 1)  2011.year = 0

       F(  1,3242.8) =    2.52
            Prob > F =    0.1122
```

4.5.3. FR Special Commands Suite

The FR Commands Suite (FR_Commands) is a special commands suite designed to work only with KFS data. The purpose of these commands is to shorten the number of commands needed to create descriptive statistics. In addition, unlike standard Stata command FR_ Commands do not abbreviate the long variable names.

4.5.3.1. Command: [bysort varname:]FR_Sum_W varlist [if] [pweight] , casewise

The command is designed to work with KFS in wide format. It reports the number of non-missing, soft missing, and hard missing observations, and the mean, min, median, and max for the varlist.

- The command allows restricting the sample using if.
- The casewise option allows reporting the statistics for the varlist using casewise (listwise) deletion.
- The command accepts the "bysort" prefix.

```
*KFS format : Wide
program drop _all
adopath + " C:\KFS_Manual_and_Data\Farhat_Robb_Commands"
use Longitudinal_wide_MI_Long_w2,clear
*Extract the original KFS data (m=0)
mi extract 0
FR_Sum_W  Assets_5 Total_Employees_5 Net_Profit_5  PO_gender_5 OO_gender_owner_5  [pweight=wgt_7_long]


                         Variable ,     N ,   N=. ,  N=.a ,          Mean ,        Min ,     Median ,             Max
                         Assets_5 ,  1745 ,   131 ,  1264 ,    711324.0346 ,     0.0000 , 35500.0000 ,   230000000.0000
                Total_Employees_5 ,  1867 ,     9 ,  1264 ,        4.0662 ,     0.0000 ,     2.0000 ,         267.0000
                     Net_Profit_5 ,  1804 ,    72 ,  1264 ,     28045.1832 , -5.000e+06 ,  2000.0000 ,     7500000.0000
                      PO_gender_5 ,  1876 ,     0 ,  1264 ,        0.7076 ,     0.0000 ,     1.0000 ,           1.0000
                OO_gender_owner_5 ,  1875 ,     1 ,  1264 ,        0.6820 ,     0.0000 ,     1.0000 ,           1.0000

FR_Sum_W  Assets_5 Total_Employees_5 Net_Profit_5  PO_gender_5 OO_gender_owner_5  [pweight=wgt_7_long], casewise

                         Variable ,     N ,   N=. ,  N=.a ,          Mean ,        Min ,     Median ,             Max
                         Assets_5 ,  1687 ,     . ,     . ,    390554.5895 ,     0.0000 , 35575.0000 ,   100213504.0000
                Total_Employees_5 ,  1687 ,     . ,     . ,        3.9991 ,     0.0000 ,     2.0000 ,         267.0000
                     Net_Profit_5 ,  1687 ,     . ,     . ,     27741.9035 , -5.000e+06 ,  2000.0000 ,     7500000.0000
                      PO_gender_5 ,  1687 ,     . ,     . ,        0.7096 ,     0.0000 ,     1.0000 ,           1.0000
                OO_gender_owner_5 ,  1687 ,     . ,     . ,        0.6855 ,     0.0000 ,     1.0000 ,           1.0000
```

```
use Cross_Sectional_wide_MI_Long_w2,clear
*Extract the original KFS data (m=0)
mi extract 0
FR_Sum_W    Assets_5  Total_Employees_5  Net_Profit_5    PO_gender_5  OO_gender_owner_5    [pweight=cswgt_final_5]  if
Home_Based_5==1

                         Variable ,     N ,   N=. ,  N=.a ,          Mean ,        Min ,     Median ,             Max
                         Assets_5 ,  1150 ,    70 ,     0 ,    123357.7373 ,     0.0000 , 15000.0000 ,    10385000.0000
                Total_Employees_5 ,  1210 ,    10 ,     0 ,        1.5221 ,     0.0000 ,     1.0000 ,          54.0000
                     Net_Profit_5 ,  1167 ,    53 ,     0 ,     10528.7729 , -2.500e+05 ,   300.0000 ,     1000000.0000
                      PO_gender_5 ,  1220 ,     . ,     . ,        0.6649 ,     0.0000 ,     1.0000 ,           1.0000
                OO_gender_owner_5 ,  1219 ,     1 ,     0 ,        0.6564 ,     0.0000 ,     1.0000 ,           1.0000

FR_Sum_W    Assets_5  Total_Employees_5  Net_Profit_5    PO_gender_5  OO_gender_owner_5    [pweight=cswgt_final_5]  if
Home_Based_5==1, casewise

                         Variable ,     N ,   N=. ,  N=.a ,          Mean ,        Min ,     Median ,             Max
                         Assets_5 ,  1103 ,     . ,     . ,    125768.9569 ,     0.0000 , 15000.0000 ,    10385000.0000
                Total_Employees_5 ,  1103 ,     . ,     . ,        1.4524 ,     0.0000 ,     1.0000 ,          44.0000
                     Net_Profit_5 ,  1103 ,     . ,     . ,     11213.8651 , -2.500e+05 ,   400.0000 ,     1000000.0000
                      PO_gender_5 ,  1103 ,     . ,     . ,        0.6678 ,     0.0000 ,     1.0000 ,           1.0000
                OO_gender_owner_5 ,  1103 ,     . ,     . ,        0.6592 ,     0.0000 ,     1.0000 ,           1.0000

bysort   c8_primary_loc_5:   FR_Sum_W     Assets_5   Total_Employees_5   Net_Profit_5     PO_gender_5   OO_gender_owner_5
[pweight=cswgt_final_5]
-------------------------------------------------------------------------------------------------------------------------
-> c8_primary_loc_5 = A residence such as a home or garage
                         Variable ,     N ,   N=. ,  N=.a ,          Mean ,        Min ,     Median ,             Max
                         Assets_5 ,  1150 ,    70 ,     0 ,    123357.7373 ,     0.0000 , 15000.0000 ,    10385000.0000
                Total_Employees_5 ,  1210 ,    10 ,     0 ,        1.5221 ,     0.0000 ,     1.0000 ,          54.0000
                     Net_Profit_5 ,  1167 ,    53 ,     0 ,     10528.7729 , -2.500e+05 ,   300.0000 ,     1000000.0000
                      PO_gender_5 ,  1220 ,     . ,     . ,        0.6649 ,     0.0000 ,     1.0000 ,           1.0000
                OO_gender_owner_5 ,  1219 ,     1 ,     0 ,        0.6564 ,     0.0000 ,     1.0000 ,           1.0000
-------------------------------------------------------------------------------------------------------------------------
-> c8_primary_loc_5 = A rented or leased space
                         Variable ,     N ,   N=. ,  N=.a ,          Mean ,        Min ,     Median ,             Max
                         Assets_5 ,   813 ,    76 ,     0 ,   2854909.7321 ,     0.0000 , 77000.0000 ,   701524992.0000
                Total_Employees_5 ,   883 ,     6 ,     0 ,        7.4640 ,     0.0000 ,     3.0000 ,         176.0000
                     Net_Profit_5 ,   845 ,    44 ,     0 ,     32028.0460 , -5.400e+07 ,  7000.0000 ,    17000000.0000
                      PO_gender_5 ,   888 ,     1 ,     0 ,        0.7171 ,     0.0000 ,     1.0000 ,           1.0000
                OO_gender_owner_5 ,   886 ,     3 ,     0 ,        0.6873 ,     0.0000 ,     1.0000 ,           1.0000
```

```
-------------------------------------------------------------------------------------------------------------------
-> c8_primary_loc_5 = Space the business purchased
                    Variable ,     N ,   N=. ,  N=.a ,          Mean ,        Min ,     Median ,             Max
                    Assets_5 ,   164 ,    19 ,     0 ,   1201078.9300 ,     0.0000 ,  4.260e+05 ,    10980000.0000
              Total_Employees_5 ,   183 ,     . ,     . ,        7.6275 ,     0.0000 ,     3.0000 ,         171.0000
                   Net_Profit_5 ,   174 ,     9 ,     0 ,     36898.0611 , -2.600e+06 ,  9000.0000 ,     2000000.0000
                    PO_gender_5 ,   183 ,     . ,     . ,        0.8171 ,     0.0000 ,     1.0000 ,           1.0000
              OO_gender_owner_5 ,   183 ,     . ,     . ,        0.7431 ,     0.0000 ,     1.0000 ,           1.0000
-------------------------------------------------------------------------------------------------------------------
-> c8_primary_loc_5 = A site where a client is located
                    Variable ,     N ,   N=. ,  N=.a ,          Mean ,        Min ,     Median ,             Max
                    Assets_5 ,   101 ,     5 ,     0 ,   2204156.7749 ,     0.0000 , 15000.0000 ,   100213504.0000
              Total_Employees_5 ,   104 ,     2 ,     0 ,        3.5565 ,     0.0000 ,     1.0000 ,         267.0000
                   Net_Profit_5 ,   101 ,     5 ,     0 ,     24593.1453 , -5.000e+06 ,  2000.0000 ,     1955400.0000
                    PO_gender_5 ,   106 ,     . ,     . ,        0.6796 ,     0.0000 ,     1.0000 ,           1.0000
              OO_gender_owner_5 ,   106 ,     . ,     . ,        0.6670 ,     0.0000 ,     1.0000 ,           1.0000
-------------------------------------------------------------------------------------------------------------------
-> c8_primary_loc_5 = Some other location? (SPECIFY)
                    Variable ,     N ,   N=. ,  N=.a ,          Mean ,        Min ,     Median ,             Max
                    Assets_5 ,     8 ,     2 ,     0 ,   2461784.5257 ,  1350.0000 , 30000.0000 ,    10342000.0000
              Total_Employees_5 ,    10 ,     . ,     . ,        2.9063 ,     0.0000 ,     2.0000 ,          10.0000
                   Net_Profit_5 ,     8 ,     2 ,     0 ,     17845.6514 , -1.000e+04 , 15000.0000 ,       50000.0000
                    PO_gender_5 ,    10 ,     . ,     . ,        0.9296 ,     0.0000 ,     1.0000 ,           1.0000
              OO_gender_owner_5 ,    10 ,     . ,     . ,        0.8970 ,     0.0000 ,     1.0000 ,           1.0000
-------------------------------------------------------------------------------------------------------------------
-> c8_primary_loc_5 = .a
                    Variable ,     N ,   N=. ,  N=.a ,          Mean ,        Min ,     Median ,             Max
                    Assets_5 ,     0 ,     0 ,  2520 ,             . ,          . ,          . ,               .
              Total_Employees_5 ,     0 ,     0 ,  2520 ,             . ,          . ,          . ,               .
                   Net_Profit_5 ,     0 ,     0 ,  2520 ,             . ,          . ,          . ,               .
                    PO_gender_5 ,     0 ,     0 ,  2520 ,             . ,          . ,          . ,               .
              OO_gender_owner_5 ,     0 ,     0 ,  2520 ,             . ,          . ,          . ,               .
Mean,Min,Median,and Max are weighted by pweight = cswgt_final_5
Mean,Min,Median and Max are calculated using nonmissing vaules for each var in the varlist independently
```

```
bysort   c8_primary_loc_5:   FR_Sum_W     Assets_5   Total_Employees_5   Net_Profit_5     PO_gender_5   OO_gender_owner_5
[pweight=cswgt_final_5], casewise
-------------------------------------------------------------------------------------------------------------------------
-> c8_primary_loc_5 = A residence such as a home or garage
                         Variable ,     N ,   N=. ,  N=.a ,          Mean ,        Min ,     Median ,             Max
                         Assets_5 ,  1103 ,     . ,     . ,    125768.9569 ,     0.0000 , 15000.0000 ,    10385000.0000
                Total_Employees_5 ,  1103 ,     . ,     . ,        1.4524 ,     0.0000 ,     1.0000 ,          44.0000
                     Net_Profit_5 ,  1103 ,     . ,     . ,    11213.8651 , -2.500e+05 ,   400.0000 ,     1000000.0000
                      PO_gender_5 ,  1103 ,     . ,     . ,        0.6678 ,     0.0000 ,     1.0000 ,           1.0000
                OO_gender_owner_5 ,  1103 ,     . ,     . ,        0.6592 ,     0.0000 ,     1.0000 ,           1.0000
-------------------------------------------------------------------------------------------------------------------------
-> c8_primary_loc_5 = A rented or leased space
                         Variable ,     N ,   N=. ,  N=.a ,          Mean ,        Min ,     Median ,             Max
                         Assets_5 ,   779 ,     . ,     . ,   2466985.9869 ,     0.0000 , 77000.0000 ,   701524992.0000
                Total_Employees_5 ,   779 ,     . ,     . ,        7.3182 ,     0.0000 ,     4.0000 ,         176.0000
                     Net_Profit_5 ,   779 ,     . ,     . ,    32362.3727 , -5.400e+07 ,  6000.0000 ,    17000000.0000
                      PO_gender_5 ,   779 ,     . ,     . ,        0.7144 ,     0.0000 ,     1.0000 ,           1.0000
                OO_gender_owner_5 ,   779 ,     . ,     . ,        0.6853 ,     0.0000 ,     1.0000 ,           1.0000
-------------------------------------------------------------------------------------------------------------------------
-> c8_primary_loc_5 = Space the business purchased
                         Variable ,     N ,   N=. ,  N=.a ,          Mean ,        Min ,     Median ,             Max
                         Assets_5 ,   158 ,     . ,     . ,   1206831.3559 ,     0.0000 ,  4.500e+05 ,    10980000.0000
                Total_Employees_5 ,   158 ,     . ,     . ,        7.7580 ,     0.0000 ,     3.0000 ,         171.0000
                     Net_Profit_5 ,   158 ,     . ,     . ,    38614.1797 , -2.600e+06 ,  9000.0000 ,     2000000.0000
                      PO_gender_5 ,   158 ,     . ,     . ,        0.8480 ,     0.0000 ,     1.0000 ,           1.0000
                OO_gender_owner_5 ,   158 ,     . ,     . ,        0.7708 ,     0.0000 ,     1.0000 ,           1.0000
-------------------------------------------------------------------------------------------------------------------------
-> c8_primary_loc_5 = A site where a client is located
                         Variable ,     N ,   N=. ,  N=.a ,          Mean ,        Min ,     Median ,             Max
                         Assets_5 ,    97 ,     . ,     . ,   2261418.1986 ,     0.0000 , 15000.0000 ,   100213504.0000
                Total_Employees_5 ,    97 ,     . ,     . ,        3.6692 ,     0.0000 ,     1.0000 ,         267.0000
                     Net_Profit_5 ,    97 ,     . ,     . ,    25811.7497 , -5.000e+06 ,  2000.0000 ,     1955400.0000
                      PO_gender_5 ,    97 ,     . ,     . ,        0.6823 ,     0.0000 ,     1.0000 ,           1.0000
                OO_gender_owner_5 ,    97 ,     . ,     . ,        0.6640 ,     0.0000 ,     1.0000 ,           1.0000
-------------------------------------------------------------------------------------------------------------------------
-> c8_primary_loc_5 = Some other location? (SPECIFY)
                         Variable ,     N ,   N=. ,  N=.a ,          Mean ,        Min ,     Median ,             Max
                         Assets_5 ,     8 ,     . ,     . ,   2461784.5257 ,  1350.0000 , 30000.0000 ,    10342000.0000
                Total_Employees_5 ,     8 ,     . ,     . ,        3.3977 ,     0.0000 ,     2.0000 ,          10.0000
                     Net_Profit_5 ,     8 ,     . ,     . ,    17845.6514 , -1.000e+04 , 15000.0000 ,       50000.0000
                      PO_gender_5 ,     8 ,     . ,     . ,        1.0000 ,     1.0000 ,     1.0000 ,           1.0000
                OO_gender_owner_5 ,     8 ,     . ,     . ,        0.9574 ,     0.5000 ,     1.0000 ,           1.0000
```

4.5.3.2. Command: [bysort varname:]FR_Sum_L varlist [if] [pweight] [, casewise]

The command is designed to work with KFS in long format. It reports the number of non-missing, soft missing, and hard missing observations, and the mean, min, median, and max for the varlist.

- The command allows restricting the sample using if.
- The casewise option allows reporting the statistics for the varlist using casewise (listwise) deletion.
- The command accepts the "bysort" prefix.

```
*KFS format : Long
program drop _all
adopath + "C:\KFS_Manual_and_Data\Farhat_Robb_Commands"
use Longitudinal_Long_MI_Long_L2,clear
*Extract the original KFS data (m=0)
mi extract 0
FR_Sum_L  Assets Total_Employees  [pweight=wgt_7_long]  if year==2005

                        Variable ,      N ,    N=. ,   N=.a ,         Mean ,          Min ,       Median ,          Max
                          Assets ,   2588 ,    196 ,     53 ,  378767.2525 ,       0.0000 ,   31800.0000 ,   2.8500e+08
                 Total_Employees ,   2711 ,     73 ,     53 ,       3.7126 ,       0.0000 ,       2.0000 ,     122.0000


Mean,Min,Median,and Max are weighted by pweight = wgt_7_long
Mean,Min,Median,and Max are calculated with the following restrictions: if year==2005

FR_Sum_L  Assets Total_Employees   [pweight=wgt_7_long] if Home_Based==1 & year==2005,casewise


                        Variable ,      N ,    N=. ,   N=.a ,         Mean ,          Min ,       Median ,          Max
                          Assets ,   1356 ,      . ,      . ,  107823.8276 ,       0.0000 ,   15000.0000 ,   1.0949e+07
                 Total_Employees ,   1356 ,      . ,      . ,       1.8066 ,       0.0000 ,       1.0000 ,      30.0000

Mean,Min,Median,and Max are weighted by pweight = wgt_7_long
Mean,Min,Median,and Max are calculated using casewise(listwise) deletion—if an observation has a missing value in any
of the variables, it is to be excluded from all the calculations
Mean,Min,Median,and Max are calculated with the following restrictions: if Home_Based==1 & year==2005
```

```
bysort year :FR_Sum_L  Assets Total_Employees  [pweight=wgt_7_long]  if Home_Based==1

-------------------------------------------------------------------------------------------------------------------
-> year = 2004
                    Variable ,      N ,    N=. ,   N=.a ,         Mean ,          Min ,       Median ,          Max
                      Assets ,   1600 ,    105 ,      0 ,   88230.4630 ,       0.0000 ,    9000.0000 ,   8.0001e+07
                Total_Employees ,   1657 ,     48 ,      0 ,       1.2321 ,       0.0000 ,       1.0000 ,      21.0000
-------------------------------------------------------------------------------------------------------------------
-> year = 2005
                    Variable ,      N ,    N=. ,   N=.a ,         Mean ,          Min ,       Median ,          Max
                      Assets ,   1387 ,     94 ,      0 ,  109519.3388 ,       0.0000 ,   15000.0000 ,   1.0949e+07
                Total_Employees ,   1446 ,     35 ,      0 ,       1.8096 ,       0.0000 ,       1.0000 ,      30.0000
-------------------------------------------------------------------------------------------------------------------
-> year = 2006
                    Variable ,      N ,    N=. ,   N=.a ,         Mean ,          Min ,       Median ,          Max
                      Assets ,   1222 ,     81 ,      0 ,  112935.0006 ,       0.0000 ,   18500.0000 , 5678000.0000
                Total_Employees ,   1277 ,     26 ,      0 ,       1.8610 ,       0.0000 ,       1.0000 ,      36.0000
-------------------------------------------------------------------------------------------------------------------
-> year = 2007
                    Variable ,      N ,    N=. ,   N=.a ,         Mean ,          Min ,       Median ,          Max
                      Assets ,   1117 ,     63 ,      0 ,  113306.3623 ,       0.0000 ,   19200.0000 , 7379000.0000
                Total_Employees ,   1162 ,     18 ,      0 ,       1.6781 ,       0.0000 ,       1.0000 ,      33.0000
-------------------------------------------------------------------------------------------------------------------
-> year = 2008
                    Variable ,      N ,    N=. ,   N=.a ,         Mean ,          Min ,       Median ,          Max
                      Assets ,   1009 ,     75 ,      0 ,  149529.3805 ,       0.0000 ,   15000.0000 ,   1.1800e+07
                Total_Employees ,   1078 ,      6 ,      0 ,       1.6550 ,       0.0000 ,       1.0000 ,      59.0000
-------------------------------------------------------------------------------------------------------------------
-> year = 2009
                    Variable ,      N ,    N=. ,   N=.a ,         Mean ,          Min ,       Median ,          Max
                      Assets ,    949 ,     57 ,      0 ,  122365.1814 ,       0.0000 ,   14500.0000 , 7915000.0000
                Total_Employees ,   1000 ,      6 ,      0 ,       1.5149 ,       0.0000 ,       1.0000 ,      54.0000
-------------------------------------------------------------------------------------------------------------------
-> year = 2010
                    Variable ,      N ,    N=. ,   N=.a ,         Mean ,          Min ,       Median ,          Max
                      Assets ,    862 ,     54 ,      0 ,  172330.4715 ,       0.0000 ,   16200.0000 ,   2.5042e+07
                Total_Employees ,    911 ,      5 ,      0 ,       1.6533 ,       0.0000 ,       1.0000 ,      53.0000
-------------------------------------------------------------------------------------------------------------------
-> year = 2011
                    Variable ,      N ,    N=. ,   N=.a ,         Mean ,          Min ,       Median ,          Max
                      Assets ,    794 ,     56 ,      0 , 1639346.8562 ,       0.0000 ,   15000.0000 ,   3.0057e+08
                Total_Employees ,    844 ,      6 ,      0 ,       1.4947 ,       0.0000 ,       1.0000 ,      56.0000
```

```
bysort year :FR_Sum_L  Assets Total_Employees  [pweight=wgt_7_long]  if Home_Based==1,casewise

-------------------------------------------------------------------------------------------------------------------
-> year = 2004
                    Variable ,      N ,    N=. ,   N=.a ,         Mean ,          Min ,       Median ,          Max
                      Assets ,   1562 ,      . ,      . ,   90136.8966 ,       0.0000 ,    9000.0000 ,   8.0001e+07
                Total_Employees ,   1562 ,      . ,      . ,       1.2316 ,       0.0000 ,       1.0000 ,      21.0000
-------------------------------------------------------------------------------------------------------------------
-> year = 2005
                    Variable ,      N ,    N=. ,   N=.a ,         Mean ,          Min ,       Median ,          Max
                      Assets ,   1356 ,      . ,      . ,  107823.8276 ,       0.0000 ,   15000.0000 ,   1.0949e+07
                Total_Employees ,   1356 ,      . ,      . ,       1.8066 ,       0.0000 ,       1.0000 ,      30.0000
-------------------------------------------------------------------------------------------------------------------
-> year = 2006
                    Variable ,      N ,    N=. ,   N=.a ,         Mean ,          Min ,       Median ,          Max
                      Assets ,   1199 ,      . ,      . ,  112050.3587 ,       0.0000 ,   18800.0000 , 5678000.0000
                Total_Employees ,   1199 ,      . ,      . ,       1.8488 ,       0.0000 ,       1.0000 ,      32.0000
-------------------------------------------------------------------------------------------------------------------
-> year = 2007
                    Variable ,      N ,    N=. ,   N=.a ,         Mean ,          Min ,       Median ,          Max
                      Assets ,   1102 ,      . ,      . ,  112897.8101 ,       0.0000 ,   19000.0000 , 7379000.0000
                Total_Employees ,   1102 ,      . ,      . ,       1.7037 ,       0.0000 ,       1.0000 ,      33.0000
-------------------------------------------------------------------------------------------------------------------
-> year = 2008
                    Variable ,      N ,    N=. ,   N=.a ,         Mean ,          Min ,       Median ,          Max
                      Assets ,   1003 ,      . ,      . ,  148781.1547 ,       0.0000 ,   15000.0000 ,   1.1800e+07
                Total_Employees ,   1003 ,      . ,      . ,       1.6866 ,       0.0000 ,       1.0000 ,      59.0000
-------------------------------------------------------------------------------------------------------------------
-> year = 2009
                    Variable ,      N ,    N=. ,   N=.a ,         Mean ,          Min ,       Median ,          Max
                      Assets ,    946 ,      . ,      . ,  122761.6723 ,       0.0000 ,   14550.0000 , 7915000.0000
                Total_Employees ,    946 ,      . ,      . ,       1.4308 ,       0.0000 ,       1.0000 ,      44.0000
-------------------------------------------------------------------------------------------------------------------
-> year = 2010
                    Variable ,      N ,    N=. ,   N=.a ,         Mean ,          Min ,       Median ,          Max
                      Assets ,    859 ,      . ,      . ,  172581.6429 ,       0.0000 ,   16000.0000 ,   2.5042e+07
                Total_Employees ,    859 ,      . ,      . ,       1.6837 ,       0.0000 ,       1.0000 ,      53.0000
-------------------------------------------------------------------------------------------------------------------
-> year = 2011
                    Variable ,      N ,    N=. ,   N=.a ,         Mean ,          Min ,       Median ,          Max
                      Assets ,    789 ,      . ,      . , 1650420.0522 ,       0.0000 ,   15000.0000 ,   3.0057e+08
                Total_Employees ,    789 ,      . ,      . ,       1.4698 ,       0.0000 ,       1.0000 ,      56.0000
```

4.5.3.3. Command: [bysort varname:]FR_Sum_MI_W varlist [if] [pweight] [, casewise]

The command is designed to work with KFS multiply imputed data in wide format. It reports the mean using imputed data, number of non-missing observations, the minimum number of observations among m, and the maximum number of observations among m for the varlist using imputed data. For non-imputed data, the command reports the mean, and the number of non-missing, soft missing, and hard missing observations.

- The command allows restricting the sample using if.
- The casewise option allows reporting the statistics for the varlist using casewise (listwise) deletion.
- The command accepts the "bysort" prefix.

```
* Using The KFS multiply imputed data
program drop _all
adopath + "C:\KFS_Manual_and_Data\Farhat_Robb_Commands"

use Cross_Sectional_wide_MI_Long_w2,clear

FR_Sum_MI_W  Assets_3 Total_Employees_3 Net_Profit_3  PO_gender_3  d9a_perc_internet_sales_3  [pweight=cswgt_final_3]

                                  |    The KFS Multiply Imputed Data     |        The KFS Non-Imputed Data        |
                        Variable ,    Mean (mi) ,Min N (mi) ,Max N (mi) ,          Mean ,      N ,    N=. ,  N=.a
                        Assets_3 , 1284905.3187 ,    2915   ,     2915  , 1133588.7931  ,   2753 ,    162 ,  2013
                Total_Employees_3 ,       4.4234 ,    2915   ,     2915  ,       4.4511  ,   2879 ,     36 ,  2013
                     Net_Profit_3 ,   45089.8197 ,    2915   ,     2915  ,   47140.2399  ,   2757 ,    158 ,  2013
                      PO_gender_3 ,       0.7025 ,    2915   ,     2915  ,       0.7025  ,   2914 ,      1 ,  2013
       d9a_perc_internet_sales_3 ,       2.4419 ,     685   ,      689  ,       2.4374  ,    679 ,     23 ,  4226


Mean(mi)and Mean are weighted by pweight = cswgt_final_3

Mean(mi)and Mean are calculated using nonmissing vaules for each var in the varlist independently

 Mean(mi): the mean using the KFS multiply imputed data
 Min N (mi): the minimum number of observations among m=1,2,3,4,5
 Max N mi: the maximum number of observations among m=1,2,3,4,5
 Mean: the mean using the KFS non-imputed data
```

```
FR_Sum_MI_W        Assets_3    Total_Employees_3    Net_Profit_3        PO_gender_3        d9a_perc_internet_sales_3
[pweight=cswgt_final_3],casewise


                                |    The KFS Multiply Imputed Data     |        The KFS Non-Imputed Data        |
                      Variable ,    Mean (mi) ,Min N (mi) ,Max N (mi) ,          Mean ,      N ,    N=. ,  N=.a
                      Assets_3 ,  693764.9357 ,     685   ,      689  ,  594893.5497  ,    609 ,      . ,     .
             Total_Employees_3 ,       4.6528 ,     685   ,      689  ,       4.7560  ,    609 ,      . ,     .
                  Net_Profit_3 ,   31778.6424 ,     685   ,      689  ,   38017.4678  ,    609 ,      . ,     .
                   PO_gender_3 ,       0.7078 ,     685   ,      689  ,       0.7125  ,    609 ,      . ,     .
       d9a_perc_internet_sales_3 ,       2.4419 ,     685   ,      689  ,       2.4333  ,    609 ,      . ,     .


Mean(mi)and Mean are weighted by pweight = cswgt_final_3
Mean(mi) and Mean are calculated using casewise(listwise) deletion—if an observation has a missing value in any of the
variables, it is to be excluded from all the calculations

 Mean(mi): the mean using the KFS multiply imputed data
 Min N (mi): the minimum number of observations among m=1,2,3,4,5
 Max N mi: the maximum number of observations among m=1,2,3,4,5
 Mean: the mean using the KFS non-imputed data
```

```
bysort Home_Based_3: FR_Sum_MI_W  Assets_3 Total_Employees_3 Net_Profit_3  PO_gender_3  d9a_perc_internet_sales_3
[pweight=cswgt_final_3]


-------------------------------------------------------------------------------------------------------------------
-> Home_Based_3 = Non Home Based
                                |    The KFS Multiply Imputed Data     |        The KFS Non-Imputed Data        |
                       Variable ,    Mean (mi) ,Min N (mi) ,Max N (mi) ,          Mean ,      N ,    N=. ,  N=.a
                       Assets_3 , 2405652.5005 ,    1433   ,     1433  , 2133035.4819  ,   1346 ,     87 ,     0
              Total_Employees_3 ,       6.9585 ,    1433   ,     1433  ,       7.0016  ,   1419 ,     14 ,     0
                   Net_Profit_3 ,   76240.5802 ,    1433   ,     1433  ,   79099.5226  ,   1352 ,     81 ,     0
                    PO_gender_3 ,       0.7205 ,    1433   ,     1433  ,       0.7205  ,   1432 ,      1 ,     0
       d9a_perc_internet_sales_3 ,       2.0967 ,     374   ,      377  ,       2.0992  ,    372 ,     13 ,  1048


-------------------------------------------------------------------------------------------------------------------
-> Home_Based_3 = Home Based
                                |    The KFS Multiply Imputed Data     |        The KFS Non-Imputed Data        |
                       Variable ,    Mean (mi) ,Min N (mi) ,Max N (mi) ,          Mean ,      N ,    N=. ,  N=.a
                       Assets_3 ,  118623.3973 ,    1482   ,     1482  ,  115172.1283  ,   1407 ,     75 ,     0
              Total_Employees_3 ,       1.7852 ,    1482   ,     1482  ,       1.7866  ,   1460 ,     22 ,     0
                   Net_Profit_3 ,   12673.4376 ,    1482   ,     1482  ,   14049.8617  ,   1405 ,     77 ,     0
                    PO_gender_3 ,       0.6838 ,    1482   ,     1482  ,       0.6838  ,   1482 ,      . ,     .
       d9a_perc_internet_sales_3 ,       2.8953 ,     311   ,      312  ,       2.8857  ,    307 ,     10 ,  1165
-------------------------------------------------------------------------------------------------------------------
```

4.5.3.4. Command: [bysort varname:]FR_Sum_MI_L varlist [if] [pweight] [, casewise]

The command is designed to work with KFS multiply imputed data in long format. It reports the mean using imputed data, number of non-missing observations, the minimum number of observations among m, and the maximum number of observations among m for the varlist using imputed data. For non-imputed data, the command reports the mean, and the number of non-missing, soft missing, and hard missing observations.

- The command allows restricting the sample using if.
- The casewise option allows reporting the statistics for the varlist using casewise (listwise) deletion.
- The command accepts the "bysort" prefix.

```
*KFS format : Long
program drop _all
adopath + "C:\KFS_Manual_and_Data\Farhat_Robb_Commands"
use Cross_Sectional_Long_MI_Long_L2,clear

FR_Sum_MI_L Assets Total_Employees  [pweight=cswgt_final] if year==2007

                                |    The KFS Multiply Imputed Data     |        The KFS Non-Imputed Data        |
                       Variable ,    Mean (mi) ,Min N (mi) ,Max N (mi) ,          Mean ,      N ,    N=. ,  N=.a
                         Assets , 1284905.3187 ,    2915   ,     2915  , 1133588.7931  ,   2753 ,    162 ,  2013
                Total_Employees ,       4.4234 ,    2915   ,     2915  ,       4.4511  ,   2879 ,     36 ,  2013

bysort year:FR_Sum_MI_L Assets  [pweight=cswgt_final]

-------------------------------------------------------------------------------------------------------------------
-> year = 2004
                                |    The KFS Multiply Imputed Data     |        The KFS Non-Imputed Data        |
                       Variable ,    Mean (mi) ,Min N (mi) ,Max N (mi) ,          Mean ,      N ,    N=. ,  N=.a
                         Assets ,  362273.5114 ,    4928   ,     4928  ,  367359.8136  ,   4499 ,    429 ,     0
-------------------------------------------------------------------------------------------------------------------
-> year = 2005
                                |    The KFS Multiply Imputed Data     |        The KFS Non-Imputed Data        |
                       Variable ,    Mean (mi) ,Min N (mi) ,Max N (mi) ,          Mean ,      N ,    N=. ,  N=.a
                         Assets ,  805040.9946 ,    3998   ,     3998  ,  778862.2360  ,   3705 ,    293 ,   930
-------------------------------------------------------------------------------------------------------------------
-> year = 2006
                                |    The KFS Multiply Imputed Data     |        The KFS Non-Imputed Data        |
                       Variable ,    Mean (mi) ,Min N (mi) ,Max N (mi) ,          Mean ,      N ,    N=. ,  N=.a
                         Assets ,  856213.5799 ,    3390   ,     3390  ,  776229.9407  ,   3190 ,    200 ,  1538
```

```
-------------------------------------------------------------------------------------------------------------------
-> year = 2007
                                |    The KFS Multiply Imputed Data     |        The KFS Non-Imputed Data        |
                       Variable ,    Mean (mi) ,Min N (mi) ,Max N (mi) ,          Mean ,      N ,    N=. ,  N=.a
                         Assets , 1284905.3187 ,    2915   ,     2915  , 1133588.7931  ,   2753 ,    162 ,  2013
-------------------------------------------------------------------------------------------------------------------
-> year = 2008
                                |    The KFS Multiply Imputed Data     |        The KFS Non-Imputed Data        |
                       Variable ,    Mean (mi) ,Min N (mi) ,Max N (mi) ,          Mean ,      N ,    N=. ,  N=.a
                         Assets ,  741097.7366 ,    2606   ,     2606  ,  764428.2122  ,   2427 ,    179 ,  2322
-------------------------------------------------------------------------------------------------------------------
-> year = 2009
                                |    The KFS Multiply Imputed Data     |        The KFS Non-Imputed Data        |
                       Variable ,    Mean (mi) ,Min N (mi) ,Max N (mi) ,          Mean ,      N ,    N=. ,  N=.a
                         Assets , 1278414.0034 ,    2408   ,     2408  , 1339926.3806  ,   2236 ,    172 ,  2520
-------------------------------------------------------------------------------------------------------------------
-> year = 2010
                                |    The KFS Multiply Imputed Data     |        The KFS Non-Imputed Data        |
                       Variable ,    Mean (mi) ,Min N (mi) ,Max N (mi) ,          Mean ,      N ,    N=. ,  N=.a
                         Assets , 1206118.5843 ,    2126   ,     2126  , 1029816.8026  ,   1984 ,    142 ,  2802
-------------------------------------------------------------------------------------------------------------------
-> year = 2011
                                |    The KFS Multiply Imputed Data     |        The KFS Non-Imputed Data        |
                       Variable ,    Mean (mi) ,Min N (mi) ,Max N (mi) ,          Mean ,      N ,    N=. ,  N=.a
                         Assets , 2959993.8255 ,    2007   ,     2007  , 3129300.2137  ,   1857 ,    150 ,  2921


Mean(mi)and Mean are weighted by pweight = cswgt_final

Mean(mi)and Mean are calculated using nonmissing vaules for each var in the varlist independently

 Mean (mi): the mean using the KFS multiply imputed data
 Min N (mi): the minimum number of observations among m=1,2,3,4,5
 Max N (mi): the maximum number of observations among m=1,2,3,4,5
 Mean: the mean using the KFS non-imputed data
```

5.1 Event History Analysis (EHA)

Event history analysis (aka, survival analysis, duration analysis, failure time analysis, and hazard analysis) is a collection of methods for analyzing time-to-event data. Time-to-event data reflect the observation of time from a specified origin (startup year) to a particular endpoint defined by the occurrence of a certain event of interest (exiting or firm closure). The Kauffman Firm Survey (KFS) data provide us records of an event of interest (firm exiting), as well as the type of event measured from a specified time origin (startup year). The endpoint consists of three mutually exclusive events of interest (permanently stopped operations, sold or merged, temporarily stopped operations) that create a competing risks situation.

Competing risks data come into view when the businesses under study can experience one, but not both, of m events of interest. For each business, we observe the time-to-event and the type of event. In addition, the occurrence of one type of event removes the business from risk of the other event types, i.e., businesses that close are no longer a target for a merger or acquisition. Thus, competing risks analysis is an extension of the ordinary survival analysis that implies one event of interest only. In this case, the standard survival analysis can be applied exactly as if we have one event of interest, treating type-m event of interest as event, and all other events of interest (m-1) as censoring events.

For the case of duration analysis, the standard approach of using Ordinary Least Squares (OLS) regression is not appropriate because it assumes that duration times are normally distributed, it may return negative predicted values, and it does not distinguish between "censored" and "uncensored" observations. These problems necessitate moving toward some kind of modeling strategy that is specifically designed for duration data. In the following subsections, we explore these modeling strategies.

Based on how we measure the time-to-event, event history analysis can utilize continuous time or discrete time models. The KFS provides us with the year in which the firm went out of business. Thus, our measurement of event time is discrete not because it is essentially discrete, but because the survey data are provided on a yearly basis and, therefore, the duration lengths are positive integers.

When modeling using duration analysis, censoring is usually an unavoidable problem. Censoring of an observation occurs when the information about its survival time is incomplete, i.e., when the annual follow-up surveys end before the event has occurred or when a business drops out from the study before an event occurs; thus, an event may not have occurred for some firms resulting in right-censored event times.

Right-censoring of data requires some assumptions about the reasons businesses censor. A non-informative (random) censoring assumes that businesses have a censoring time that is independent of the event of interest; thus, under the non-

informative censoring assumption, a censored observation has the same risk of closure as firms that have completed the follow-up survey. An informative censoring assumes that censoring time is possibly related to the event of interest.

For firms that were still in business at the end of the study period (2011), right-censoring is not problematic because it is a non-informative (random) censoring.

For firms that we do not have information for regarding the time of exit, but we do know when they last responded to a follow-up, we can use the last time they responded as the time at which the survival time is right-censored.

5.2 Event History Data Structures

EHA data can be either a single episode event history dataset or a multi-episode event history dataset.

All the KFS files (expect the original MRP data file) have two important variables for survival analysis. The duration variable, "duration," which measures time to occurrence of an event or the time at which the observation is censored, and the event variable "event" are coded 1 if the event occurred and 0 if the observation was censored.

5.2.1 Multi Episode - Longitudinal Data

For the KFS longitudinal data (panel), the multi-episode event data (stsplit command) gives us the same number of observations as does the longitudinal data in the long format (reshape command), yet the multi-episode event data and the longitudinal data in the long format are not the same.

```
*Multiple spells multi-episode event history format
* Use Longitudinal Data
use KFS8_L7_w1,clear
tab Duration event
* sts is not supported by svy
* Use all data based on the baseline weights
stset Duration [pweight=wgt_7_long] , failure(event==1)  id(mprid)
*stsplit with  every(#) option splits episodes into two or more episodes at the
implied time points since being at risk
stsplit split,every(1)
*Check what stsplit did?
list mprid Duration event _st _d _t _t0 split b1_bus_start_0  if mprid<10000498,
nol sepby(mprid)
```

```
      +-------------------------------------------------------------------------+
      |    mprid   Duration   event   _st   _d   _t   _t0   split   b1_bus~0 |
      |-------------------------------------------------------------------------|
   1. | 10000016          1       .     1    0    1     0       0          3 |
   2. | 10000016          2       1     1    1    2     1       1          3 |
      |-------------------------------------------------------------------------|
   3. | 10000090          1       .     1    0    1     0       0          3 |
   4. | 10000090          2       .     1    0    2     1       1          3 |
   5. | 10000090          3       .     1    0    3     2       2          3 |
   6. | 10000090          4       .     1    0    4     3       3          3 |
   7. | 10000090          5       .     1    0    5     4       4          3 |
   8. | 10000090          6       .     1    0    6     5       5          3 |
   9. | 10000090          7       .     1    0    7     6       6          3 |
  10. | 10000090          8       0     1    0    8     7       7          3 |
      |-------------------------------------------------------------------------|
  11. | 10000391          1       .     1    0    1     0       0          3 |
  12. | 10000391          2       .     1    0    2     1       1          3 |
  13. | 10000391          3       .     1    0    3     2       2          3 |
  14. | 10000391          4       .     1    0    4     3       3          3 |
  15. | 10000391          5       .     1    0    5     4       4          3 |
  16. | 10000391          6       1     1    1    6     5       5          3 |
      +-------------------------------------------------------------------------+

*Describe survival-time data
stdescribe

                                |-------------- per subject --------------|
                      unweighted   unweighted           unweighted
Category                 total        mean         min     median        max
-------------------------------------------------------------------------------
no. of subjects           3140
no. of records           18286    5.823567           1          8          8

(first) entry time                       0           0          0          0
(final) exit time                 5.823567           1          8          8

subjects with gap            0
time on gap if gap           0           .           .          .          .
time at risk             18286    5.823567           1          8          8

failures                  1496    .4764331           0          0          1
-------------------------------------------------------------------------------



gen year=2004+_t0
gen outcome=event

save L_multi_episode,replace
```

Where : _t0 is the time at which the observation “entered” the data, _t the time on which the observation “exited” the data (either by the event occurring or through censoring), _d is whether an event occurs at end of span, _st determines if time span should be included in the analysis and b1_bus_start_0 is a (non-time-varying) covariate.

5.2.2 Single Episode - Longitudinal Data

For a single episode event history dataset, a single line of data (one "record") is sufficient to describe an entire case, as long as we do not have repeated events or gaps (drop outs).

```
*Single-record-per-subject survival data
* Use Longitudinal Data
use KFS8_L7_w1,clear
tab Duration event

   Time in |
years from |  Indicator variable:
  start to | 1=uncensored(closure)
closure or |      0= censored
 censoring |         0          1 |     Total
-----------+----------------------+----------
         1 |         0        303 |       303
         2 |         0        283 |       283
         3 |         0        224 |       224
         4 |         0        238 |       238
         5 |         0        164 |       164
         6 |         0        153 |       153
         7 |        14        131 |       145
         8 |     1,630          0 |     1,630
-----------+----------------------+----------
     Total |     1,644      1,496 |     3,140

* sts is not supported by svy
* Use all data based on the baseline weights
stset Duration [pweight=wgt_7_long] , failure(event==1)  id(mprid)
*Check what stset did?
list mprid Duration event _t0 _t _d b1_bus_start_0  if mprid<10000498, nol
sepby(mprid)

      +-------------------------------------------------------------+
      |    mprid   Duration   event   _t0   _t   _d   b1_bus~0 |
      |-------------------------------------------------------------|
   1. | 10000016          2       1     0    2    1          3 |
      |-------------------------------------------------------------|
   2. | 10000090          8       0     0    8    0          3 |
      |-------------------------------------------------------------|
   3. | 10000391          6       1     0    6    1          3 |
      +-------------------------------------------------------------+
stdescribe


                              |-------------- per subject --------------|
                     unweighted   unweighted            unweighted
Category                  total         mean         min     median        max
-------------------------------------------------------------------------------
no. of subjects            3140
no. of records             3140            1           1          1          1
(first) entry time                         0           0          0          0
(final) exit time                   5.823567           1          8          8
subjects with gap             0
time on gap if gap            0            .           .          .          .
time at risk              18286     5.823567           1          8          8
failures                   1496     .4764331           0          0          1
-------------------------------------------------------------------------------
```

Where: _t0 is the time at which the observation "entered" the data, _t the time on which the observation "exited" the data (either by the event occurring or through censoring), _d is whether an event occurs at end of span and b1_bus_start_0 is a (non-time-varying) covariate.

For the KFS longitudinal data, using either a single episode event history dataset (non-time-varying) or a multi-episode event history dataset (time-varying) will give us the same results for non-time-varying survival analysis.

5.2.3 Multi Episode - Cross Sectional Data

Using the KFS cross-sectional data in survival analysis required more complicated data preparation. In multi-episode event data, stata assumes no gaps (drop outs). The problem with gaps is that firms that drop out at time t+1 and report back at time t+3 cannot be consider at risk of failure during time t+2. After all, we observe them at t+3 because they survive at time t+2.

```
*Multiple spells multi-episode event history format
use KFS8_CS_w1,clear
* sts is not supported by svy
* Use all data based on the baseline weights
replace Duration=Duration+1 if event==1
replace Duration=Duration+1 if event==0 & status_7=="Temporarily Stopped"
stset Duration [pweight=cswgt_final_0] , failure(event==1)  id(mprid)
gen Dur=Duration
gen outcome=event
*stsplit with  every(#) option splits episodes into two or more episodes at the
implied time points since being at risk
stsplit split,every(1)
*Check what stsplit did?
list mprid Duration event _st _d _t _t0 split b1_bus_start_0  if mprid==10000760
| mprid==10000799, nol sepby(mprid)
```

```
     +-------------------------------------------------------------------------+
     |    mprid   Duration   event   _st   _d   _t   _t0   split   b1_bus~0 |
     |-------------------------------------------------------------------------|
51.  | 10000760          1       .     1    0    1     0       0          3 |
52.  | 10000760          2       .     1    0    2     1       1          3 |
53.  | 10000760          3       .     1    0    3     2       2          3 |
54.  | 10000760          4       .     1    0    4     3       3          3 |
55.  | 10000760          5       .     1    0    5     4       4          3 |
56.  | 10000760          6       .     1    0    6     5       5          3 |
57.  | 10000760          7       .     1    0    7     6       6          3 |
58.  | 10000760          8       0     1    0    8     7       7          3 |
     |-------------------------------------------------------------------------|
59.  | 10000799          1       .     1    0    1     0       0          4 |
60.  | 10000799          2       .     1    0    2     1       1          4 |
61.  | 10000799          3       .     1    0    3     2       2          4 |
62.  | 10000799          4       1     1    1    4     3       3          4 |
     +-------------------------------------------------------------------------+
```

The problem with using stsplit with the KFS cross-sectional data is that stata assumes no gaps (drop outs). Take, for example, the firm with MPRID 10000799. This firm dropped out in 2006 and reappeared in 2007. For this firm, we should have a gap in 2006. To account for the gaps, we need to drop the years where firms did not respond.

```
*Multiple spells multi-episode event history format
use KFS8_CS_w1,clear
* sts is not supported by svy
* Use all data based on the baseline weights
replace Duration=Duration+1 if event==1
replace Duration=Duration+1 if event==0 & status_7=="Temporarily Stopped"
stset Duration [pweight=cswgt_final_0] , failure(event==1)  id(mprid)
gen Dur=Duration
gen outcome=event
*stsplit with  every(#) option splits episodes into two or more episodes at the
implied time points since being at risk
stsplit split,every(1)

forvalue i=1/7 {
drop if status_`i'=="Refusal"  & _t0==`i' & outcome==1
}

forvalue i=0/7 {
gen Dstatus_`i'=1
replace Dstatus_`i'=0 if status_`i'=="Refusal"
}

egen tot=rsum(Dstatus_*) if outcome==0
drop Dstatus*

forvalue i=1/7 {
drop if status_`i'=="Refusal"  & _t0==`i' & outcome==0 & tot<Dur
}

list mprid Duration event _st _d _t _t0 split b1_bus_start_0  if mprid==10000760
| mprid==10000799, nol sepby(mprid)
```

```
      +-------------------------------------------------------------------------+
      |    mprid   Duration   event   _st   _d   _t   _t0   split   b1_bus~0 |
      |-------------------------------------------------------------------------|
  51. | 10000760          1       .     1    0    1     0       0          3 |
  52. | 10000760          2       .     1    0    2     1       1          3 |
  53. | 10000760          3       .     1    0    3     2       2          3 |
  54. | 10000760          4       .     1    0    4     3       3          3 |
  55. | 10000760          5       .     1    0    5     4       4          3 |
  56. | 10000760          6       .     1    0    6     5       5          3 |
  57. | 10000760          7       .     1    0    7     6       6          3 |
  58. | 10000760          8       0     1    0    8     7       7          3 |
      |-------------------------------------------------------------------------|
  59. | 10000799          1       .     1    0    1     0       0          4 |
  60. | 10000799          2       .     1    0    2     1       1          4 |
  61. | 10000799          4       1     1    1    4     3       3          4 |
      +-------------------------------------------------------------------------+
```

```
*Describe survival-time data
stdescribe

                                |-------------- per subject --------------|
                      unweighted   unweighted            unweighted
Category                 total        mean         min     median        max
-------------------------------------------------------------------------------
no. of subjects           4928
no. of records           27270    5.533685           1          6          8

(first) entry time                       0           0          0          0
(final) exit time                 6.044643           1          7          8

subjects with gap         1292
time on gap if gap        2518    1.744976           1          1          6
time at risk             27270    5.533685           1          6          8

failures                  2190    .4443994           0          0          1
-------------------------------------------------------------------------------

gen year=2004+_t0

save CS_multi_episode,replace
```

Now, our multi-episode cross sectional data is ready for survival analysis because we dropped all gaps.

5.2.4 Multi Episode - Time Varying Covariates

The single episode event history format or multi-episode event history formats we created in example 5.2.1, 5.2.2, and 5.2.3 are useful for survival analysis methods that do not allow time-varying covariates (for example, non-parametric and parametric survival models).

For survival analysis methods that allow time-varying covariates (for example Cox competing risks and discrete-time methods), we need to have our data in a one record for each interval of time in which covariates are constant (long form).

Given that the hardest task in undertaking event history analysis is to setup the data in in the appropriate format, we will provide the Stata codes that will create the appropriate data format for survival analysis with time varying covariates for both the longitudinal and cross-sectional original data, as well as for the multiply imputed data.

While the cross-sectional analysis of survival data is useful in obtaining an accurate picture of survival rates, it is not practical for longitudinal analysis, especially when we consider time-varying covariates. This is because cross-sectional observations have a significant number of missing values (gaps), KFS did not collect data if the firm reported being sold, merged, temporarily stopped, and a high possibility of informative censoring; meanwhile, we do not have this problem for longitudinal observations.

It is a common practice in complex survey-based studies to consider businesses who respond to all follow-ups up to the last follow-up (attrition sample) (Chambers and Skinner, 2003; Lepkowski, 1989). Carrying longitudinal (panel) survival analysis will require limiting the analysis to the longitudinal observations in KFS and using the longitudinal weights.

The Stata code will create the following files:

Longitudinal_Long_Survival_Ready Notes: ➢ Already declared to be survival-time data (do not use stset or stset, clear)	Non-Imputed Data: Can be used for panel analysis and survival analysis methods that allow time-varying and nontime-varying covariates
Longitudinal_Long_MI_Survival_Ready Notes: ➢ Already declared to be survival-time data (do not use stset or stset, clear) ➢ Already declared to be multiple-imputation data (do not use mi set)	Imputed Data : Can be used for panel analysis and survival analysis methods that allow time-varying and nontime-varying covariates
Cross_Sectional_Long_Survival_Ready Notes: ➢ Already declared to be survival-time data (do not use stset or stset, clear)	Non-Imputed Data: Can be used for panel analysis and survival analysis methods that allow nontime-varying covariates
Cross_Sectional_Long_MI_Survival_Ready Notes: ➢ Already declared to be survival-time data (do not use stset or stset, clear) ➢ Already declared to be multiple-imputation data (do not use mi set)	Imputed Data : Can be used for panel analysis and survival analysis methods that allow nontime-varying covariates

5.2.4.1 Stata Code: Longitudinal_Long_Survival_Ready

```
use KFS8_L7_L1,clear
rename Duration dur
rename event eve
merge 1:1 mprid year using L_multi_episode, keepusing(Duration event _st _d _t
_t0  classf_0 classf_1 classf_2 classf_3 classf_4 classf_5 classf_6 classf_7)
*Competing Risks

gen Competing=_d
forvalue i=1/7{
replace Competing =classf_`i' if dur==`i' & _d==1
}

drop _merge

replace Competing =classf_7 if Duration==7 & _d==0 & classf_7==5
merge m:1 mprid   using Kfs8_enclave_14oct13, keepusing(a10_out_of_business_*)
keep if _merge==3
tab Competing if event<.

forvalue i=1/7{
replace Competing =a10_out_of_business_`i' if dur==`i' & _d==1 &
a10_out_of_business_`i'<. & Competing==4
}

drop classf* dur eve _merge a10_*
recode Competing (3=4)
recode Competing (5=3)
label variable Competing "CompetingRisks"
label define CompetingRisks 0 "0: No Event" 1 "1: Sold" 2 "2: Merged" 3 "3:
Temporarily Stopped" 4 "4: Out of Business"
label values Competing CompetingRisks
tab Competing if event<.

        CompetingRisks |      Freq.     Percent        Cum.
-----------------------+-----------------------------------
           0: No Event |      1,630       51.91       51.91
               1: Sold |        139        4.43       56.34
             2: Merged |         61        1.94       58.28
3: Temporarily Stopped |         14        0.45       58.73
    4: Out of Business |      1,296       41.27      100.00
-----------------------+-----------------------------------
                 Total |      3,140      100.00

*Describe survival-time data
stdescribe
```

```
                                |-------------- per subject --------------|
                     unweighted   unweighted            unweighted
Category                total        mean         min     median        max
-------------------------------------------------------------------------------
no. of subjects          3140
no. of records          18286    5.823567           1          8          8

(first) entry time             0           0          0          0          0
(final) exit time       5.823567           1          8          8

subjects with gap           0
time on gap if gap          0           .           .          .          .
time at risk            18286    5.823567           1          8          8

failures                 1496    .4764331           0          0          1
-------------------------------------------------------------------------------

save Longitudinal_Long_Survival_Ready,replace
```

5.2.4.2 Stata Code: Longitudinal_Long_MI_Survival_Ready

```
use Longitudinal_Long_MI_Long_L2,clear
rename Duration dur
rename event eve
merge m:1 mprid year using L_multi_episode, keepusing(Duration event _st _d _t
_t0 classf_0 classf_1 classf_2 classf_3 classf_4 classf_5 classf_6 classf_7)
*Competing Risks

gen Competing=_d
forvalue i=1/7{
replace Competing =classf_`i' if dur==`i' & _d==1
}

drop _merge

replace Competing =classf_7 if Duration==7 & _d==0 & classf_7==5
merge m:1 mprid   using Kfs8_enclave_14oct13, keepusing(a10_out_of_business_*)
keep if _merge==3
tab Competing if event<.

forvalue i=1/7{
replace Competing =a10_out_of_business_`i' if dur==`i' & _d==1 &
a10_out_of_business_`i'<. & Competing==4
}
tab Competing if event<.

drop classf* dur eve _merge a10_*
recode Competing (3=4)
recode Competing (5=3)
label variable Competing "CompetingRisks"
label define CompetingRisks 0 "0: No Event" 1 "1: Sold" 2 "2: Merged" 3 "3:
Temporarily Stopped" 4 "4: Out of Business"
label values Competing CompetingRisks
mi xeq :tab Competing if event<.

m=0 data:
-> tab Competing if event<.
```

```
        CompetingRisks |      Freq.     Percent        Cum.
-----------------------+-----------------------------------
           0: No Event |      1,630       51.91       51.91
               1: Sold |        139        4.43       56.34
             2: Merged |         61        1.94       58.28
3: Temporarily Stopped |         14        0.45       58.73
    4: Out of Business |      1,296       41.27      100.00
-----------------------+-----------------------------------
                 Total |      3,140      100.00

m=1 data:
-> tab Competing if event<.

        CompetingRisks |      Freq.     Percent        Cum.
-----------------------+-----------------------------------
           0: No Event |      1,630       51.91       51.91
               1: Sold |        139        4.43       56.34
             2: Merged |         61        1.94       58.28
3: Temporarily Stopped |         14        0.45       58.73
    4: Out of Business |      1,296       41.27      100.00
-----------------------+-----------------------------------
                 Total |      3,140      100.00

*Describe survival-time data
mi xeq : stdescribe

         failure _d:  event == 1
   analysis time _t:  Duration
                 id:  mprid
             weight:  [pweight=wgt_7_long]

                                |-------------- per subject --------------|
                      unweighted   unweighted            unweighted
Category                 total        mean         min     median        max
-------------------------------------------------------------------------------
no. of subjects           3140
no. of records           18286    5.823567           1          8          8

(first) entry time                       0           0          0          0
(final) exit time                 5.823567           1          8          8

subjects with gap            0
time on gap if gap           0           .           .          .          .
time at risk             18286    5.823567           1          8          8

failures                  1496    .4764331           0          0          1
-------------------------------------------------------------------------------
```

```
m=1 data:
-> stdescribe

         failure _d:  event == 1
   analysis time _t:  Duration
                 id:  mprid
             weight:  [pweight=wgt_7_long]

                                |-------------- per subject --------------|
                      unweighted   unweighted            unweighted
Category                 total        mean         min     median        max
-------------------------------------------------------------------------------
no. of subjects           3140
no. of records           18286    5.823567           1          8          8

(first) entry time                       0           0          0          0
(final) exit time                 5.823567           1          8          8

subjects with gap            0
time on gap if gap           0           .           .          .          .
time at risk             18286    5.823567           1          8          8

failures                  1496    .4764331           0          0          1
-------------------------------------------------------------------------------

save Longitudinal_Long_MI_Survival_Ready,replace
```

5.2.4.3 Stata Code: Cross_Sectional_Long_Survival_Ready

```
use KFS8_CS_L1,clear
rename Duration dur
rename event eve
keep if cswgt_final<.
drop if status=="Hard Missing Value"
drop *wgt*
merge 1:1 mprid year using CS_multi_episode, keepusing(Duration event
cswgt_final_0 _st _d _t _t0 classf_0 classf_1 classf_2 classf_3 classf_4
classf_5 classf_6 classf_7)
*Competing Risks

gen Competing=_d
forvalue i=1/7{
replace Competing =classf_`i' if dur==`i' & _d==1
}

drop _merge

replace Competing =classf_7 if Duration==8 & _d==0 & classf_7==5
merge m:1 mprid   using Kfs8_enclave_14oct13, keepusing(a10_out_of_business_*)
keep if _merge==3
tab Competing if event<.

forvalue i=1/7{
replace Competing =a10_out_of_business_`i' if dur==`i' & _d==1 &
a10_out_of_business_`i'<. & Competing==4
}
tab Competing if event<.


recode Competing (3=4)
recode Competing (5=3)
replace Competing=5 if Competing==0 & dur<8
drop classf* dur eve _merge a10_*

label variable Competing "CompetingRisks"
label define CompetingRisks 0 "0: No Event" 1 "1: Sold" 2 "2: Merged" 3 "3:
Temporarily Stopped" 4 "4: Out of Business" 5 "5: Drop-Outs"
label values Competing CompetingRisks
tab Competing if event<.

        CompetingRisks |      Freq.     Percent        Cum.
-----------------------+-----------------------------------
           0: No Event |      2,032       41.23       41.23
               1: Sold |        215        4.36       45.60
             2: Merged |         74        1.50       47.10
3: Temporarily Stopped |         30        0.61       47.71
    4: Out of Business |      1,901       38.58       86.28
          5: Drop-Outs |        676       13.72      100.00
-----------------------+-----------------------------------
                 Total |      4,928      100.00

save Cross_Sectional_Long_Survival_Ready,replace
```

5.2.4.4 Stata Code: Cross_Sectional_Long_MI_Survival_Ready

```
use Cross_Sectional_Long_MI_Long_L2,clear
rename Duration dur
rename event eve
keep if cswgt_final<.
drop if status=="Hard Missing Value"
drop *wgt*
merge m:1 mprid year using CS_multi_episode, keepusing(Duration event
cswgt_final_0 _st _d _t _t0 classf_0 classf_1 classf_2 classf_3 classf_4
classf_5 classf_6 classf_7)
*Competing Risks

gen Competing=_d
forvalue i=1/7{
replace Competing =classf_`i' if dur==`i' & _d==1
}

drop _merge

replace Competing =classf_7 if Duration==8 & _d==0 & classf_7==5
merge m:1 mprid   using Kfs8_enclave_14oct13, keepusing(a10_out_of_business_*)
keep if _merge==3
tab Competing if event<.

forvalue i=1/7{
replace Competing =a10_out_of_business_`i' if dur==`i' & _d==1 &
a10_out_of_business_`i'<. & Competing==4
}

bysort _mi_m: tab Competing if event<.


recode Competing (3=4)
recode Competing (5=3)
replace Competing=5 if Competing==0 & dur<8
drop classf* dur eve _merge a10_*

label variable Competing "CompetingRisks"
label define CompetingRisks 0 "0: No Event" 1 "1: Sold" 2 "2: Merged" 3 "3:
Temporarily Stopped" 4 "4: Out of Business" 5 "5: Drop-Outs"
label values Competing CompetingRisks

mi xeq :tab Competing if event<.

m=0 data:
-> tab Competing if event<.

        CompetingRisks |      Freq.     Percent        Cum.
-----------------------+-----------------------------------
           0: No Event |      2,032       41.23       41.23
               1: Sold |        215        4.36       45.60
             2: Merged |         74        1.50       47.10
3: Temporarily Stopped |         30        0.61       47.71
    4: Out of Business |      1,901       38.58       86.28
          5: Drop-Outs |        676       13.72      100.00
-----------------------+-----------------------------------
                 Total |      4,928      100.00
```

```
m=1 data:
-> tab Competing if event<.

        CompetingRisks |      Freq.     Percent        Cum.
-----------------------+-----------------------------------
           0: No Event |      2,032       41.23       41.23
               1: Sold |        215        4.36       45.60
             2: Merged |         74        1.50       47.10
3: Temporarily Stopped |         30        0.61       47.71
    4: Out of Business |      1,901       38.58       86.28
          5: Drop-Outs |        676       13.72      100.00
-----------------------+-----------------------------------
                 Total |      4,928      100.00

mi xeq :stdescribe

m=0 data:
-> stdescribe

         failure _d:  event == 1
   analysis time _t:  Duration
                 id:  mprid
             weight:  [pweight=cswgt_final_0]

                                |-------------- per subject --------------|
                      unweighted   unweighted            unweighted
Category                 total        mean         min     median        max
-------------------------------------------------------------------------------
no. of subjects           4928
no. of records           27270    5.533685           1          6          8
(first) entry time                       0           0          0          0
(final) exit time                 6.044643           1          7          8
subjects with gap         1292
time on gap if gap        2518    1.744976           1          1          6
time at risk             27270    5.533685           1          6          8
failures                  2190    .4443994           0          0          1
-------------------------------------------------------------------------------

m=1 data:
-> stdescribe

         failure _d:  event == 1
   analysis time _t:  Duration
                 id:  mprid
             weight:  [pweight=cswgt_final_0]

                                |-------------- per subject --------------|
                      unweighted   unweighted            unweighted
Category                 total        mean         min     median        max
-------------------------------------------------------------------------------
no. of subjects           4928
no. of records           27270    5.533685           1          6          8
(first) entry time                       0           0          0          0
(final) exit time                 6.044643           1          7          8
subjects with gap         1292
time on gap if gap        2518    1.744976           1          1          6
time at risk             27270    5.533685           1          6          8
failures                  2190    .4443994           0          0          1
-------------------------------------------------------------------------------

save Cross_Sectional_Long_MI_Survival_Ready,replace
```

For Stata, all EHA data formats are equal; the following example shows how Stata will give the same results regardless of the data format.

```
*Multiple spells multi-episode event history format
*Cross Sectional
use CS_multi_episode,clear
sts list

           Beg.          Net          Survivor
  Time    Total   Fail   Lost           Function
-------------------------------------------------
     1  73278.4      0   8424              1.0000
     2  64854.3   4697   2994              0.9276
     3  57163.2   5982   1174              0.8305
     4  50006.7   5347 -226.6              0.7417
     5  44886.3   5759  -1096              0.6465
     6  40222.8   4350  146.9              0.5766
     7  35725.7   3900  -1127              0.5137
     8  32952.4   3985  2.9e+04            0.4516
-------------------------------------------------

use Cross_Sectional_Long_Survival_Ready,clear
sts list

           Beg.          Net          Survivor
  Time    Total   Fail   Lost           Function
-------------------------------------------------
     1  73278.4      0   8424              1.0000
     2  64854.3   4697   2994              0.9276
     3  57163.2   5982   1174              0.8305
     4  50006.7   5347 -226.6              0.7417
     5  44886.3   5759  -1096              0.6465
     6  40222.8   4350  146.9              0.5766
     7  35725.7   3900  -1127              0.5137
     8  32952.4   3985  2.9e+04            0.4516
-------------------------------------------------

use Cross_Sectional_Long_MI_Survival_Ready,clear
mi xeq 0: sts list

           Beg.          Net          Survivor
  Time    Total   Fail   Lost           Function
-------------------------------------------------
     1  73278.4      0   8424              1.0000
     2  64854.3   4697   2994              0.9276
     3  57163.2   5982   1174              0.8305
     4  50006.7   5347 -226.6              0.7417
     5  44886.3   5759  -1096              0.6465
     6  40222.8   4350  146.9              0.5766
     7  35725.7   3900  -1127              0.5137
     8  32952.4   3985  2.9e+04            0.4516
-------------------------------------------------
```

```
*OR

mi extract 0
sts list

          Beg.          Net          Survivor
  Time    Total   Fail   Lost           Function
-------------------------------------------------
     1  73278.4      0   8424             1.0000
     2  64854.3   4697   2994             0.9276
     3  57163.2   5982   1174             0.8305
     4  50006.7   5347 -226.6             0.7417
     5  44886.3   5759  -1096             0.6465
     6  40222.8   4350  146.9             0.5766
     7  35725.7   3900  -1127             0.5137
     8  32952.4   3985  2.9e+04           0.4516
-------------------------------------------------

*Single-record-per-subject survival data
* Use Longitudinal Data
use KFS8_L7_w1,clear
stset Duration [pweight=wgt_7_long] , failure(event==1)  id(mprid)
sts list


          Beg.          Net          Survivor
  Time    Total   Fail   Lost           Function
-------------------------------------------------
     1  73278.4   7952      0             0.8915
     2  65326.3   8158      0             0.7802
     3  57168.4   5503      0             0.7051
     4  51665.6   5644      0             0.6280
     5  46021.5   4150      0             0.5714
     6    41872   3601      0             0.5223
     7  38271.3   3062  317.3             0.4805
     8    34892      0  3.5e+04           0.4805
-------------------------------------------------

*Multiple spells multi-episode event history format
stsplit split,every(1)
sts list

          Beg.          Net          Survivor
  Time    Total   Fail   Lost           Function
-------------------------------------------------
     1  73278.4   7952      0             0.8915
     2  65326.3   8158      0             0.7802
     3  57168.4   5503      0             0.7051
     4  51665.6   5644      0             0.6280
     5  46021.5   4150      0             0.5714
     6    41872   3601      0             0.5223
     7  38271.3   3062  317.3             0.4805
     8    34892      0  3.5e+04           0.4805
-------------------------------------------------
```

```
use  Longitudinal_Long_Survival_Ready,clear
sts list

           Beg.          Net          Survivor
  Time    Total   Fail   Lost           Function
-------------------------------------------------
     1  73278.4   7952      0             0.8915
     2  65326.3   8158      0             0.7802
     3  57168.4   5503      0             0.7051
     4  51665.6   5644      0             0.6280
     5  46021.5   4150      0             0.5714
     6    41872   3601      0             0.5223
     7  38271.3   3062  317.3             0.4805
     8    34892      0  3.5e+04           0.4805
-------------------------------------------------


use  Longitudinal_Long_MI_Survival_Ready,clear
mi xeq 0:sts list

           Beg.          Net          Survivor
  Time    Total   Fail   Lost           Function
-------------------------------------------------
     1  73278.4   7952      0             0.8915
     2  65326.3   8158      0             0.7802
     3  57168.4   5503      0             0.7051
     4  51665.6   5644      0             0.6280
     5  46021.5   4150      0             0.5714
     6    41872   3601      0             0.5223
     7  38271.3   3062  317.3             0.4805
     8    34892      0  3.5e+04           0.4805
-------------------------------------------------
```

5.2.5 The Construction of The "Duration" and "event" Variables

To understand how the variables "duration" and "event" were constructed, the following table shows some of the patterns of right-censored cases in KFS using the cross-sectional surveys. Keep in mind that censoring is defined by the researcher; in other words, one researcher's censoring may be another researcher's event of interest. In the following table, we define the event as closure (out of business, merged, or sold), and we assume the event happing in the beginning of the year.

Beginning of Survey Observation	Episode ↓								End of Survey Observation	n	Duration	event
	→	→	→	→	→	→	→	→		1587	8	0
	→	e								303	1	1
	→	→	e							275	2	1
	→	→	→	→	e					199	4	1
	→	→	→	e						177	3	1
	→	→	→	→	→	e				141	5	1
	→	→	→	→	→	→	e			136	6	1
	→	d	d	d	d	d	d	d		124	1	0
	→	→	→	→	→	→	→	e		115	7	1
	→	d	e							85	2	1
	→	→	d	d	d	d	d	d		74	2	0
	→	→	→	d	d	d	d	d		66	3	0
	→	→	d	e						61	3	1
	→	→	→	→	d	d	d	d		61	4	0
	→	→	→	→	→	→	→	d		58	7	0
	→	→	→	→	→	→	d	d		53	6	0
	→	→	d	d	→	→	e			1	6	1
	→	→	d	t	t	d	e			1	6	1
	→	→	d	t	d	d	d	t		1	7	0
	→	→	→	→	→	t	t	→		1	8	0
	→	→	→	d	t	d	d	d		1	5	0
	→	t	t	t	t	t	e	→		1	6	1
	2004	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**				

e : Event d: Refusal t: temporarily stopped

5.3 Nonparametric Analysis : Kaplan-Meier and Life Tables

Because parametric analysis of duration data requires the assumption of an underlying distribution, we first subject the KFS data to a nonparametric analysis of duration before moving on to the semi-parametric and parametric analyses. The nonparametric methods are useful for the specification analysis of parametric models, displaying data on duration, and preliminary analyses.

The nonparametric duration models enable the calculation of nonparametric estimates of the survival and hazard functions without assuming an underlying distribution or how independent variables change survival experiences. Thus, it avoids the potentially large errors brought about by making incorrect assumptions about the distribution.

The Kaplan-Meier product-limit method is more suitable for duration data with a precise measure of event time. On the other hand, the life-table method is more suitable for duration data with a crude measure of event time.

Let $t_1 < t_2 <..< t_k$ represent the unique event times, S(t) the probability of surviving to t_i or beyond (the survival function), d_i the number of deaths at time t_i and n_i the number "at risk" just prior to time t_i. Then the Kaplan–Meier estimator is the nonparametric maximum likelihood estimate of

$$\hat{S}(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i}$$

The life-table method enables the calculation of nonparametric estimates of the survival and hazard functions without assuming an underlying distribution or how independent variables change survival experiences; thus, it avoids the potentially large errors brought about by making incorrect assumptions about distribution. Unlike the Kaplan-Meier product-limit method, the life-table survival estimates can still be calculated even if the exact survival or censoring times are not known for each firm, as long as the number of firms who exit or are censored within each time interval is known. In addition, the nonparametric duration models allow comparison of two or more groups of survival data (Stratified Test) and test for the association of survival time with time-constant covariates, without producing estimates of parameters.

The life tables are constructed by partitioning event time into intervals and then counting for each time-interval of the following: the number of surviving firms at the start of the interval, the number of failures during the interval (uncensored observations), and the number of firms who are lost for other reasons (i.e., mergers) during the interval (censored observations). Let $t_1 < t_2 < \ldots < t_k$ represent the unique event times, n_j the number of observations that fall into each of the time-intervals $\left[t_{i-1}, t_i\right)$, $i = 1,2,..3,k+1$, where $t_0 = 0$, d_j is the number of firms exiting at time t_j that

fall into each of the time-intervals $\left[t_{i-1}, t_i\right]$, c_j is the number of censored observations that fall into each of the time intervals $\left[t_{i-1}, t_i\right]$, $r_j = n_j - \frac{c_j}{2}$ is the size of the risk set – firms at risk are those that have not experienced an event nor have they been censored prior to time t_j - and $q_j = \frac{d_j}{r_j}$ is the proportion of firms exited in the risk set . The probability of surviving to t_i or beyond (the survival function) is given by the following expression:

$$S(t_i) = \Pr(T > t) = \prod_{j=1}^{i-1} \left(1 - q_j\right)$$

where $S(t_1) = 1$ and T denotes a non-negative random variable representing the survival time.

The hazard rate is the risk of exit for a firm that has survived to the beginning of the respective interval:

$$\lambda\left(t_{im}\right) = \frac{2q_j}{w(2 - q_j)}$$

where for the i[th] interval, t_{im} is the midpoint and w is the interval width.

The hazard function is a time to failure function that gives the instantaneous probability of the failure, given that it has not yet occurred. It is worth noting that because the Kaplan-Meier product-limit method uses the actual survival time of cases versus categorizing time into equal intervals, the survival rates are identical to the ones produced using life-tables method.

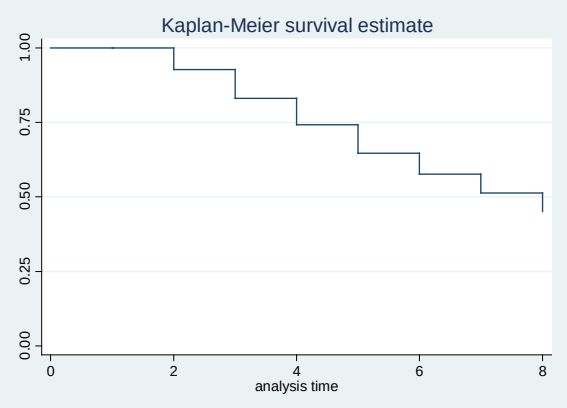
The measure advantage of the nonparametric methods is their ability to estimate survival, failure, and hazard rates.

Examples 5.1 Kaplan-Meier

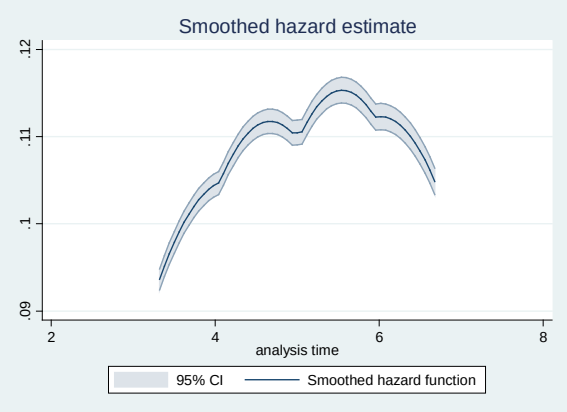
Each file we created has two variables: duration, which measure time to occurrence of an event or the time at which the observation is censored, and the variable event, which takes the value 1 for an event and 0 otherwise (the censored cases). All wide format files are single-record-per-subject survival data, which allows us to conduct duration analysis using wide format data.

If the censoring of businesses that did not respond (drop-out or lost to follow-up) is due to non-informative censoring and the number of these observations is high, then ignoring these observations will lead to loss of efficiency and power due to fewer observations. Nonetheless, standard practice in survival analysis does not ignore those right-censored observations because they could contain information about firm survival. To consider businesses that drop out, we need to use the baseline weights (cswgt_final_0).

```
*Kaplan-Meier
*Multiple spells multi-episode event history format
use CS_multi_episode,clear
* List the survivor or cumulative hazard function
sts list,  survival

         failure _d:  event == 1
   analysis time _t:  Duration
                 id:  mprid
             weight:  [pweight=cswgt_final_0]

           Beg.          Net          Survivor
  Time    Total   Fail   Lost           Function
-------------------------------------------------
     1  73278.4      0   8424             1.0000
     2  64854.3   4697   2994             0.9276
     3  57163.2   5982   1174             0.8305
     4  50006.7   5347 -226.6             0.7417
     5  44886.3   5759  -1096             0.6465
     6  40222.8   4350  146.9             0.5766
     7  35725.7   3900  -1127             0.5137
     8  32952.4   3985  2.9e+04           0.4516
-------------------------------------------------

sts graph, survival tmax(8)
```

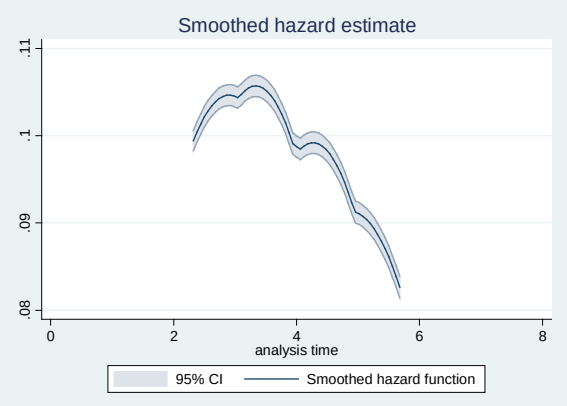


```
sts graph, hazard ci
```



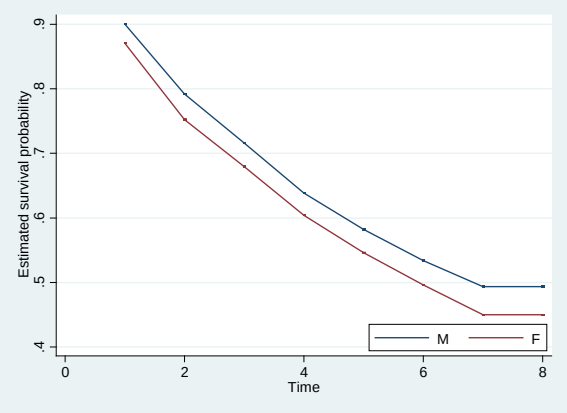

```
*Does survival differ by   gender?
sts list,  survival by(PO_gender_0)

           Beg.          Net          Survivor
  Time    Total   Fail   Lost           Function
-------------------------------------------------
PO_gender_0=0
     1  22590.4      0   2284             1.0000
     2  20305.9   1767   1301             0.9130
     3    17238   1917   43.3             0.8115
     4  15278.2   1545 -59.88             0.7294
     5  13793.2   1690 -46.01             0.6401
     6  12149.2   1393 -194.4             0.5667
     7  10951.1   1283 -350.4             0.5003
     8  10018.3   1258   8760             0.4375
PO_gender_0=1
     1  50648.4      0   6117             1.0000
     2  44531.6   2930   1693             0.9342
     3  39908.4   4066   1131             0.8390
     4  34711.8   3802 -166.7             0.7471
     5  31076.3   4069  -1050             0.6493
     6  28056.8   2958  324.6             0.5808
     7  24774.6   2617 -776.3             0.5195
     8  22934.1   2727  2.0e+04           0.4577
-------------------------------------------------

*sts test tests the equality of survivor functions across two or more groups
* Since our data are pweighted, cox test is the only possibility.
 sts test PO_gender_0 ,      cox

Cox regression-based test for equality of survival curves

            |   Events         Events       Relative
PO_gender_0 |  observed       expected       hazard
------------+-------------------------------------
0           |  10851.63       10401.58       1.0437
1           |  23168.75       23618.81       0.9813
------------+-------------------------------------
Total       |  34020.39       34020.39       1.0000

             Wald chi2(1) =       1.34
                  Pr>chi2 =     0.2470
```

```
* Use Longitudinal Data

use KFS8_L7_w1,clear
* Declare data to be survival-time data
* Use all data based on the baseline weights
* sts is not supported by svy
stset Duration [pweight=wgt_7_long] , failure(event==1)
* List the survivor or cumulative hazard function
sts list,  survival


          Beg.          Net          Survivor
  Time    Total   Fail   Lost           Function
-------------------------------------------------
     1  73278.4   7952      0             0.8915
     2  65326.3   8158      0             0.7802
     3  57168.4   5503      0             0.7051
     4  51665.6   5644      0             0.6280
     5  46021.5   4150      0             0.5714
     6    41872   3601      0             0.5223
     7  38271.3   3062  317.3             0.4805
     8    34892      0  3.5e+04           0.4805
-------------------------------------------------
* hazard rate
stptime ,at(1(1)8)

Cohort     |  person-time   failures        rate   [95% Conf. Interval]
-----------+-------------------------------------------------------------
(0   -   1]|    73278.441    7952.18   .10852003   .0959773    .1232116
(1   -   2]|    65326.262    8157.86   .12487871   .1103941    .1418692
(2   -   3]|    57168.403    5502.83   .09625643   .0833742    .1117589
(3   -   4]|    51665.577     5644.1   .10924296   .0947397     .126675
(4   -   5]|    46021.477    4149.52   .09016494    .076156     .107602
(5   -   6]|    41871.953    3600.68   .08599255   .0723237    .1030876
(6   -   7]|    38271.277    3061.94   .08000615   .0663536    .0973987
     >   7 |    34892.039          0           0          .           .
-----------+-------------------------------------------------------------
     total |    408495.43    38069.1   .09319346   .0880935     .098673


sts graph, hazard ci
```

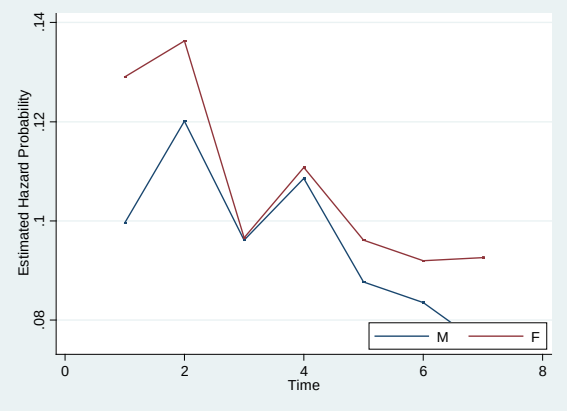


```
*Does survival differ by   gender?
sts list,  survival by(PO_gender_0)


           Beg.          Net          Survivor
  Time    Total   Fail   Lost            Function
-------------------------------------------------
PO_gender_0=0
     1  22206.1   2865      0              0.8710
     2  19340.9   2636      0              0.7523
     3  16705.2   1614      0              0.6796
     4  15091.1   1672      0              0.6043
     5  13418.9   1290      0              0.5462
     6  12128.6   1116      0              0.4959
     7  11012.8   1020    124              0.4500
     8  9868.97      0   9869              0.4500
PO_gender_0=1
     1  51072.3   5087      0              0.9004
     2  45985.4   5522      0              0.7923
     3  40463.2   3889      0              0.7161
     4  36574.5   3972      0              0.6384
     5  32602.5   2859      0              0.5824
     6  29743.3   2485      0              0.5337
     7  27258.5   2042  193.3              0.4937
     8  25023.1      0  2.5e+04            0.4937
-------------------------------------------------

stptime ,at(1(1)8) by(PO_gender_0)

PO_gend~0  |  person-time   failures        rate   [95% Conf. Interval]
-----------+-------------------------------------------------------------
0          |
(0   -   1]|    22206.112    2865.21   .12902821   .1046017    .1611529
(1   -   2]|    19340.897    2635.73   .13627769   .1092393    .1724256
(2   -   3]|    16705.164    1614.08    .0966214   .0735635    .1295946
(3   -   4]|    15091.088    1672.14   .11080344   .0854434    .1464177
(4   -   5]|    13418.944     1290.3   .09615492   .0707055    .1343472
(5   -   6]|    12128.646    1115.84   .09200043   .0676996    .1283101
(6   -   7]|    11012.805    1019.85   .09260588   .0658467    .1346459
(7   -   8]|    9868.9653          0           0           .           .
-----------+-------------------------------------------------------------
1          |
(0   -   1]|    51072.328    5086.96   .09960312   .0857377    .1164286
(1   -   2]|    45985.365    5522.13   .12008443   .1036269    .1400131
(2   -   3]|    40463.239    3888.75   .09610575   .0812573    .1145534
(3   -   4]|    36574.489    3971.96   .10859909   .0916847    .1296734
(4   -   5]|    32602.533    2859.23   .08769952   .0717561    .1084065
(5   -   6]|    29743.307    2484.84   .08354266   .0678262    .1041545
(6   -   7]|    27258.471    2042.09   .07491569   .0600609    .0947093
(7   -   8]|    25023.074          0           0           .           .
-----------+-------------------------------------------------------------
     total |    408495.43    38069.1   .09319346   .0880935     .098673

*sts test tests the equality of survivor functions across two or more groups
* Since our data are pweighted, cox test is the only possibility.
 sts test PO_gender_0 ,      cox
```

```
Cox regression-based test for equality of survival curves
             |   Events         Events       Relative
PO_gender_0 |  observed       expected       hazard
------------+-------------------------------------
0           |  12213.16       11215.63       1.0907
1           |  25855.94       26853.47       0.9644
------------+-------------------------------------
Total       |  38069.10       38069.10       1.0000

             Wald chi2(1) =       3.99
                  Pr>chi2 =     0.0458

sts generate s = s, by(PO_gender_0)

sort _t
graph twoway (line s _t if PO_gender_0 == 1, sort connect(l)) (line s _t if
PO_gender_0 == 0, sort connect(l)), ///
       legend(pos(5) ring(0) lab(1 "M") lab(2 "F")) ///
       xtitle("Time") ytitle("Estimated survival probability")
```




```
sts generate h = h, by(PO_gender_0)
sort _t
graph twoway (line h _t if PO_gender_0 == 1, sort connect(l)) (line h _t if
PO_gender_0 == 0, sort connect(l)), ///
       legend(pos(5) ring(0) lab(1 "M") lab(2 "F")) ///
       xtitle("Time") ytitle("Estimated Hazard Probability")
```



When using the full sample with the baseline weights (cswgt_final_0), it is important to understand how Stata deals with businesses that are censored at the same time that others fail. The STATA convention is that failures occur before censoring, thus censored observations are in risk set at time t. If we were to assume that censoring occurred before failures, we need to modify our duration variable to reflect such ordering.

Examples 5.2 Life tables

```
*Life tables
* Use Longitudinal Data
use KFS8_L7_w1,clear
* Life tables for survival data
* ltable is not supported by svy nor pweight.
* fweights are allowed : may not use noninteger
gen lfweights= round(wgt_7_long)
/*noadjust suppresses the actuarial adjustment for deaths and censored
observations. The default is to consider the adjusted number at risk at the
start of the interval as the total at the start minus (the number dead or
censored)/2. If noadjust is specified, the number at risk is simply the total at
the start, corresponding to the standard Kaplan–Meier assumption.*/
ltable  Duration event [fw=lfweights], survival  hazard i (1 2 3 4 5 6 7 8)
noadjust
```

Interval		Beg. Total	Deaths	Lost	Survival	Std. Error	[95% Conf.	Int.]
1	2	73258	7955	0	0.8914	0.0011	0.8891	0.8936
2	3	65303	8157	0	0.7801	0.0015	0.7770	0.7830
3	4	57146	5504	0	0.7049	0.0017	0.7016	0.7082
4	5	51642	5640	0	0.6279	0.0018	0.6244	0.6314
5	6	46002	4147	0	0.5713	0.0018	0.5677	0.5749
6	7	41855	3600	0	0.5222	0.0018	0.5186	0.5258
7	8	38255	3067	319	0.4803	0.0018	0.4767	0.4839
8	.	34869	0	34869	0.4803	0.0018	0.4767	0.4839

Interval		Beg. Total	Cum. Failure	Std. Error	Hazard	Std. Error	[95% Conf.	Int.]
1	2	73258	0.1086	0.0011	0.1086	0.0012	0.1062	0.1110
2	3	65303	0.2199	0.0015	0.1249	0.0014	0.1222	0.1276
3	4	57146	0.2951	0.0017	0.0963	0.0013	0.0938	0.0989
4	5	51642	0.3721	0.0018	0.1092	0.0015	0.1064	0.1121
5	6	46002	0.4287	0.0018	0.0901	0.0014	0.0874	0.0929
6	7	41855	0.4778	0.0018	0.0860	0.0014	0.0832	0.0888
7	8	38255	0.5197	0.0018	0.0802	0.0014	0.0774	0.0830
8	.	34869	0.5197	0.0018	.	.	.	.

```
* estimate separate functions for each group formed by varlist
ltable  Duration event [fw=lfweights], survival i (1 2 3 4 5 6 7 8) noadjust
by(PO_gender_0)

                Beg.                                        Std.
   Interval     Total   Deaths   Lost    Survival    Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
PO_gender_0 = 0
    1     2     22209     2870      0     0.8708    0.0023     0.8663    0.8751
    2     3     19339     2632      0     0.7523    0.0029     0.7465    0.7579
    3     4     16707     1614      0     0.6796    0.0031     0.6734    0.6857
    4     5     15093     1675      0     0.6042    0.0033     0.5977    0.6106
    5     6     13418     1289      0     0.5461    0.0033     0.5396    0.5527
    6     7     12129     1116      0     0.4959    0.0034     0.4893    0.5024
    7     8     11013     1022    125     0.4499    0.0033     0.4433    0.4564
    8     .      9866        0   9866     0.4499    0.0033     0.4433    0.4564
PO_gender_0 = 1
    1     2     51049     5085      0     0.9004    0.0013     0.8978    0.9030
    2     3     45964     5525      0     0.7922    0.0018     0.7886    0.7957
    3     4     40439     3890      0     0.7160    0.0020     0.7120    0.7198
    4     5     36549     3965      0     0.6383    0.0021     0.6341    0.6424
    5     6     32584     2858      0     0.5823    0.0022     0.5780    0.5866
    6     7     29726     2484      0     0.5336    0.0022     0.5293    0.5380
    7     8     27242     2045    194     0.4936    0.0022     0.4892    0.4979
    8     .     25003        0  25003     0.4936    0.0022     0.4892    0.4979
-------------------------------------------------------------------------------

use Longitudinal_Long_Survival_Ready,clear
gen lfweights= round(wgt_7_long)
ltable  Duration event [fw=lfweights], survival  hazard i (1 2 3 4 5 6 7 8)
noadjust

                Beg.                                        Std.
   Interval     Total   Deaths   Lost    Survival    Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
    1     2     73258     7955      0     0.8914    0.0011     0.8891    0.8936
    2     3     65303     8157      0     0.7801    0.0015     0.7770    0.7830
    3     4     57146     5504      0     0.7049    0.0017     0.7016    0.7082
    4     5     51642     5640      0     0.6279    0.0018     0.6244    0.6314
    5     6     46002     4147      0     0.5713    0.0018     0.5677    0.5749
    6     7     41855     3600      0     0.5222    0.0018     0.5186    0.5258
    7     8     38255     3067    319     0.4803    0.0018     0.4767    0.4839
    8     .     34869        0  34869     0.4803    0.0018     0.4767    0.4839
-------------------------------------------------------------------------------



                Beg.     Cum.     Std.                Std.
   Interval     Total   Failure   Error    Hazard     Error    [95% Conf. Int.]
-------------------------------------------------------------------------------
    1     2     73258    0.1086  0.0011    0.1086    0.0012    0.1062    0.1110
    2     3     65303    0.2199  0.0015    0.1249    0.0014    0.1222    0.1276
    3     4     57146    0.2951  0.0017    0.0963    0.0013    0.0938    0.0989
    4     5     51642    0.3721  0.0018    0.1092    0.0015    0.1064    0.1121
    5     6     46002    0.4287  0.0018    0.0901    0.0014    0.0874    0.0929
    6     7     41855    0.4778  0.0018    0.0860    0.0014    0.0832    0.0888
    7     8     38255    0.5197  0.0018    0.0802    0.0014    0.0774    0.0830
    8     .     34869    0.5197  0.0018         .         .         .         .
-------------------------------------------------------------------------------
```

Examples 5.3 Survival, Failure and Hazard Rates Using Logit Regression

We can use a discrete time logistic model to estimate survival rates as well as hazard rates.

```
* We can use discrete time logistic model to estimate  Survival & Hazard

use Longitudinal_Long_Survival_Ready,clear
tab year , gen(D)
logit _d i.D1 i.D2 i.D3 i.D4 i.D5 i.D6 i.D7  [pweight=wgt_7_long] if year<2011,
nocons
predict Hazard if year<2011
*preserve and restore deal with the programming problem where the user's data
must be changed to achieve the desired result but, when the program concludes,
the programmer wishes to undo the change done to the data.

preserve
gen Cyear=year+1 if year<2011
collapse (mean) Hazard [pweight=wgt_7_long], by(Cyear)
gen Survival = 1
replace Survival = (1 - Hazard)*Survival if Cyear == 2005
replace Survival = (1 - Hazard)*Survival[_n-1] if Cyear > 2005
format Survival Hazard %6.4f
gen Interval0=Cyear-2004
gen Interval1=Interval0+1
list Cyear Interval0 Interval1  Survival Hazard if Cyear<., separator(7)
restore
     +-------------------------------------------------+
     | Cyear   Interv~0   Interv~1   Survival   Hazard |
     |-------------------------------------------------|
  1. |  2005          1          2     0.8915   0.1085 |
  2. |  2006          2          3     0.7802   0.1249 |
  3. |  2007          3          4     0.7051   0.0963 |
  4. |  2008          4          5     0.6280   0.1092 |
  5. |  2009          5          6     0.5714   0.0902 |
  6. |  2010          6          7     0.5223   0.0860 |
  7. |  2011          7          8     0.4805   0.0800 |
     +-------------------------------------------------+

*Does survival differ by gender?

use Longitudinal_Long_Survival_Ready,clear
* Declare data to be panel data
xtset mprid year
*Creat nontime-varying covariates
bysort mprid (year): gen PO_gender_0=PO_gender[1]
*Report whether variables vary over time
stvary  PO_gender_0  PO_gender
tab year , gen(D)
logit _d i.D1 i.D2 i.D3 i.D4 i.D5 i.D6 i.D7 PO_gender_0 [pweight=wgt_7_long] if
year<2011, nocons
preserve
predict Hazard if year<2011
gen Cyear=year+1 if year<2011
collapse (mean) Hazard [pweight=wgt_7_long], by(Cyear PO_gender_0)
sort PO_gender_0 Cyear
gen Survival = 1
by PO_gender_0:replace Survival = (1 - Hazard)*Survival if Cyear == 2005
by PO_gender_0:replace Survival = (1 - Hazard)*Survival[_n-1] if Cyear > 2005
format Survival Hazard %6.4f
```

```
gen Interval0=Cyear-2004
gen Interval1=Interval0+1
list PO_gender_0 Cyear Interval0 Interval1  Survival Hazard if Cyear<.,
separator(7)
restore
```

```
     +-------------------------------------------------------------+
     | PO_gen~0   Cyear   Interv~0   Interv~1   Survival   Hazard |
     |-------------------------------------------------------------|
  1. |        0    2005          1          2     0.8820   0.1180 |
  2. |        0    2006          2          3     0.7623   0.1357 |
  3. |        0    2007          3          4     0.6823   0.1049 |
  4. |        0    2008          4          5     0.6012   0.1189 |
  5. |        0    2009          5          6     0.5420   0.0984 |
  6. |        0    2010          6          7     0.4912   0.0939 |
  7. |        0    2011          7          8     0.4482   0.0874 |
     |-------------------------------------------------------------|
  9. |        1    2005          1          2     0.8956   0.1044 |
 10. |        1    2006          2          3     0.7878   0.1203 |
 11. |        1    2007          3          4     0.7148   0.0927 |
 12. |        1    2008          4          5     0.6396   0.1052 |
 13. |        1    2009          5          6     0.5841   0.0868 |
 14. |        1    2010          6          7     0.5357   0.0828 |
 15. |        1    2011          7          8     0.4945   0.0770 |
     +-------------------------------------------------------------+
```

```
use Cross_Sectional_Long_Survival_Ready,clear
tab year , gen(D)
logit _d  i.D2 i.D3 i.D4 i.D5 i.D6 i.D7 i.D8 [pweight=cswgt_final_0] if
year>2004 , nocons
*preserve and restore deal with the programming problem where the user's data
must be changed to achieve the desired result but, when the program concludes,
the programmer wishes to undo the change done to the data.
preserve
predict Hazard
gen Cyear=year
collapse (mean) Hazard [pweight=cswgt_final_0], by(Cyear)
gen Survival = 1
replace Survival = (1 - Hazard)*Survival if Cyear == 2005
replace Survival = (1 - Hazard)*Survival[_n-1] if Cyear > 2005
format Survival Hazard %6.4f
gen Interval0=Cyear-2004
gen Interval1=Interval0+1
drop if Cyear==2004
list Cyear Interval0 Interval1  Survival Hazard if Cyear<., separator(7)
restore
```

```
     +-------------------------------------------------+
     | Cyear   Interv~0   Interv~1   Survival   Hazard |
     |-------------------------------------------------|
  1. |  2005          1          2     0.9276   0.0724 |
  2. |  2006          2          3     0.8305   0.1047 |
  3. |  2007          3          4     0.7417   0.1069 |
  4. |  2008          4          5     0.6465   0.1283 |
  5. |  2009          5          6     0.5766   0.1082 |
  6. |  2010          6          7     0.5137   0.1092 |
  7. |  2011          7          8     0.4516   0.1209 |
     +-------------------------------------------------+
```

Examples 5.4 Survival, Failure and Hazard Rates Using Cox Regression

We can use Cox proportional hazard regression to estimate survival rates as well as hazard rates.

```
*We can use Cox proportional hazard regression to estimate survival rates as
well as hazard rates.
use Longitudinal_Long_Survival_Ready,clear
preserve
stcox, estimate
predict Survival,basesurv
predict Hazard ,basehc
collapse (mean) Survival Hazard, by(year)
format Survival Hazard    %6.4f
replace year=year+1
drop if year==2012
list year  Survival Hazard   , separator(7)
restore
```

```
     +-------------------------+
     | year   Survival   Hazard |
     |-------------------------|
  1. | 2005     0.8915   0.1085 |
  2. | 2006     0.7802   0.1249 |
  3. | 2007     0.7051   0.0963 |
  4. | 2008     0.6280   0.1092 |
  5. | 2009     0.5714   0.0902 |
  6. | 2010     0.5223   0.0860 |
  7. | 2011     0.4805   0.0800 |
     +-------------------------+
```

```
use Cross_Sectional_Long_Survival_Ready,clear
preserve
stcox, estimate
predict Survival,basesurv
predict Hazard ,basehc
collapse (mean) Survival Hazard, by(year)
format Survival Hazard    %6.4f
drop if year==2004
list year  Survival Hazard   , separator(7)
restore
```

```
     +-------------------------+
     | year   Survival   Hazard |
     |-------------------------|
  1. | 2005     0.9276   0.0724 |
  2. | 2006     0.8305   0.1047 |
  3. | 2007     0.7417   0.1069 |
  4. | 2008     0.6465   0.1283 |
  5. | 2009     0.5766   0.1082 |
  6. | 2010     0.5137   0.1092 |
  7. | 2011     0.4516   0.1209 |
     +-------------------------+
```

5.4 Semiparametric Analysis of Duration

The Cox proportional hazard regression model (Cox, 1972) and the associated partial likelihood theory of estimation provide flexible methods of regression for duration data. Cox's method does not require a particular probability distribution to represent the survival time, is able to incorporate time-dependent covariates, controls for tied data (two or more firms exit at the same time), and accommodates both continuous and discrete measurement of event times, as well as competing risks.

The Cox proportional hazard rate can be written as follows:

$$\lambda(t, x, \beta) = \lambda_0\left(t\right)e^{\mathbf{x}'\boldsymbol{\beta}}$$

where λ_0 is known as the baseline hazard and depends on t but not x, $\mathbf{x}$ a vector of covariates and $\boldsymbol{\beta}$ is a vector of parameters. In a competing risks setup, each of the cause-specific hazards λ_j may be modeled by the Cox regression model. Let *m* be distinct types of events of interest indexed by $j \in \left\{1, 2, \ldots, m\right\}$, then the cause-specific hazard λ_j is:

$\lambda_j(t, x_j, \beta_j) = \lambda_{0j}\left(t\right)e^{\mathbf{x}'_j\boldsymbol{\beta}_j}$ and the maximum likelihood function is:

$$L = \prod_{j=1}^{m}\prod_{i=1}^{k_j}\frac{e^{\mathbf{x}'_j\boldsymbol{\beta}_j}}{\sum_{k\in R}e^{\mathbf{x}'_{jk}\boldsymbol{\beta}_j}}$$

where k_j is the number of times businesses exited due to event of interest type *j*, and R is the risk set.

The maximum likelihood function enables us to estimate $\lambda_j(t, x_j, \beta_j)$ by treating all other events ($m-1$) as censored observations.

Examples 5.5 Cox Regression: Nontime-Varying Covariates

```
*Longitudinal data in wide format:Nontime-Varying Covariates
use KFS8_L7_w1,clear
svyset    [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_0)
stset Duration  , failure(event==1)
gen LnAssets_0=ln( Assets_0+1)
*Breslow method to handle tied failures; the default
*efron,exactm, exactp ,vce, shared are not allowed with the svy prefix
svy: stcox  LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0 i.Comp_advantage_0
i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0  OO_age_owner_0
OO_race_white_owner_0 ,    cformat(%6.3f)  sformat(%6.3f)   nolstretch

Number of strata   =         6                  Number of obs      =      2872
Number of PSUs     =      2872                  Population size    = 66738.821
                                                 Design df          =      2866
                                                 F(   9,   2858)    =      7.96
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  LnAssets_0 |      0.971      0.008   -3.634   0.000        0.955       0.986
1.Home_Bas~0 |      0.911      0.056   -1.508   0.132        0.807       1.028
1.Sole_Pro~0 |      0.873      0.055   -2.155   0.031        0.771       0.988
1.Comp_adv~0 |      0.985      0.061   -0.241   0.810        0.872       1.113
 1.Have_IP_0 |      0.890      0.070   -1.480   0.139        0.764       1.038
1.OO_D_edu~0 |      0.781      0.047   -4.102   0.000        0.694       0.879
OO_work_ex~0 |      0.982      0.003   -5.482   0.000        0.975       0.988
OO_age_own~0 |      1.004      0.003    1.492   0.136        0.999       1.010
OO_race_wh~0 |      0.919      0.077   -1.018   0.309        0.780       1.082
------------------------------------------------------------------------------

*Longitudinal data in Multi Episode format:Nontime-Varying Covariates
use L_multi_episode, clear
gen LnAssets_0=ln( Assets_0+1)
svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_0)
svy: stcox LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0 i.Comp_advantage_0
i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0   OO_age_owner_0
OO_race_white_owner_0  ,    cformat(%6.3f)  sformat(%6.3f)   nolstretch

Number of strata   =         6                  Number of obs      =     16775
Number of PSUs     =      2872                  Population size    = 374120.14
                                                 Design df          =      2866
                                                 F(   9,   2858)    =      7.96
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  LnAssets_0 |      0.971      0.008   -3.634   0.000        0.955       0.986
1.Home_Bas~0 |      0.911      0.056   -1.508   0.132        0.807       1.028
1.Sole_Pro~0 |      0.873      0.055   -2.155   0.031        0.771       0.988
1.Comp_adv~0 |      0.985      0.061   -0.241   0.810        0.872       1.113
 1.Have_IP_0 |      0.890      0.070   -1.480   0.139        0.764       1.038
1.OO_D_edu~0 |      0.781      0.047   -4.102   0.000        0.694       0.879
OO_work_ex~0 |      0.982      0.003   -5.482   0.000        0.975       0.988
OO_age_own~0 |      1.004      0.003    1.492   0.136        0.999       1.010
OO_race_wh~0 |      0.919      0.077   -1.018   0.309        0.780       1.082
------------------------------------------------------------------------------
```

```
*Longitudinal data in long format:Time-Varying Covariates
*Can be used for Nontime-Varying Covariates analysis
use Longitudinal_Long_Survival_Ready,clear
*Creat nontime-varying covariates (at baseline)
*Declare data to be panel data
xtset mprid year
preserve
gen LnAssets=ln( Assets+1)
bysort mprid (year): gen LnAssets_0=LnAssets[1]
bysort mprid (year): gen Home_Based_0=Home_Based[1]
bysort mprid (year): gen Sole_Proprietorship_0=Sole_Proprietorship[1]
bysort mprid (year): gen Comp_advantage_0=Comp_advantage[1]
bysort mprid (year): gen Have_IP_0=Have_IP[1]
bysort mprid (year): gen OO_D_education_owner_0=OO_D_education_owner[1]
bysort mprid (year): gen OO_work_exp_owner_0=OO_work_exp_owner[1]
bysort mprid (year): gen OO_age_owner_0=OO_age_owner[1]
bysort mprid (year): gen OO_race_white_owner_0=OO_race_white_owner[1]
*Declare survey design for dataset
svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
svy: stcox LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0 i.Comp_advantage_0
i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0   OO_age_owner_0
OO_race_white_owner_0 ,    cformat(%6.3f)  sformat(%6.3f)   nolstretch
restore

Number of strata   =         6                  Number of obs      =     16775
Number of PSUs     =      2872                  Population size    = 374120.14
                                                 Design df          =      2866
                                                 F(   9,   2858)    =      7.96
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  LnAssets_0 |      0.971      0.008   -3.634   0.000        0.955       0.986
1.Home_Bas~0 |      0.911      0.056   -1.508   0.132        0.807       1.028
1.Sole_Pro~0 |      0.873      0.055   -2.155   0.031        0.771       0.988
1.Comp_adv~0 |      0.985      0.061   -0.241   0.810        0.872       1.113
 1.Have_IP_0 |      0.890      0.070   -1.480   0.139        0.764       1.038
1.OO_D_edu~0 |      0.781      0.047   -4.102   0.000        0.694       0.879
OO_work_ex~0 |      0.982      0.003   -5.482   0.000        0.975       0.988
OO_age_own~0 |      1.004      0.003    1.492   0.136        0.999       1.010
OO_race_wh~0 |      0.919      0.077   -1.018   0.309        0.780       1.082
------------------------------------------------------------------------------

*Multiply Imputed (MI) Longitudinal data in long format:Time-Varying Covariates
*Can be used for Nontime-Varying Covariates analysis
use Longitudinal_Long_MI_Survival_Ready,clear
*Creat nontime-varying covariates (at baseline)
mi xtset mprid year
*preserve
*mi xeq is not a memory-efficient
/*
mi xeq:bysort    mprid (year): gen Home_Based_0=Home_Based[1]
mi xeq:bysort    mprid (year): gen Sole_Proprietorship_0=Sole_Proprietorship[1]
mi xeq:bysort    mprid (year): gen Comp_advantage_0=Comp_advantage[1]
mi xeq:bysort    mprid (year): gen Have_IP_0=Have_IP[1]
mi xeq:bysort    mprid (year): gen
OO_D_education_owner_0=OO_D_education_owner[1]
mi xeq:bysort    mprid (year): gen OO_work_exp_owner_0=OO_work_exp_owner[1]
mi xeq:bysort    mprid (year): gen OO_age_owner_0=OO_age_owner[1]
mi xeq:bysort    mprid (year): gen OO_race_white_owner_0=OO_race_white_owner[1]
```

```
*/
* We can use more memory-efficient way to achieve the same results as mi xeq
sort  _mi_m mprid  year
gen LnAssets=ln( Assets+1)
bysort _mi_m  mprid (year): gen LnAssets_0=LnAssets[1]
bysort _mi_m  mprid (year): gen Home_Based_0=Home_Based[1]
bysort _mi_m  mprid (year): gen Sole_Proprietorship_0=Sole_Proprietorship[1]
bysort _mi_m  mprid (year): gen Comp_advantage_0=Comp_advantage[1]
bysort _mi_m  mprid (year): gen Have_IP_0=Have_IP[1]
bysort _mi_m  mprid (year): gen OO_D_education_owner_0=OO_D_education_owner[1]
bysort _mi_m  mprid (year): gen OO_work_exp_owner_0=OO_work_exp_owner[1]
bysort _mi_m  mprid (year): gen OO_age_owner_0=OO_age_owner[1]
bysort _mi_m  mprid (year): gen OO_race_white_owner_0=OO_race_white_owner[1]
mi svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
mi xeq 0   :svy: stcox LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0  OO_race_white_owner_0 ,  cformat(%6.3f)  sformat(%6.3f)
nolstretch

Number of strata   =         6                  Number of obs      =     16775
Number of PSUs     =      2872                  Population size    = 374120.14
                                                 Design df          =      2866
                                                 F(   9,   2858)    =      7.96
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |               Linearized
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  LnAssets_0 |      0.971      0.008   -3.634   0.000        0.955       0.986
1.Home_Bas~0 |      0.911      0.056   -1.508   0.132        0.807       1.028
1.Sole_Pro~0 |      0.873      0.055   -2.155   0.031        0.771       0.988
1.Comp_adv~0 |      0.985      0.061   -0.241   0.810        0.872       1.113
 1.Have_IP_0 |      0.890      0.070   -1.480   0.139        0.764       1.038
1.OO_D_edu~0 |      0.781      0.047   -4.102   0.000        0.694       0.879
OO_work_ex~0 |      0.982      0.003   -5.482   0.000        0.975       0.988
OO_age_own~0 |      1.004      0.003    1.492   0.136        0.999       1.010
OO_race_wh~0 |      0.919      0.077   -1.018   0.309        0.780       1.082
------------------------------------------------------------------------------


*Multiply Imputed (MI)
mi estimate:svy: stcox LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0  OO_race_white_owner_0 ,   cformat(%6.3f)  sformat(%6.3f)
nolstretch
mi estimate ,hr  cformat(%6.3f)  sformat(%6.3f)   nolstretch
restore
```

```
Multiple-imputation estimates                   Imputations       =          5
Survey: Cox regression                          Number of obs     =      18286

Number of strata  =         6                   Population size   = 408495.43
Number of PSUs    =      3140
                                                Average RVI       =     0.0083
                                                Largest FMI       =     0.0244
                                                Complete DF       =       3134
DF adjustment:   Small sample                   DF:     min       =    2119.31
                                                        avg       =    2890.95
                                                        max       =    3127.49
Model F test:       Equal FMI                   F(   9, 3109.1)   =       8.11
Within VCE type:   Linearized                   Prob > F          =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  LnAssets_0 |      0.969      0.008   -3.973   0.000        0.954       0.984
1.Home_Bas~0 |      0.919      0.054   -1.439   0.150        0.819       1.031
1.Sole_Pro~0 |      0.851      0.051   -2.671   0.008        0.756       0.958
1.Comp_adv~0 |      0.995      0.060   -0.089   0.929        0.884       1.119
 1.Have_IP_0 |      0.908      0.068   -1.292   0.196        0.785       1.051
1.OO_D_edu~0 |      0.798      0.046   -3.922   0.000        0.713       0.893
OO_work_ex~0 |      0.982      0.003   -5.572   0.000        0.976       0.988
OO_age_own~0 |      1.005      0.003    1.825   0.068        1.000       1.011
OO_race_wh~0 |      0.907      0.072   -1.231   0.218        0.776       1.060
------------------------------------------------------------------------------

*Cross Sectional data in Multi Episode format:Nontime-Varying Covariates
use CS_multi_episode, clear
gen LnAssets_0=ln( Assets_0+1)
svyset  mprid  [pweight=cswgt_final_0] , strata(sampleinfo_samplestrata_0)
svy: stcox LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0 i.Comp_advantage_0
i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0   OO_age_owner_0
OO_race_white_owner_0 ,cformat(%6.3f)  sformat(%6.3f)   nolstretch

Number of strata   =         6                Number of obs      =     24673
Number of PSUs     =      4415                Population size    = 361135.56
                                               Design df          =      4409
                                               F(   9,   4401)    =      8.88
                                               Prob > F           =    0.0000

------------------------------------------------------------------------------
             |               Linearized
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  LnAssets_0 |      0.969      0.007   -4.580   0.000        0.956       0.982
1.Home_Bas~0 |      0.882      0.045   -2.480   0.013        0.798       0.974
1.Sole_Pro~0 |      0.952      0.050   -0.935   0.350        0.859       1.055
1.Comp_adv~0 |      1.002      0.052    0.031   0.975        0.904       1.110
 1.Have_IP_0 |      0.911      0.058   -1.464   0.143        0.804       1.032
1.OO_D_edu~0 |      0.830      0.042   -3.683   0.000        0.752       0.917
OO_work_ex~0 |      0.983      0.003   -6.050   0.000        0.978       0.989
OO_age_own~0 |      1.005      0.002    1.917   0.055        1.000       1.010
OO_race_wh~0 |      0.939      0.061   -0.977   0.329        0.826       1.066
------------------------------------------------------------------------------
```

```
*Cross Sectional data in long format:Time-Varying Covariates
*Can be used for Nontime-Varying Covariates analysis
use Cross_Sectional_Long_Survival_Ready,clear
*Assuming Noninformative Censoring. Conditioning on the explanatory variables,
the fact that a firm is censored at time t does not give any information about
the firm's hazard at time t.
*That is, firms(Drop-Outs) are not censored because they are at higher or lower
risk of an event
*Creat nontime-varying covariates (at baseline)
xtset mprid year
preserve
*mi xeq is not a memory-efficient
*We can use more memory-efficient way to achieve the same results as mi xeq
gen LnAssets=ln( Assets+1)
bysort _mi_m  mprid (year): gen LnAssets_0=LnAssets[1]
bysort mprid (year): gen Home_Based_0=Home_Based[1]
bysort mprid (year): gen Sole_Proprietorship_0=Sole_Proprietorship[1]
bysort mprid (year): gen Comp_advantage_0=Comp_advantage[1]
bysort mprid (year): gen Have_IP_0=Have_IP[1]
bysort mprid (year): gen OO_D_education_owner_0=OO_D_education_owner[1]
bysort mprid (year): gen OO_work_exp_owner_0=OO_work_exp_owner[1]
bysort mprid (year): gen OO_age_owner_0=OO_age_owner[1]
bysort mprid (year): gen OO_race_white_owner_0=OO_race_white_owner[1]
svyset mprid  [pweight=cswgt_final_0] , strata(sampleinfo_samplestrata)
svy: stcox LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0 i.Comp_advantage_0
i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0   OO_age_owner_0
OO_race_white_owner_0 ,cformat(%6.3f)  sformat(%6.3f)   nolstretch
restore


Number of strata   =         6                 Number of obs      =     24673
Number of PSUs     =      4415                 Population size    = 361135.56
                                                Design df          =      4409
                                                F(   9,   4401)    =      8.88
                                                Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  LnAssets_0 |      0.969      0.007   -4.580   0.000        0.956       0.982
1.Home_Bas~0 |      0.882      0.045   -2.480   0.013        0.798       0.974
1.Sole_Pro~0 |      0.952      0.050   -0.935   0.350        0.859       1.055
1.Comp_adv~0 |      1.002      0.052    0.031   0.975        0.904       1.110
 1.Have_IP_0 |      0.911      0.058   -1.464   0.143        0.804       1.032
1.OO_D_edu~0 |      0.830      0.042   -3.683   0.000        0.752       0.917
OO_work_ex~0 |      0.983      0.003   -6.050   0.000        0.978       0.989
OO_age_own~0 |      1.005      0.002    1.917   0.055        1.000       1.010
OO_race_wh~0 |      0.939      0.061   -0.977   0.329        0.826       1.066
------------------------------------------------------------------------------
```

```
*Multiply Imputed (MI) Cross Sectional data in long format:Time-Varying
Covariates
*Can be used for Nontime-Varying Covariates analysis
use Cross_Sectional_Long_MI_Survival_Ready,clear
*Creat nontime-varying covariates (at baseline)
mi xtset mprid year
*preserve
*mi xeq is not a memory-efficient
*We can use more memory-efficient way to achieve the same results as mi xeq
sort  _mi_m mprid  year
gen LnAssets=ln( Assets+1)
bysort _mi_m  mprid (year): gen LnAssets_0=LnAssets[1]
bysort _mi_m  mprid (year): gen Home_Based_0=Home_Based[1]
bysort _mi_m  mprid (year): gen Sole_Proprietorship_0=Sole_Proprietorship[1]
bysort _mi_m  mprid (year): gen Comp_advantage_0=Comp_advantage[1]
bysort _mi_m  mprid (year): gen Have_IP_0=Have_IP[1]
bysort _mi_m  mprid (year): gen OO_D_education_owner_0=OO_D_education_owner[1]
bysort _mi_m  mprid (year): gen OO_work_exp_owner_0=OO_work_exp_owner[1]
bysort _mi_m  mprid (year): gen OO_age_owner_0=OO_age_owner[1]
bysort _mi_m  mprid (year): gen OO_race_white_owner_0=OO_race_white_owner[1]
mi svyset  mprid  [pweight=cswgt_final_0] , strata(sampleinfo_samplestrata)
mi xeq 0   :svy: stcox LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0  OO_race_white_owner_0 ,  cformat(%6.3f)  sformat(%6.3f)
nolstretch


Number of strata   =         6                  Number of obs      =     24673
Number of PSUs     =      4415                  Population size    = 361135.56
                                                 Design df          =      4409
                                                 F(   9,   4401)    =      8.88
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  LnAssets_0 |      0.969      0.007   -4.580   0.000        0.956       0.982
1.Home_Bas~0 |      0.882      0.045   -2.480   0.013        0.798       0.974
1.Sole_Pro~0 |      0.952      0.050   -0.935   0.350        0.859       1.055
1.Comp_adv~0 |      1.002      0.052    0.031   0.975        0.904       1.110
 1.Have_IP_0 |      0.911      0.058   -1.464   0.143        0.804       1.032
1.OO_D_edu~0 |      0.830      0.042   -3.683   0.000        0.752       0.917
OO_work_ex~0 |      0.983      0.003   -6.050   0.000        0.978       0.989
OO_age_own~0 |      1.005      0.002    1.917   0.055        1.000       1.010
OO_race_wh~0 |      0.939      0.061   -0.977   0.329        0.826       1.066
------------------------------------------------------------------------------
```

```
*Multiply Imputed (MI)
mi estimate:svy: stcox i.Home_Based_0 i.Sole_Proprietorship_0 i.Comp_advantage_0
i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0   OO_age_owner_0
OO_race_white_owner_0 ,cformat(%6.3f)  sformat(%6.3f)   nolstretch
mi estimate ,hr  cformat(%6.3f)  sformat(%6.3f)   nolstretch
*restore

Multiple-imputation estimates                   Imputations        =         5
Survey: Cox regression                          Number of obs      =     27270

Number of strata  =         6                   Population size    = 399089.66
Number of PSUs    =      4928
                                                Average RVI        =    0.0064
                                                Largest FMI        =    0.0220
                                                Complete DF        =      4922
DF adjustment:   Small sample                   DF:     min        =   3067.48
                                                        avg        =   4571.57
                                                        max        =   4914.20
Model F test:       Equal FMI                   F(   9, 4886.1)    =      9.08
Within VCE type:   Linearized                   Prob > F           =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  LnAssets_0 |      0.970      0.006   -4.601   0.000        0.958       0.983
1.Home_Bas~0 |      0.895      0.043   -2.317   0.021        0.814       0.983
1.Sole_Pro~0 |      0.932      0.046   -1.431   0.153        0.845       1.027
1.Comp_adv~0 |      0.994      0.050   -0.111   0.911        0.902       1.096
 1.Have_IP_0 |      0.932      0.056   -1.176   0.240        0.829       1.048
1.OO_D_edu~0 |      0.832      0.040   -3.833   0.000        0.758       0.914
OO_work_ex~0 |      0.984      0.003   -6.222   0.000        0.979       0.989
OO_age_own~0 |      1.005      0.002    2.214   0.027        1.001       1.010
OO_race_wh~0 |      0.941      0.058   -0.992   0.321        0.834       1.061
------------------------------------------------------------------------------
```

Examples 5.6 Cox Competing Risks: Nontime-Varying Covariates

If all closures are not the same (out of business, merge, acquired, and temporarily stopped), it might not be appropriate to lump closures all together or one might just want a more refined analysis (closure through merger or closed down).

```
* Studying Out of Business only.
use Longitudinal_Long_Survival_Ready,clear
*Longitudinal data in long format:Cox Competing Risks Time-Varying Covariates
*Can be used for Nontime-Varying Covariates analysis
*Creat nontime-varying covariates (at baseline)
*Declare data to be panel data
xtset mprid year
preserve
bysort mprid (year): gen Home_Based_0=Home_Based[1]
bysort mprid (year): gen Sole_Proprietorship_0=Sole_Proprietorship[1]
bysort mprid (year): gen Comp_advantage_0=Comp_advantage[1]
bysort mprid (year): gen Have_IP_0=Have_IP[1]
bysort mprid (year): gen OO_D_education_owner_0=OO_D_education_owner[1]
bysort mprid (year): gen OO_work_exp_owner_0=OO_work_exp_owner[1]
bysort mprid (year): gen OO_age_owner_0=OO_age_owner[1]
bysort mprid (year): gen OO_race_white_owner_0=OO_race_white_owner[1]
*Declare survey design for dataset
svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
```

```
*It causes Stata to forget the st markers, making the data no longer st data to
Stata.  The data remain unchanged.  It is not necessary to stset, clear before
doing another stset.
stset,clear
* Studying Out of Business only:Competing==4
stset Duration [pweight=wgt_7_long] , failure(Competing==4)  id(mprid)
svy: stcox i.Home_Based_0 i.Sole_Proprietorship_0 i.Comp_advantage_0 i.Have_IP_0
i.OO_D_education_owner_0  OO_work_exp_owner_0   OO_age_owner_0
OO_race_white_owner_0 ,    cformat(%6.3f)  sformat(%6.3f)   nolstretch
restore

Number of strata   =         6                 Number of obs      =     17995
Number of PSUs     =      3077                 Population size    = 401614.07
                                               Design df          =      3071
                                               F(   8,   3064)    =      7.65
                                               Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.Home_Bas~0 |      1.102      0.070    1.528   0.127        0.973       1.248
1.Sole_Pro~0 |      0.898      0.059   -1.645   0.100        0.790       1.021
1.Comp_adv~0 |      0.931      0.060   -1.105   0.269        0.821       1.056
 1.Have_IP_0 |      0.905      0.074   -1.224   0.221        0.771       1.062
1.OO_D_edu~0 |      0.752      0.047   -4.515   0.000        0.665       0.851
OO_work_ex~0 |      0.983      0.003   -5.011   0.000        0.976       0.989
OO_age_own~0 |      1.005      0.003    1.722   0.085        0.999       1.011
OO_race_wh~0 |      0.863      0.073   -1.745   0.081        0.731       1.018
------------------------------------------------------------------------------




*Multiply Imputed (MI) Longitudinal data in long format:Cox Competing Risks
Time-Varying Covariates
*Can be used for Nontime-Varying Covariates analysis
use Longitudinal_Long_MI_Survival_Ready,clear
*Creat nontime-varying covariates (at baseline)
mi xtset mprid year
preserve
*mi xeq is not a memory-efficient
* We can use more memory-efficient way to achieve the same results as mi xeq
sort  _mi_m mprid  year
bysort _mi_m  mprid (year): gen Home_Based_0=Home_Based[1]
bysort _mi_m  mprid (year): gen Sole_Proprietorship_0=Sole_Proprietorship[1]
bysort _mi_m  mprid (year): gen Comp_advantage_0=Comp_advantage[1]
bysort _mi_m  mprid (year): gen Have_IP_0=Have_IP[1]
bysort _mi_m  mprid (year): gen OO_D_education_owner_0=OO_D_education_owner[1]
bysort _mi_m  mprid (year): gen OO_work_exp_owner_0=OO_work_exp_owner[1]
bysort _mi_m  mprid (year): gen OO_age_owner_0=OO_age_owner[1]
bysort _mi_m  mprid (year): gen OO_race_white_owner_0=OO_race_white_owner[1]
mi svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
mi stset,clear
* Studying Out of Business only:Competing==4
mi stset Duration [pweight=wgt_7_long] , failure(Competing==4)  id(mprid)
mi xeq 0   :svy: stcox i.Home_Based_0 i.Sole_Proprietorship_0 i.Comp_advantage_0
i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0   OO_age_owner_0
OO_race_white_owner_0 ,  cformat(%6.3f)  sformat(%6.3f)   nolstretch
```

```
Number of strata   =         6                  Number of obs      =     17995
Number of PSUs     =      3077                  Population size    = 401614.07
                                                 Design df          =      3071
                                                 F(   8,   3064)    =      7.65
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.Home_Bas~0 |      1.102      0.070    1.528   0.127        0.973       1.248
1.Sole_Pro~0 |      0.898      0.059   -1.645   0.100        0.790       1.021
1.Comp_adv~0 |      0.931      0.060   -1.105   0.269        0.821       1.056
 1.Have_IP_0 |      0.905      0.074   -1.224   0.221        0.771       1.062
1.OO_D_edu~0 |      0.752      0.047   -4.515   0.000        0.665       0.851
OO_work_ex~0 |      0.983      0.003   -5.011   0.000        0.976       0.989
OO_age_own~0 |      1.005      0.003    1.722   0.085        0.999       1.011
OO_race_wh~0 |      0.863      0.073   -1.745   0.081        0.731       1.018
------------------------------------------------------------------------------
```

```
*Multiply Imputed (MI)
cap mi estimate:svy: stcox i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0  OO_race_white_owner_0 ,   cformat(%6.3f)  sformat(%6.3f)
nolstretch
mi estimate ,hr  cformat(%6.3f)  sformat(%6.3f)   nolstretch
restore
```

```
Multiple-imputation estimates                  Imputations        =         5
Survey: Cox regression                         Number of obs      =     18286

Number of strata  =         6                  Population size    = 408495.43
Number of PSUs    =      3140
                                               Average RVI        =    0.0083
                                               Largest FMI        =    0.0272
                                               Complete DF        =      3134
DF adjustment:   Small sample                  DF:     min        =   1970.13
                                                       avg        =   2869.35
                                                       max        =   3131.69
Model F test:       Equal FMI                  F(   8, 3105.3)    =      7.88
Within VCE type:   Linearized                  Prob > F           =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.Home_Bas~0 |      1.098      0.069    1.503   0.133        0.972       1.241
1.Sole_Pro~0 |      0.895      0.058   -1.732   0.083        0.788       1.015
1.Comp_adv~0 |      0.927      0.059   -1.190   0.234        0.818       1.051
 1.Have_IP_0 |      0.913      0.074   -1.133   0.257        0.779       1.069
1.OO_D_edu~0 |      0.747      0.047   -4.661   0.000        0.661       0.845
OO_work_ex~0 |      0.983      0.003   -5.013   0.000        0.976       0.989
OO_age_own~0 |      1.006      0.003    1.892   0.059        1.000       1.012
OO_race_wh~0 |      0.851      0.070   -1.947   0.052        0.724       1.001
------------------------------------------------------------------------------
```

```
*Multiply Imputed (MI) Cross Sectional data in long format:Time-Varying
Covariates
*Can be used for Nontime-Varying Covariates analysis
use Cross_Sectional_Long_MI_Survival_Ready,clear
*Creat nontime-varying covariates (at baseline)
mi xtset mprid year
*preserve
*mi xeq is not a memory-efficient
*We can use more memory-efficient way to achieve the same results as mi xeq
sort  _mi_m mprid  year
bysort _mi_m  mprid (year): gen Home_Based_0=Home_Based[1]
bysort _mi_m  mprid (year): gen Sole_Proprietorship_0=Sole_Proprietorship[1]
bysort _mi_m  mprid (year): gen Comp_advantage_0=Comp_advantage[1]
bysort _mi_m  mprid (year): gen Have_IP_0=Have_IP[1]
bysort _mi_m  mprid (year): gen OO_D_education_owner_0=OO_D_education_owner[1]
bysort _mi_m  mprid (year): gen OO_work_exp_owner_0=OO_work_exp_owner[1]
bysort _mi_m  mprid (year): gen OO_age_owner_0=OO_age_owner[1]
bysort _mi_m  mprid (year): gen OO_race_white_owner_0=OO_race_white_owner[1]
mi svyset  mprid  [pweight=cswgt_final_0] , strata(sampleinfo_samplestrata)
mi stset,clear
* Studying Out of Business only:Competing==4
mi stset Duration [pweight=cswgt_final_0] , failure(Competing==4)  id(mprid)
mi xeq 0   :svy: stcox i.Home_Based_0 i.Sole_Proprietorship_0 i.Comp_advantage_0
i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0   OO_age_owner_0
OO_race_white_owner_0 ,  cformat(%6.3f)  sformat(%6.3f)   nolstretch
Number of strata   =         6                 Number of obs      =     26757
Number of PSUs     =      4813                 Population size    =  390810.6
                                                Design df          =      4807
                                                F(   8,   4800)    =      8.47
                                                Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.Home_Bas~0 |      1.040      0.053    0.758   0.449        0.940       1.150
1.Sole_Pro~0 |      0.992      0.053   -0.158   0.874        0.893       1.101
1.Comp_adv~0 |      0.939      0.050   -1.185   0.236        0.847       1.042
 1.Have_IP_0 |      0.964      0.063   -0.569   0.569        0.849       1.094
1.OO_D_edu~0 |      0.771      0.040   -4.962   0.000        0.696       0.855
OO_work_ex~0 |      0.985      0.003   -5.412   0.000        0.979       0.990
OO_age_own~0 |      1.005      0.003    2.092   0.036        1.000       1.010
OO_race_wh~0 |      0.872      0.057   -2.093   0.036        0.767       0.991
------------------------------------------------------------------------------




*Multiply Imputed (MI)
cap mi estimate:svy: stcox i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0  OO_race_white_owner_0 ,   cformat(%6.3f)  sformat(%6.3f)
nolstretch
mi estimate ,hr  cformat(%6.3f)  sformat(%6.3f)   nolstretch
*restore
```

```
Multiple-imputation estimates                   Imputations       =          5
Survey: Cox regression                          Number of obs     =      27270

Number of strata  =         6                   Population size   = 399089.66
Number of PSUs    =      4928
                                                Average RVI       =     0.0052
                                                Largest FMI       =     0.0157
                                                Complete DF       =       4922
DF adjustment:   Small sample                   DF:     min       =    3740.27
                                                        avg       =    4659.68
                                                        max       =    4919.39
Model F test:       Equal FMI                   F(   8, 4894.0)   =       8.62
Within VCE type:   Linearized                   Prob > F          =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.Home_Bas~0 |      1.039      0.053    0.757   0.449        0.941       1.147
1.Sole_Pro~0 |      0.983      0.052   -0.330   0.742        0.886       1.090
1.Comp_adv~0 |      0.930      0.049   -1.381   0.167        0.840       1.031
 1.Have_IP_0 |      0.967      0.062   -0.531   0.595        0.852       1.096
1.OO_D_edu~0 |      0.776      0.040   -4.904   0.000        0.702       0.859
OO_work_ex~0 |      0.984      0.003   -5.571   0.000        0.979       0.990
OO_age_own~0 |      1.006      0.003    2.287   0.022        1.001       1.011
OO_race_wh~0 |      0.876      0.056   -2.056   0.040        0.772       0.994
------------------------------------------------------------------------------
```

```
* Studying Drop Outs only:Competing==5

*Multiply Imputed (MI) Cross Sectional data in long format:Time-Varying
Covariates
*Can be used for Nontime-Varying Covariates analysis
use Cross_Sectional_Long_MI_Survival_Ready,clear
*Creat nontime-varying covariates (at baseline)
mi xtset mprid year
*preserve
*mi xeq is not a memory-efficient
*We can use more memory-efficient way to achieve the same results as mi xeq
sort  _mi_m mprid  year
bysort _mi_m  mprid (year): gen Home_Based_0=Home_Based[1]
bysort _mi_m  mprid (year): gen Sole_Proprietorship_0=Sole_Proprietorship[1]
bysort _mi_m  mprid (year): gen Comp_advantage_0=Comp_advantage[1]
bysort _mi_m  mprid (year): gen Have_IP_0=Have_IP[1]
bysort _mi_m  mprid (year): gen OO_D_education_owner_0=OO_D_education_owner[1]
bysort _mi_m  mprid (year): gen OO_work_exp_owner_0=OO_work_exp_owner[1]
bysort _mi_m  mprid (year): gen OO_age_owner_0=OO_age_owner[1]
bysort _mi_m  mprid (year): gen OO_race_white_owner_0=OO_race_white_owner[1]
mi svyset  mprid  [pweight=cswgt_final_0] , strata(sampleinfo_samplestrata)
mi stset,clear
* Studying Drop Outs only:Competing==5
mi stset Duration [pweight=cswgt_final_0] , failure(Competing==5)  id(mprid)
mi xeq 0   :svy: stcox i.Home_Based_0 i.Sole_Proprietorship_0 i.Comp_advantage_0
i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0   OO_age_owner_0
OO_race_white_owner_0 ,  cformat(%6.3f)  sformat(%6.3f)   nolstretch
```

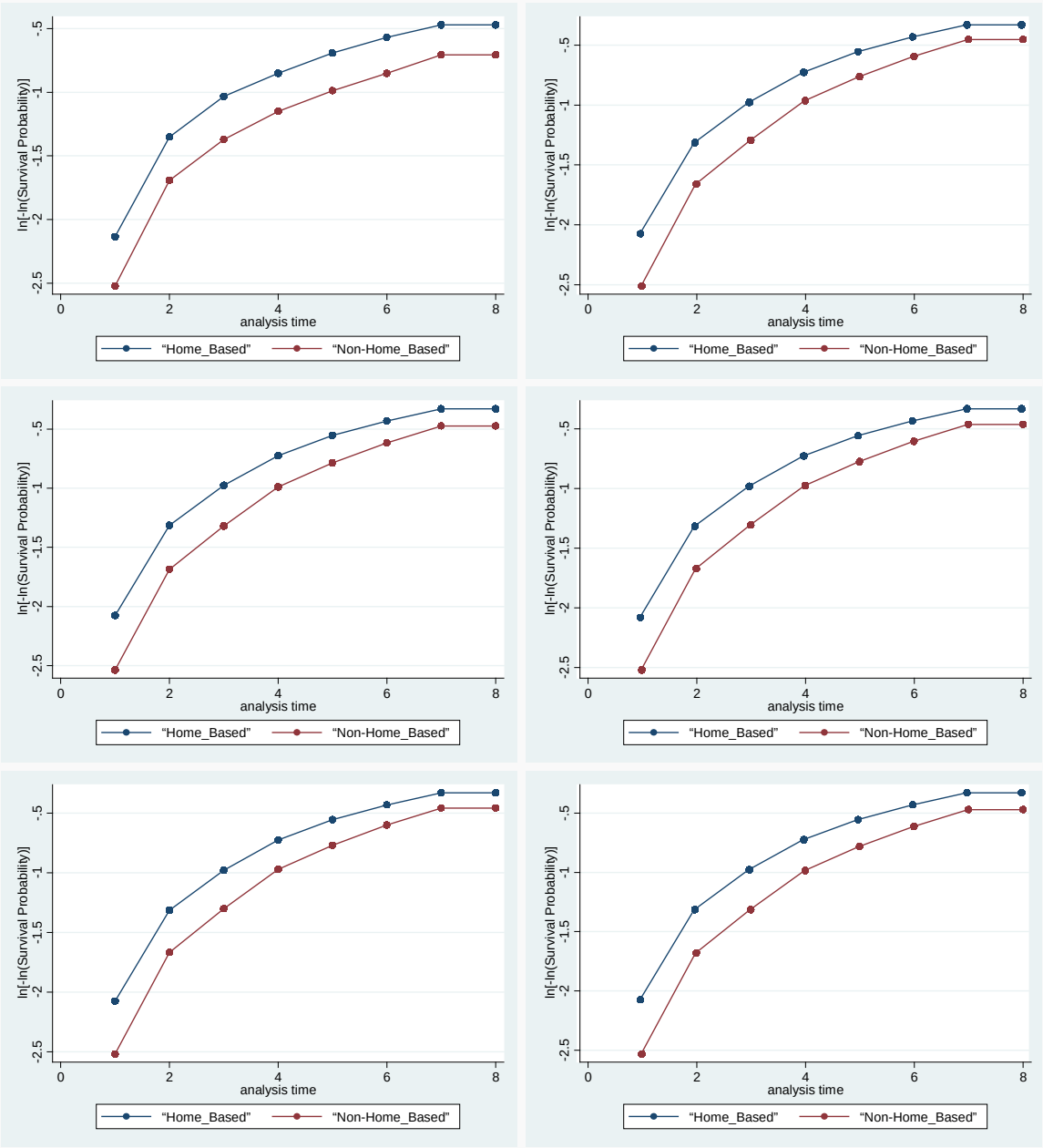
```
Number of strata   =         6                  Number of obs      =     26989
Number of PSUs     =      4897                  Population size    = 394702.25
                                                 Subpop. no. of obs =     18040
                                                 Subpop. size       = 256411.06
                                                 Design df          =      4891
                                                 F(   8,   4884)    =      7.81
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.Home_Bas~0 |      0.956      0.026   -1.654   0.098        0.906       1.008
1.Sole_Pro~0 |      0.970      0.028   -1.056   0.291        0.917       1.026
1.Comp_adv~0 |      0.989      0.028   -0.400   0.689        0.936       1.045
 1.Have_IP_0 |      0.933      0.034   -1.914   0.056        0.870       1.002
1.OO_D_edu~0 |      0.883      0.024   -4.485   0.000        0.837       0.933
OO_work_ex~0 |      0.993      0.002   -4.663   0.000        0.990       0.996
OO_age_own~0 |      1.000      0.001   -0.145   0.885        0.997       1.002
OO_race_wh~0 |      0.918      0.031   -2.559   0.011        0.860       0.980
------------------------------------------------------------------------------

*Multiply Imputed (MI)
cap mi estimate:svy: stcox i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0  OO_race_white_owner_0 ,   cformat(%6.3f)  sformat(%6.3f)
nolstretch
mi estimate ,hr  cformat(%6.3f)  sformat(%6.3f)   nolstretch



Multiple-imputation estimates                  Imputations        =         5
Survey: Cox regression                         Number of obs      =     27270

Number of strata  =         6                  Population size    = 399089.66
Number of PSUs    =      4928                  Subpop. no. of obs =     18321
                                                 Subpop. size       = 260798.46
                                                 Average RVI        =    0.0050
                                                 Largest FMI        =    0.0250
                                                 Complete DF        =      4922
DF adjustment:   Small sample                  DF:     min        =   2769.96
                                                       avg        =   4600.60
                                                       max        =   4919.55
Model F test:       Equal FMI                  F(   8, 4895.6)    =      7.91
Within VCE type:   Linearized                  Prob > F           =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.Home_Bas~0 |      0.953      0.025   -1.795   0.073        0.905       1.004
1.Sole_Pro~0 |      0.969      0.027   -1.131   0.258        0.916       1.024
1.Comp_adv~0 |      0.985      0.027   -0.543   0.587        0.933       1.040
 1.Have_IP_0 |      0.937      0.033   -1.842   0.065        0.874       1.004
1.OO_D_edu~0 |      0.886      0.024   -4.465   0.000        0.840       0.934
OO_work_ex~0 |      0.993      0.001   -4.759   0.000        0.990       0.996
OO_age_own~0 |      1.000      0.001    0.137   0.891        0.998       1.003
OO_race_wh~0 |      0.916      0.030   -2.712   0.007        0.859       0.976
------------------------------------------------------------------------------
```

Examples 5.7 Cox Regression: Time-Varying Covariates

```
*Multiply Imputed (MI) Longitudinal data in long format:Cox Competing Risks
Time-Varying Covariates
use Longitudinal_Long_MI_Survival_Ready,clear
gen LnAssets=ln( Assets+1)
mi xtset mprid year
*preserve
mi svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
*Non Imputed Data
mi xeq 0   :svy: stcox LnAssets  i.Home_Based i.Sole_Proprietorship
i.Comp_advantage i.Have_IP i.OO_D_education_owner  OO_work_exp_owner
OO_age_owner  OO_race_white_owner OO_gender_owner  Total_Employees
Gender_similarity ,  cformat(%6.3f)  sformat(%6.3f)   nolstretch


Number of strata   =         6                  Number of obs      =     16329
Number of PSUs     =      3086                  Population size    = 363391.12
                                                 Design df          =      3080
                                                 F(  12,   3069)    =      9.89
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      0.937      0.008   -7.741   0.000        0.922       0.953
1.Home_Based |      0.823      0.057   -2.810   0.005        0.718       0.943
1.Sole_Pro~p |      0.799      0.061   -2.913   0.004        0.687       0.929
1.Comp_adv~e |      0.812      0.056   -3.040   0.002        0.710       0.929
   1.Have_IP |      0.937      0.084   -0.727   0.467        0.785       1.118
1.OO_D_edu~r |      0.833      0.056   -2.697   0.007        0.730       0.951
OO_work_ex~r |      0.984      0.004   -4.126   0.000        0.976       0.992
OO_age_owner |      1.002      0.003    0.602   0.547        0.995       1.009
OO_race_wh~r |      1.059      0.102    0.595   0.552        0.876       1.280
OO_gender_~r |      0.990      0.081   -0.118   0.906        0.843       1.163
Total_Empl~s |      0.998      0.005   -0.352   0.725        0.988       1.009
Gender_sim~y |      1.236      0.251    1.043   0.297        0.830       1.840
------------------------------------------------------------------------------

*Multiply Imputed (MI)
cap mi estimate:svy: stcox LnAssets  i.Home_Based i.Sole_Proprietorship
i.Comp_advantage i.Have_IP i.OO_D_education_owner  OO_work_exp_owner
OO_age_owner  OO_race_white_owner OO_gender_owner  Total_Employees
Gender_similarity ,   cformat(%6.3f)  sformat(%6.3f)   nolstretch
mi estimate ,hr  cformat(%6.3f)  sformat(%6.3f)   nolstretch
*restore
```

```
Multiple-imputation estimates                   Imputations       =          5
Survey: Cox regression                          Number of obs     =      18286

Number of strata  =         6                   Population size   = 408495.43
Number of PSUs    =      3140
                                                Average RVI       =     0.0061
                                                Largest FMI       =     0.0213
                                                Complete DF       =       3134
DF adjustment:   Small sample                   DF:     min       =    2285.97
                                                        avg       =    2951.26
                                                        max       =    3128.03
Model F test:       Equal FMI                   F(  12, 3123.0)   =      11.95
Within VCE type:   Linearized                   Prob > F          =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      0.942      0.007   -8.070   0.000        0.929       0.956
1.Home_Based |      0.903      0.054   -1.714   0.087        0.804       1.015
1.Sole_Pro~p |      0.791      0.052   -3.586   0.000        0.697       0.899
1.Comp_adv~e |      0.831      0.049   -3.147   0.002        0.740       0.933
   1.Have_IP |      0.977      0.073   -0.305   0.760        0.844       1.132
1.OO_D_edu~r |      0.817      0.047   -3.497   0.000        0.729       0.915
OO_work_ex~r |      0.984      0.003   -4.913   0.000        0.977       0.990
OO_age_owner |      1.003      0.003    1.067   0.286        0.997       1.009
OO_race_wh~r |      0.948      0.076   -0.666   0.505        0.810       1.109
OO_gender_~r |      0.957      0.066   -0.641   0.521        0.835       1.096
Total_Empl~s |      0.997      0.005   -0.530   0.596        0.988       1.007
Gender_sim~y |      1.409      0.244    1.981   0.048        1.003       1.977
------------------------------------------------------------------------------
```

```
*Graphical check of the proportional hazards assumptions
mi xeq 0/5: stphplot, by(Home_Based) adj(LnAssets   Sole_Proprietorship
Comp_advantage Have_IP OO_D_education_owner  OO_work_exp_owner   OO_age_owner
OO_race_white_owner OO_gender_owner  Total_Employees Gender_similarity ) ///
nonegative noltime legend(lab(1 “Home_Based”) lab(2 “Non-Home_Based”) )
*the transformed survival curves are approximately parallel,No evidence against
the assumption of proportional hazards
```




```
*Multiply Imputed (MI) Cross Sectional data in long format:Time-Varying
Covariates
use Cross_Sectional_Long_Survival_Ready,clear
gen LnAssets=ln( Assets+1)
xtset mprid year
svyset  mprid  [pweight=cswgt_final_0] , strata(sampleinfo_samplestrata)
svy: stcox LnAssets  i.Home_Based i.Sole_Proprietorship i.Comp_advantage
i.Have_IP i.OO_D_education_owner  OO_work_exp_owner   OO_age_owner
OO_race_white_owner OO_gender_owner  Total_Employees Gender_similarity ,
cformat(%6.3f)  sformat(%6.3f)   nolstretch
```

```
Number of strata   =         6                  Number of obs      =     22123
Number of PSUs     =      4782                  Population size    = 321338.74
                                                 Design df          =      4776
                                                 F(   0,   4776)    =         .
                                                 Prob > F           =         .

------------------------------------------------------------------------------
             |             Linearized
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      1.000  (omitted)
1.Home_Based |      1.000  (omitted)
1.Sole_Pro~p |      1.000  (omitted)
1.Comp_adv~e |      1.000  (omitted)
   1.Have_IP |      1.000  (omitted)
1.OO_D_edu~r |      1.000  (omitted)
OO_work_ex~r |      1.000  (omitted)
OO_age_owner |      1.000  (omitted)
OO_race_wh~r |      1.000  (omitted)
OO_gender_~r |      1.000  (omitted)
Total_Empl~s |      1.000  (omitted)
Gender_sim~y |      1.000  (omitted)
------------------------------------------------------------------------------
```

```
stvary LnAssets  Home_Based Sole_Proprietorship Comp_advantage Have_IP
OO_D_education_owner  OO_work_exp_owner   OO_age_owner  OO_race_white_owner
OO_gender_owner  Total_Employees Gender_similarity

               subjects for whom the variable is
                                                  never    always   sometimes
    variable |  constant    varying              missing   missing   missing
-------------+------------------------------------------------------------
    LnAssets |       778       4044                 1767       106      3055
  Home_Based |      4513        415                 2595         0      2333
Sole_Propr~p |      4642        286                 2595         0      2333
Comp_advan~e |      2508       2411                 2512         9      2407
     Have_IP |      3711       1217                 2563         0      2365
OO_D_educa~r |      4554        359                 2565        15      2348
OO_work_ex~r |      3920        998                 2572        10      2346
OO_age_owner |       712       4206                 2566        10      2352
OO_race_wh~r |      4692        221                 2567        15      2346
OO_gender_~r |      4307        618                 2567         3      2358
Total_Empl~s |      1284       3630                 2371        14      2543
Gender_sim~y |      4344        581                 2567         3      2358
```

Using cross-sectional data to study time-varying covariates is not possible in KFS because of gaps (missing data), and KFS did not collect data in the event year. This is not a problem for the longitudinal data because we do not have gaps, and we assume that the event takes place in the beginning of the year.

Examples 5.8 Cox Competing Risks: Time-Varying Covariates

If all closures are not the same (out of business, merge, acquired, and temporarily stopped), it might not be appropriate to lump closures all together or one might just want a more refined analysis (closure through merger or closed down).

```
* Studying Sold Businesses only.
use Longitudinal_Long_Survival_Ready,clear
*Longitudinal data in long format:Cox Competing Risks Time-Varying Covariates
*Declare data to be panel data
xtset mprid year
gen LnAssets=ln( Assets+1)
*Declare survey design for dataset
svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
*It causes Stata to forget the st markers, making the data no longer st data to
Stata.  The data remain unchanged.  It is not necessary to stset, clear before
doing another stset.
stset,clear
* Studying Sold Businesses only:Competing==1
stset Duration [pweight=wgt_7_long] , failure(Competing==1)  id(mprid)
svy: stcox LnAssets  i.Home_Based i.Sole_Proprietorship i.Comp_advantage
i.Have_IP i.OO_D_education_owner  OO_work_exp_owner   OO_age_owner
OO_race_white_owner OO_gender_owner  Total_Employees Gender_similarity ,
cformat(%6.3f)  sformat(%6.3f)   nolstretch


Number of strata   =         6                  Number of obs      =     16329
Number of PSUs     =      3086                  Population size    = 363391.12
                                                 Design df          =      3080
                                                 F(  12,   3069)    =      2.93
                                                 Prob > F           =    0.0005

------------------------------------------------------------------------------
             |               Linearized
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      1.076      0.040     1.981   0.048        1.001       1.157
1.Home_Based |      0.402      0.102    -3.601   0.000        0.245       0.660
1.Sole_Pro~p |      0.916      0.244    -0.330   0.742        0.544       1.543
1.Comp_adv~e |      1.163      0.255     0.688   0.491        0.757       1.787
   1.Have_IP |      1.086      0.266     0.337   0.736        0.672       1.755
1.OO_D_edu~r |      1.077      0.233     0.341   0.733        0.704       1.647
OO_work_ex~r |      0.968      0.014    -2.248   0.025        0.940       0.996
OO_age_owner |      1.011      0.012     0.901   0.368        0.987       1.036
OO_race_wh~r |      0.860      0.266    -0.487   0.626        0.469       1.577
OO_gender_~r |      0.970      0.259    -0.115   0.908        0.575       1.637
Total_Empl~s |      0.991      0.010    -0.966   0.334        0.972       1.010
Gender_sim~y |      0.914      0.552    -0.150   0.881        0.279       2.990
------------------------------------------------------------------------------

*Multiply Imputed (MI) Longitudinal data in long format:Cox Competing Risks
Time-Varying Covariates
use Longitudinal_Long_MI_Survival_Ready,clear
mi xtset mprid year
gen LnAssets=ln( Assets+1)
mi svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
mi stset,clear
* Studying Sold Businesses only:Competing==1
mi stset Duration [pweight=wgt_7_long] , failure(Competing==1)  id(mprid)
```

```
mi xeq 0   :svy: stcox LnAssets  i.Home_Based i.Sole_Proprietorship
i.Comp_advantage i.Have_IP i.OO_D_education_owner  OO_work_exp_owner
OO_age_owner  OO_race_white_owner OO_gender_owner  Total_Employees
Gender_similarity ,   cformat(%6.3f)  sformat(%6.3f)   nolstretch

Number of strata   =         6                 Number of obs      =     16329
Number of PSUs     =      3086                 Population size    = 363391.12
                                               Design df          =      3080
                                               F(  12,   3069)    =      2.93
                                               Prob > F           =    0.0005

------------------------------------------------------------------------------
             |             Linearized
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      1.076      0.040    1.981   0.048        1.001       1.157
1.Home_Based |      0.402      0.102   -3.601   0.000        0.245       0.660
1.Sole_Pro~p |      0.916      0.244   -0.330   0.742        0.544       1.543
1.Comp_adv~e |      1.163      0.255    0.688   0.491        0.757       1.787
   1.Have_IP |      1.086      0.266    0.337   0.736        0.672       1.755
1.OO_D_edu~r |      1.077      0.233    0.341   0.733        0.704       1.647
OO_work_ex~r |      0.968      0.014   -2.248   0.025        0.940       0.996
OO_age_owner |      1.011      0.012    0.901   0.368        0.987       1.036
OO_race_wh~r |      0.860      0.266   -0.487   0.626        0.469       1.577
OO_gender_~r |      0.970      0.259   -0.115   0.908        0.575       1.637
Total_Empl~s |      0.991      0.010   -0.966   0.334        0.972       1.010
Gender_sim~y |      0.914      0.552   -0.150   0.881        0.279       2.990
------------------------------------------------------------------------------
```

```
*Multiply Imputed (MI)
cap mi estimate:svy: stcox LnAssets  i.Home_Based i.Sole_Proprietorship
i.Comp_advantage i.Have_IP i.OO_D_education_owner  OO_work_exp_owner
OO_age_owner  OO_race_white_owner OO_gender_owner  Total_Employees
Gender_similarity ,   cformat(%6.3f)  sformat(%6.3f)   nolstretch
mi estimate ,hr  cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                   Imputations        =         5
Survey: Cox regression                          Number of obs      =     18286

Number of strata  =         6                   Population size    = 408495.43
Number of PSUs    =      3140
                                                Average RVI        =    0.0108
                                                Largest FMI        =    0.0302
                                                Complete DF        =      3134
DF adjustment:   Small sample                   DF:     min        =   1816.67
                                                        avg        =   2714.36
                                                        max        =   3130.23
Model F test:       Equal FMI                   F(  12, 3104.6)    =      3.46
Within VCE type:   Linearized                   Prob > F           =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      1.070      0.037    1.975   0.048        1.001       1.145
1.Home_Based |      0.393      0.095   -3.849   0.000        0.244       0.632
1.Sole_Pro~p |      0.833      0.217   -0.703   0.482        0.499       1.388
1.Comp_adv~e |      1.206      0.260    0.867   0.386        0.790       1.841
   1.Have_IP |      1.034      0.244    0.142   0.887        0.651       1.644
1.OO_D_edu~r |      1.143      0.242    0.628   0.530        0.754       1.731
OO_work_ex~r |      0.970      0.014   -2.196   0.028        0.943       0.997
OO_age_owner |      1.009      0.012    0.756   0.450        0.986       1.032
OO_race_wh~r |      0.832      0.244   -0.628   0.530        0.469       1.477
OO_gender_~r |      0.970      0.251   -0.117   0.907        0.584       1.612
Total_Empl~s |      0.997      0.008   -0.452   0.651        0.982       1.012
Gender_sim~y |      0.896      0.522   -0.189   0.850        0.286       2.808
------------------------------------------------------------------------------
```

5.5 Parametric Analysis of Duration

Parametric analysis of duration models can be used when we know, or are willing to assume, a parametric form for the distribution of survival times and the event times are continuous. Substituting in a more reasonable distributional assumption for the disturbances leads to the adoption of parametric proportional hazards (PH) duration models, such as the exponential, gompertz, and weibull. An alternative to the parametric proportional hazards models is to use parametric accelerated failure time models (AFT).

Parametric analysis of duration can easily accommodate competing risks. Let T denote a continuous non-negative random variable representing survival time; suppose there are m distinct types of events of interest indexed by $j \in \{1,2,\ldots,m\}$ and let $\mathbf{x}$ be a vector of covariates. Then, the overall hazard rate is:

$$\lambda(t,\mathbf{x}) = \lim_{\Delta \to 0} \frac{\Pr\{t \le T < t+\Delta \mid T \ge t\}}{\Delta}$$

and the overall probability of surviving (all types of events of interest) up to time t is:

$$S(t,\mathbf{x}) = e^{-\Gamma(t,\mathbf{x})}$$

where the cumulative (or integrated) hazard $\Gamma(t,\mathbf{x}) = \int_0^t \lambda(u,\mathbf{x})\,du$. Then, the cause-specific hazard rate for the event of interest type j is:

$$\lambda_j(t,\mathbf{x}) = \lim_{\Delta \to 0} \frac{\Pr\{t \le T < t+\Delta, J = j \mid T \ge t\}}{\Delta}$$

and the survival function for the event of interest type j is:

$$S_j(t,\mathbf{x}) = e^{-\Gamma_j(t,\mathbf{x})}$$

where the cumulative (or integrated) cause-specific hazard $\Gamma_j(t,\mathbf{x}) = \int_0^t \lambda_j(u,\mathbf{x})\,du$. Therefore, the maximum likelihood function is:

$$L = \prod_{i=1}^{n}\prod_{j=1}^{m} \lambda_j(t_i,\mathbf{x}_i)^{d_{ji}} e^{-\Gamma_j(t_i,\mathbf{x}_i)}$$

where n is the number of observations, d_{ij} is an indicator variable that takes the value of 1 if the businesses exited due to the event of interest type j and 0 otherwise. The maximum likelihood function enables us to estimate $\lambda_j(t,\mathbf{x})$ by treating all other events ($m-1$) as censored observations. For survival distribution, any distribution defined for $t \in [0,\infty]$ can serve as a survival distribution. In addition, with a simple transformation

of t we can consider any distribution defined over $y \in \left[-\infty, \infty\right]$. Let $T = e^y$, so that $y = Log(T)$.

Given that there are multiple parametric models (weibull, log-logistic, log-normal), we use the Akaike information criterion (AIC) and Schwarz's Bayesian Criterion (SBC, also known as BIC) to determine which model is more desirable. The AIC and SBC statistics are calculated as follows:

$$AIC = -2 \times Log\left(L\right) + 2 \times p$$
$$SBC = -2 \times Log\left(L\right) + Log(n) \times p$$

where L is the Log Likelihood statistic, p is the number of parameters, and n is the number of observations. Lower values of the statistics (AIC, SBC) indicate a more desirable model.

Examples 5.9 Parametric Regression: Nontime-Varying Covariates

```
*Longitudinal data in wide format:Nontime-Varying Covariates
use KFS8_L7_w1,clear
gen LnAssets_0=ln( Assets_0+1)

svyset    [pweight=wgt_7_long] , strata(sampleinfo_samplestrata_0)
stset Duration [pweight=wgt_7_long] , failure(event==1)
*Breslow method to handle tied failures; the default
*efron,exactm, exactp ,vce, shared are not allowed with the svy prefix
cap svy: stcox LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0   OO_race_white_owner_0 ,     cformat(%6.3f)  sformat(%6.3f)
nolstretch
estimates store stcoxph
*Parametric proportional hazards (PH)
cap svy: streg LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0   OO_race_white_owner_0 ,  distribution(exponential)
cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store explph
cap svy: streg LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0   OO_race_white_owner_0 ,  distribution(weibull)
cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store weiph
cap svy: streg LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0   OO_race_white_owner_0 ,  distribution(gompertz)
cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store gomph

*Parametric accelerated failure time models
cap svy: streg LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0   OO_race_white_owner_0 ,  distribution(exponential) time
cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store expaft
cap svy: streg LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0   OO_race_white_owner_0 ,  distribution(weibull)     time
cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store weiaft
cap svy: streg LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0   OO_race_white_owner_0 ,  distribution(loglogistic) time
cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store logisaft
cap svy: streg LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0   OO_race_white_owner_0 ,  distribution(lognormal)   time
cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store norlaft
```

```
estimates table explph weiph gomph expaft weiaft  logisaft norlaft,stats(b  )
b(%6.4f)    star

------------------------------------------------------------------------------------------------------
    Variable |   explph       weiph        gomph        expaft       weiaft      logisaft     norlaft
-------------+----------------------------------------------------------------------------------------
_t           |
  LnAssets_0 | -0.0313***   -0.0324***   -0.0311***    0.0313***    0.0276***    0.0337***    0.0317***
             |
Home_Based_0 |
           1 | -0.0926      -0.0958      -0.0923       0.0926       0.0817       0.1194       0.1172
             |
Sole_Propr~0 |
           1 | -0.1436*     -0.1508*     -0.1427*      0.1436*      0.1286*      0.1450*      0.1296*
             |
Comp_advan~0 |
           1 | -0.0153      -0.0164      -0.0152       0.0153       0.0139       0.0187       0.0159
             |
   Have_IP_0 |
           1 | -0.1218      -0.1278      -0.1210       0.1218       0.1090       0.1253       0.1209
             |
OO_D_educa~0 |
           1 | -0.2623***   -0.2723***   -0.2610***    0.2623***    0.2323***    0.2513***    0.2418***
             |
OO_work_ex~0 | -0.0197***   -0.0204***   -0.0196***    0.0197***    0.0174***    0.0177***    0.0165***
OO_age_own~0 |  0.0049       0.0049       0.0048      -0.0049      -0.0042      -0.0030      -0.0025
OO_race_wh~0 | -0.0890      -0.0931      -0.0885       0.0890       0.0794       0.0933       0.0960
       _cons | -1.7811***   -2.0756***   -1.7536***    1.7811***    1.7705***    1.2545***    1.2851***
-------------+----------------------------------------------------------------------------------------
ln_p         |
       _cons |               0.1590***                                0.1590***
-------------+----------------------------------------------------------------------------------------
gamma        |
       _cons |                           -0.0090
-------------+----------------------------------------------------------------------------------------
ln_gam       |
       _cons |                                                                    -0.3611***
-------------+----------------------------------------------------------------------------------------
ln_sig       |
       _cons |                                                                                 0.1564***
-------------+----------------------------------------------------------------------------------------
Statistics   |
          b  |
------------------------------------------------------------------------------------------------------
legend: * p<0.05; ** p<0.01; *** p<0.001

estimates table stcoxph,stats(b  ) b(%6.4f)    star

----------------------------
    Variable |  stcoxph
-------------+-------------
  LnAssets_0 | -0.0299***
             |
1.Home_Bas~0 | -0.0930
             |
1.Sole_Pro~0 | -0.1361*
             |
1.Comp_adv~0 | -0.0150
             |
 1.Have_IP_0 | -0.1160
             |
1.OO_D_edu~0 | -0.2468***
OO_work_ex~0 | -0.0185***
OO_age_own~0 |  0.0044
OO_race_wh~0 | -0.0848
-------------+-------------
           b |
----------------------------
legend: * p<0.05; ** p<0.01; *** p<0.001
```

```
*Multiply Imputed (MI) Longitudinal data in long format:Time-Varying Covariates
*Can be used for Nontime-Varying Covariates analysis
use Longitudinal_Long_MI_Survival_Ready,clear
*Creat nontime-varying covariates (at baseline)
mi xtset mprid year
gen LnAssets=ln( Assets+1)

*mi xeq is not a memory-efficient
* We can use more memory-efficient way to achieve the same results as mi xeq
sort  _mi_m mprid  year
bysort _mi_m  mprid (year): gen LnAssets_0=LnAssets[1]
bysort _mi_m  mprid (year): gen Home_Based_0=Home_Based[1]
bysort _mi_m  mprid (year): gen Sole_Proprietorship_0=Sole_Proprietorship[1]
bysort _mi_m  mprid (year): gen Comp_advantage_0=Comp_advantage[1]
bysort _mi_m  mprid (year): gen Have_IP_0=Have_IP[1]
bysort _mi_m  mprid (year): gen OO_D_education_owner_0=OO_D_education_owner[1]
bysort _mi_m  mprid (year): gen OO_work_exp_owner_0=OO_work_exp_owner[1]
bysort _mi_m  mprid (year): gen OO_age_owner_0=OO_age_owner[1]
bysort _mi_m  mprid (year): gen OO_race_white_owner_0=OO_race_white_owner[1]
mi svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)

*Multiply Imputed (MI)

cap mi estimate,post:svy: stcox LnAssets_0 i.Home_Based_0
i.Sole_Proprietorship_0 i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0
OO_work_exp_owner_0  OO_age_owner_0   OO_race_white_owner_0 ,     cformat(%6.3f)
sformat(%6.3f)   nolstretch
estimates store stcoxph
*Parametric proportional hazards (PH)
cap mi estimate,post hr:svy: streg LnAssets_0 i.Home_Based_0
i.Sole_Proprietorship_0 i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0
OO_work_exp_owner_0  OO_age_owner_0   OO_race_white_owner_0 ,
distribution(exponential)  cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store explph
cap mi estimate,post hr:svy: streg LnAssets_0 i.Home_Based_0
i.Sole_Proprietorship_0 i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0
OO_work_exp_owner_0  OO_age_owner_0   OO_race_white_owner_0 ,
distribution(weibull)      cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store weiph
cap mi estimate,post hr:svy: streg LnAssets_0 i.Home_Based_0
i.Sole_Proprietorship_0 i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0
OO_work_exp_owner_0  OO_age_owner_0   OO_race_white_owner_0 ,
distribution(gompertz)     cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store gomph
*Parametric accelerated failure time models
cap mi estimate,post hr:svy: streg LnAssets_0 i.Home_Based_0
i.Sole_Proprietorship_0 i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0
OO_work_exp_owner_0  OO_age_owner_0   OO_race_white_owner_0 ,
distribution(exponential) time cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store expaft
cap mi estimate,post hr:svy: streg LnAssets_0 i.Home_Based_0
i.Sole_Proprietorship_0 i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0
OO_work_exp_owner_0  OO_age_owner_0   OO_race_white_owner_0 ,
distribution(weibull)     time  cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store weiaft
cap mi estimate,post hr:svy: streg LnAssets_0 i.Home_Based_0
i.Sole_Proprietorship_0 i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0
OO_work_exp_owner_0  OO_age_owner_0   OO_race_white_owner_0 ,
distribution(loglogistic) time  cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store logisaft
cap mi estimate,post hr:svy: streg LnAssets_0 i.Home_Based_0
i.Sole_Proprietorship_0 i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0
```

```
OO_work_exp_owner_0  OO_age_owner_0   OO_race_white_owner_0 ,
distribution(lognormal)   time  cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store norlaft
estimates table explph weiph gomph expaft weiaft  logisaft norlaft,stats(b  )
b(%6.4f)    star

-----------------------------------------------------------------------------------------------
    Variable |   explph       weiph        gomph        expaft       weiaft      logisaft     norlaft
-------------+---------------------------------------------------------------------------------
_t           |
  LnAssets_0 | -0.0333***   -0.0345***   -0.0331***    0.0333***    0.0297***    0.0353***    0.0330***
             |
Home_Based_0 |
           1 | -0.0830      -0.0857      -0.0827       0.0830       0.0737       0.1163       0.1149
             |
Sole_Propr~0 |
           1 | -0.1708**    -0.1788**    -0.1693**     0.1708**     0.1538**     0.1775**     0.1608**
             |
Comp_advan~0 |
           1 | -0.0052      -0.0060      -0.0050       0.0052       0.0052       0.0095       0.0062
             |
   Have_IP_0 |
           1 | -0.1023      -0.1074      -0.1013       0.1023       0.0923       0.1008       0.0922
             |
OO_D_educa~0 |
           1 | -0.2403***   -0.2490***   -0.2385***    0.2403***    0.2141***    0.2343***    0.2241***
             |
OO_work_ex~0 | -0.0192***   -0.0199***   -0.0191***    0.0192***    0.0171***    0.0175***    0.0163***
OO_age_own~0 |  0.0056       0.0058       0.0056      -0.0056      -0.0049      -0.0037      -0.0033
OO_race_wh~0 | -0.1026      -0.1068      -0.1018       0.1026       0.0919       0.1130       0.1141
       _cons | -1.7910***   -2.0682***   -1.7507***    1.7910***    1.7789***    1.2489***    1.2940***
-------------+---------------------------------------------------------------------------------
ln_p         |
       _cons |              0.1507***                              0.1507***
-------------+---------------------------------------------------------------------------------
gamma        |
       _cons |                          -0.0133
-------------+---------------------------------------------------------------------------------
ln_gam       |
       _cons |                                                                  -0.3546***
-------------+---------------------------------------------------------------------------------
ln_sig       |
       _cons |                                                                               0.1610***
-------------+---------------------------------------------------------------------------------
legend: * p<0.05; ** p<0.01; *** p<0.001

estimates table stcoxph,stats(b  ) b(%6.4f)    star

----------------------------
    Variable |  stcoxph
-------------+--------------
  LnAssets_0 | -0.0316***
             |
1.Home_Bas~0 | -0.0843
             |
1.Sole_Pro~0 | -0.1613**
             |
1.Comp_adv~0 | -0.0053
             |
 1.Have_IP_0 | -0.0961
             |
1.OO_D_edu~0 | -0.2257***
OO_work_ex~0 | -0.0181***
OO_age_own~0 |  0.0051
OO_race_wh~0 | -0.0980
-------------+--------------
           b |
----------------------------
legend: * p<0.05; ** p<0.01; *** p<0.001
```

Examples 5.10 Parametric Regression: Time-Varying Covariates

```
*Multiply Imputed (MI) Longitudinal data in long format:Time-Varying Covariates
use Longitudinal_Long_MI_Survival_Ready,clear
mi xtset mprid year
gen LnAssets=ln( Assets+1)
mi svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)

mi xeq 0:svy: stcox LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner    OO_race_white_owner  ,     cformat(%6.3f)  sformat(%6.3f)
nolstretch
estimates store stcoxph
*Parametric proportional hazards (PH)
mi xeq 0:svy: streg LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner    OO_race_white_owner  ,  distribution(exponential)
cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store explph
mi xeq 0:svy: streg LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner    OO_race_white_owner  ,  distribution(weibull)
cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store weiph
mi xeq 0:svy: streg LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner    OO_race_white_owner  ,  distribution(gompertz)
cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store gomph
*Parametric accelerated failure time models
mi xeq 0:svy: streg LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner    OO_race_white_owner  ,  distribution(exponential) time
cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store expaft
mi xeq 0:svy: streg LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner    OO_race_white_owner  ,  distribution(weibull)     time
cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store weiaft
mi xeq 0:svy: streg LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner    OO_race_white_owner  ,  distribution(loglogistic) time
cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store logisaft
mi xeq 0:svy: streg LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner    OO_race_white_owner  ,  distribution(lognormal)   time
cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store norlaft
estimates table explph weiph gomph expaft weiaft  logisaft norlaft,stats(b  )
b(%6.4f)    star
```

```
------------------------------------------------------------------------------------------------------
    Variable |   explph       weiph        gomph        expaft       weiaft      logisaft     norlaft
-------------+----------------------------------------------------------------------------------------
_t           |
    LnAssets | -0.0700***   -0.0722***   -0.0676***    0.0700***    0.0659***    0.0741***    0.0693***
             |
  Home_Based |
           1 | -0.1660*     -0.1657*     -0.1686*      0.1660*      0.1512*      0.1885**     0.1850**
             |
Sole_Propr~p |
           1 | -0.1707*     -0.1696*     -0.1723*      0.1707*      0.1547*      0.1868*      0.1679*
             |
Comp_advan~e |
           1 | -0.0881      -0.0627      -0.1284       0.0881       0.0572       0.0792       0.0828
             |
     Have_IP |
           1 | -0.0743      -0.0728      -0.0762       0.0743       0.0664       0.1035       0.1075
             |
OO_D_educa~r |
           1 | -0.2072**    -0.2121**    -0.1991**     0.2072**     0.1934**     0.2155**     0.2040**
             |
OO_work_ex~r | -0.0175***   -0.0176***   -0.0173***    0.0175***    0.0160***    0.0156***    0.0141***
OO_age_owner | -0.0006      -0.0015       0.0003       0.0006       0.0014       0.0030       0.0032
OO_race_wh~r |  0.0652       0.0667       0.0650      -0.0652      -0.0608      -0.0388      -0.0114
       _cons | -1.4217***   -1.5514***   -1.2877***    1.4217***    1.4148***    0.8231***    0.8803***
-------------+----------------------------------------------------------------------------------------
ln_p         |
       _cons |               0.0922***                              0.0922***
-------------+----------------------------------------------------------------------------------------
gamma        |
       _cons |                          -0.0543***
-------------+----------------------------------------------------------------------------------------
ln_gam       |
       _cons |                                                                    -0.2799***
-------------+----------------------------------------------------------------------------------------
ln_sig       |
       _cons |                                                                                 0.2420***
-------------+----------------------------------------------------------------------------------------
                                          legend: * p<0.05; ** p<0.01; *** p<0.001
```

```
estimates table stcoxph,stats(b  ) b(%6.4f)    star


---------------------------
    Variable |  stcoxph
-------------+-------------
    LnAssets | -0.0660***
             |
1.Home_Based | -0.1812**
             |
1.Sole_Pro~p | -0.1801*
             |
1.Comp_adv~e | -0.2161**
             |
   1.Have_IP | -0.0778
             |
1.OO_D_edu~r | -0.1814**
OO_work_ex~r | -0.0165***
OO_age_owner |  0.0012
OO_race_wh~r |  0.0678
-------------+-------------
           b |
---------------------------
legend: * p<0.05; ** p<0.01; *** p<0.001
```

```
*Multiply Imputed (MI) Longitudinal data in long format:Time-Varying Covariates
use Longitudinal_Long_MI_Survival_Ready,clear
mi xtset mprid year
gen LnAssets=ln( Assets+1)
mi svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)

*Multiply Imputed (MI)
mi estimate,post:svy: stcox LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner    OO_race_white_owner  ,     cformat(%6.3f)  sformat(%6.3f)
nolstretch
estimates store stcoxph
*Parametric proportional hazards (PH)
mi estimate,post hr:svy: streg LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner    OO_race_white_owner  ,  distribution(exponential)
cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store explph
mi estimate,post hr:svy: streg LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner    OO_race_white_owner  ,  distribution(weibull)
cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store weiph
mi estimate,post hr:svy: streg LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner    OO_race_white_owner  ,  distribution(gompertz)
cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store gomph
*Parametric accelerated failure time models
mi estimate,post hr:svy: streg LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner    OO_race_white_owner  ,  distribution(exponential) time
cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store expaft
mi estimate,post hr:svy: streg LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner    OO_race_white_owner  ,  distribution(weibull)     time
cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store weiaft
mi estimate,post hr:svy: streg LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner    OO_race_white_owner  ,  distribution(loglogistic) time
cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store logisaft
mi estimate,post hr:svy: streg LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner    OO_race_white_owner  ,  distribution(lognormal)   time
cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store norlaft
estimates table explph weiph gomph expaft weiaft  logisaft norlaft,stats(b  )
b(%6.4f)    star
```

```
------------------------------------------------------------------------------------------------------
    Variable |   explph       weiph        gomph        expaft       weiaft      logisaft     norlaft
-------------+----------------------------------------------------------------------------------------
_t           |
    LnAssets | -0.0640***   -0.0675***   -0.0634***    0.0640***    0.0577***    0.0665***    0.0631***
             |
  Home_Based |
           1 | -0.0913      -0.0925      -0.0917       0.0913       0.0792       0.1337*      0.1342*
             |
Sole_Propr~p |
           1 | -0.1724**    -0.1695*     -0.1732**     0.1724**     0.1450*      0.1759**     0.1624**
             |
Comp_advan~e |
           1 | -0.0833      -0.0407      -0.0942       0.0833       0.0348       0.0594       0.0638
             |
     Have_IP |
           1 | -0.0281      -0.0265      -0.0284       0.0281       0.0227       0.0527       0.0521
             |
OO_D_educa~r |
           1 | -0.2099***   -0.2175***   -0.2078***    0.2099***    0.1862***    0.2011***    0.1909***
             |
OO_work_ex~r | -0.0169***   -0.0170***   -0.0168***    0.0169***    0.0145***    0.0142***    0.0133***
OO_age_owner |  0.0026       0.0010       0.0028      -0.0026      -0.0009       0.0013       0.0012
OO_race_wh~r | -0.0530      -0.0502      -0.0532       0.0530       0.0430       0.0661       0.0727
       _cons | -1.4097***   -1.6419***   -1.3719***    1.4097***    1.4053***    0.7761***    0.8381***
-------------+----------------------------------------------------------------------------------------
ln_p         |
       _cons |               0.1556***                              0.1556***
-------------+----------------------------------------------------------------------------------------
gamma        |
       _cons |                           -0.0146
-------------+----------------------------------------------------------------------------------------
ln_gam       |
       _cons |                                                                    -0.3690***
-------------+----------------------------------------------------------------------------------------
ln_sig       |
       _cons |                                                                                0.1489***
-------------+----------------------------------------------------------------------------------------
                                                    legend: * p<0.05; ** p<0.01; *** p<0.001
```

```
estimates table stcoxph,stats(b  ) b(%6.4f)    star
```

```
----------------------------
    Variable |  stcoxph
-------------+--------------
    LnAssets | -0.0619***
             |
1.Home_Based | -0.1045
             |
1.Sole_Pro~p | -0.1805**
             |
1.Comp_adv~e | -0.1861**
             |
   1.Have_IP | -0.0275
             |
1.OO_D_edu~r | -0.1900***
OO_work_ex~r | -0.0159***
OO_age_owner |  0.0028
OO_race_wh~r | -0.0481
-------------+--------------
           b |
----------------------------
legend: * p<0.05; ** p<0.01; *** p<0.001
```

5.6 Discrete Time Models of Duration

So far we have assumed that event time is measured on a continuum and that events can occur at any time. However, the KFS provides us with the year in which the firm went out of business was sold, merged, temporary stopped, or dropped out. Thus, our measurement of event time is discrete and, in any given year, all events happened in the same time (ties).

We assume that the event of interest is known to happen within an interval of time indexed by t=1,2,3,..,*n* and let the probability of an event's occurrence or failure be λ_i and the probability of nonoccurrence be $1 - \lambda_i$. Also, we assume that this probability is a function of covariates

$$\lambda_{it} = \beta_0 + \beta_1 x_{1i} + .. + \beta_n x_{ni}$$

Now we need functional form for λ_{it}. The discrete time models are estimated by maximum likelihood using logit (proportional odds model) or complementary log-log (proportional hazards model).

The discrete time proportional odds model can be written as $Log\left[\frac{\lambda_i}{1-\lambda_i}\right] = \beta_0 + \beta_1 x_{1i} + .. + \beta_n x_{ni}$, where the left side transformation is the logit function.

The discrete time proportional hazard model can be written as $Log[-\log(1 - \lambda_i)] = \beta_0 + \beta_1 x_{1i} + .. + \beta_n x_{ni}$, where the left side transformation is the complementary log-log function.

The estimation procedure for both models requires manipulating the data so that each discrete time point for each business appears as a separate observation (episode splitting). The binary dependent variable in these models will take the value of 0 in all years. If the business's survival time is censored or if the business's survival time is not censored, the binary dependent variable is equal to 0 for all but the last year (event year). In our files, we already have this variable under the name of "_d."

Examples 5.11 Discrete Time Models: Nontime-Varying Covariates

```
use Longitudinal_Long_MI_Survival_Ready,clear
*Creat nontime-varying covariates (at baseline)
mi xtset mprid year
gen LnAssets=ln( Assets+1)
*mi xeq is not a memory-efficient
* We can use more memory-efficient way to achieve the same results as mi xeq
sort  _mi_m mprid  year
bysort _mi_m  mprid (year): gen LnAssets_0=LnAssets[1]
bysort _mi_m  mprid (year): gen Home_Based_0=Home_Based[1]
bysort _mi_m  mprid (year): gen Sole_Proprietorship_0=Sole_Proprietorship[1]
bysort _mi_m  mprid (year): gen Comp_advantage_0=Comp_advantage[1]
bysort _mi_m  mprid (year): gen Have_IP_0=Have_IP[1]
bysort _mi_m  mprid (year): gen OO_D_education_owner_0=OO_D_education_owner[1]
bysort _mi_m  mprid (year): gen OO_work_exp_owner_0=OO_work_exp_owner[1]
bysort _mi_m  mprid (year): gen OO_age_owner_0=OO_age_owner[1]
bysort _mi_m  mprid (year): gen OO_race_white_owner_0=OO_race_white_owner[1]
mi svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
*Describing duration dependence :For a non-parametric baseline, we need to
create year dummy variables
tab year , gen(D)
mi xeq 0/5:tab year _d
```

```
           |          _d
      year |         0          1 |     Total
-----------+----------------------+----------
      2004 |     2,837        303 |     3,140
      2005 |     2,554        283 |     2,837
      2006 |     2,330        224 |     2,554
      2007 |     2,092        238 |     2,330
      2008 |     1,928        164 |     2,092
      2009 |     1,775        153 |     1,928
      2010 |     1,644        131 |     1,775
      2011 |     1,630          0 |     1,630
-----------+----------------------+----------
     Total |    16,790      1,496 |    18,286
```

```
*For ML estimation of the discrete time logistic model we use logit
mi xeq 0: svy:  logit _d LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0  OO_race_white_owner_0 i.D1 i.D2 i.D3 i.D4 i.D5 i.D6 i.D7 if
year<2011  ,  or   nocons cformat(%6.3f)  sformat(%6.3f) nolstretch
```

```
Survey: Logistic regression

Number of strata   =         6                  Number of obs      =     15275
Number of PSUs     =      2872                  Population size    = 342091.18
                                                 Design df          =      2866
                                                 F(  16,   2851)    =    279.46
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |               Linearized
          _d | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  LnAssets_0 |      0.967      0.009   -3.598   0.000        0.949       0.985
1.Home_Bas~0 |      0.900      0.062   -1.512   0.131        0.786       1.032
1.Sole_Pro~0 |      0.858      0.061   -2.157   0.031        0.747       0.986
1.Comp_adv~0 |      0.983      0.069   -0.245   0.807        0.857       1.128
 1.Have_IP_0 |      0.879      0.077   -1.484   0.138        0.741       1.042
1.OO_D_edu~0 |      0.759      0.051   -4.109   0.000        0.665       0.865
OO_work_ex~0 |      0.980      0.004   -5.501   0.000        0.972       0.987
OO_age_own~0 |      1.005      0.003    1.509   0.131        0.999       1.012
OO_race_wh~0 |      0.909      0.086   -1.013   0.311        0.756       1.093
        1.D1 |      0.222      0.044   -7.557   0.000        0.150       0.328
        1.D2 |      0.264      0.052   -6.718   0.000        0.179       0.389
        1.D3 |      0.206      0.042   -7.717   0.000        0.138       0.308
        1.D4 |      0.240      0.049   -6.933   0.000        0.161       0.360
        1.D5 |      0.195      0.041   -7.751   0.000        0.129       0.295
        1.D6 |      0.188      0.040   -7.913   0.000        0.124       0.285
        1.D7 |      0.165      0.035   -8.399   0.000        0.109       0.252
------------------------------------------------------------------------------
```

```
*For ML estimation of the discrete time complementary log-log model we use
cloglog
mi xeq 0: svy:  cloglog _d LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0  OO_race_white_owner_0 i.D1 i.D2 i.D3 i.D4 i.D5 i.D6 i.D7 if
year<2011  ,   eform    nocons cformat(%6.3f)  sformat(%6.3f) nolstretch
```

```
Survey: Complementary log-log regression

Number of strata   =         6                  Number of obs      =     15275
Number of PSUs     =      2872                  Population size    = 342091.18
                                                 Design df          =      2866
                                                 F(  16,   2851)    =    325.43
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
          _d |     exp(b)   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  LnAssets_0 |      0.969      0.008   -3.624   0.000        0.952       0.986
1.Home_Bas~0 |      0.906      0.059   -1.514   0.130        0.797       1.030
1.Sole_Pro~0 |      0.864      0.058   -2.180   0.029        0.758       0.985
1.Comp_adv~0 |      0.985      0.065   -0.228   0.820        0.865       1.122
 1.Have_IP_0 |      0.883      0.073   -1.501   0.133        0.751       1.039
1.OO_D_edu~0 |      0.770      0.049   -4.104   0.000        0.679       0.872
OO_work_ex~0 |      0.981      0.004   -5.472   0.000        0.974       0.988
OO_age_own~0 |      1.005      0.003    1.472   0.141        0.998       1.011
OO_race_wh~0 |      0.913      0.081   -1.027   0.304        0.767       1.087
        1.D1 |      0.203      0.038   -8.474   0.000        0.141       0.294
        1.D2 |      0.239      0.044   -7.695   0.000        0.166       0.344
        1.D3 |      0.189      0.037   -8.616   0.000        0.129       0.276
        1.D4 |      0.219      0.042   -7.847   0.000        0.150       0.320
        1.D5 |      0.180      0.036   -8.651   0.000        0.122       0.266
        1.D6 |      0.174      0.035   -8.796   0.000        0.118       0.257
        1.D7 |      0.154      0.031   -9.284   0.000        0.103       0.228
------------------------------------------------------------------------------
```

```
*Multiply Imputed (MI)
mi estimate :svy: logit _d LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0  OO_race_white_owner_0 i.D1 i.D2 i.D3 i.D4 i.D5 i.D6 i.D7 if
year<2011  ,  or   nocons cformat(%6.3f)  sformat(%6.3f) nolstretch
mi estimate,hr  cformat(%6.3f)  sformat(%6.3f) nolstretch
```

```
Multiple-imputation estimates                   Imputations       =          5
Survey: Logistic regression                     Number of obs     =      16656

Number of strata  =         6                   Population size   = 373603.39
Number of PSUs    =      3140
                                                Average RVI       =     0.0047
                                                Largest FMI       =     0.0244
                                                Complete DF       =       3134
DF adjustment:   Small sample                   DF:     min       =    2119.00
                                                        avg       =    2894.91
                                                        max       =    3127.22
Model F test:       Equal FMI                   F(  16, 3128.2)   =     304.58
Within VCE type:   Linearized                   Prob > F          =     0.0000

------------------------------------------------------------------------------
          _d | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  LnAssets_0 |      0.965      0.009   -3.927   0.000        0.948       0.982
1.Home_Bas~0 |      0.909      0.060   -1.444   0.149        0.799       1.035
1.Sole_Pro~0 |      0.834      0.057   -2.676   0.007        0.730       0.953
1.Comp_adv~0 |      0.994      0.067   -0.094   0.925        0.870       1.135
 1.Have_IP_0 |      0.898      0.074   -1.294   0.196        0.764       1.057
1.OO_D_edu~0 |      0.776      0.050   -3.929   0.000        0.684       0.881
OO_work_ex~0 |      0.980      0.004   -5.590   0.000        0.973       0.987
OO_age_own~0 |      1.006      0.003    1.839   0.066        1.000       1.012
OO_race_wh~0 |      0.895      0.081   -1.225   0.221        0.750       1.069
        1.D1 |      0.225      0.043   -7.805   0.000        0.154       0.327
        1.D2 |      0.267      0.051   -6.891   0.000        0.183       0.388
        1.D3 |      0.201      0.040   -8.090   0.000        0.136       0.296
        1.D4 |      0.233      0.046   -7.335   0.000        0.157       0.343
        1.D5 |      0.189      0.038   -8.192   0.000        0.127       0.281
        1.D6 |      0.180      0.037   -8.393   0.000        0.121       0.269
        1.D7 |      0.167      0.035   -8.658   0.000        0.112       0.251
------------------------------------------------------------------------------
```

```
mi estimate : svy:  cloglog _d LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0  OO_race_white_owner_0 i.D1 i.D2 i.D3 i.D4 i.D5 i.D6 i.D7 if
year<2011  ,   eform    nocons cformat(%6.3f)  sformat(%6.3f) nolstretch
mi estimate,hr  cformat(%6.3f)  sformat(%6.3f) nolstretch
```

```
Multiple-imputation estimates                   Imputations       =          5
Survey: Complementary log-log regression        Number of obs     =      16656

Number of strata  =         6                   Population size   = 373603.39
Number of PSUs    =      3140
                                                Average RVI       =     0.0047
                                                Largest FMI       =     0.0247
                                                Complete DF       =       3134
DF adjustment:   Small sample                   DF:     min       =    2102.81
                                                        avg       =    2902.14
                                                        max       =    3127.41
Model F test:       Equal FMI                   F(  16, 3128.2)   =     355.53
Within VCE type:   Linearized                   Prob > F          =     0.0000

------------------------------------------------------------------------------
          _d | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  LnAssets_0 |      0.967      0.008   -3.961   0.000        0.951       0.983
1.Home_Bas~0 |      0.914      0.057   -1.450   0.147        0.809       1.032
1.Sole_Pro~0 |      0.841      0.054   -2.703   0.007        0.742       0.954
1.Comp_adv~0 |      0.995      0.063   -0.085   0.932        0.878       1.127
 1.Have_IP_0 |      0.902      0.071   -1.307   0.191        0.773       1.053
1.OO_D_edu~0 |      0.787      0.048   -3.920   0.000        0.698       0.887
OO_work_ex~0 |      0.981      0.003   -5.568   0.000        0.974       0.988
OO_age_own~0 |      1.005      0.003    1.798   0.072        1.000       1.011
OO_race_wh~0 |      0.900      0.076   -1.237   0.216        0.762       1.063
        1.D1 |      0.206      0.037   -8.777   0.000        0.144       0.293
        1.D2 |      0.241      0.043   -7.903   0.000        0.170       0.343
        1.D3 |      0.184      0.035   -9.027   0.000        0.128       0.266
        1.D4 |      0.212      0.040   -8.288   0.000        0.147       0.306
        1.D5 |      0.174      0.033   -9.134   0.000        0.120       0.254
        1.D6 |      0.167      0.032   -9.318   0.000        0.114       0.243
        1.D7 |      0.155      0.030   -9.591   0.000        0.106       0.227
------------------------------------------------------------------------------
```

*the estimated coefficients are similar to those for the cox model

*Describing duration dependence :we can use log(time)

```
gen log_time=log(year)
mi xeq 0: svy:  logit _d LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0  OO_race_white_owner_0 log_time if year<2011  ,  or   nocons
cformat(%6.3f)  sformat(%6.3f) nolstretch
```

```
Number of strata   =         6                  Number of obs      =     15275
Number of PSUs     =      2872                  Population size    = 342091.18
                                                 Design df          =      2866
                                                 F(  10,   2857)    =    430.63
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
          _d | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  LnAssets_0 |      0.966      0.009   -3.577   0.000        0.948       0.985
1.Home_Bas~0 |      0.900      0.064   -1.494   0.135        0.783       1.034
1.Sole_Pro~0 |      0.855      0.062   -2.173   0.030        0.742       0.985
1.Comp_adv~0 |      0.983      0.070   -0.247   0.805        0.854       1.130
 1.Have_IP_0 |      0.876      0.078   -1.489   0.137        0.736       1.043
1.OO_D_edu~0 |      0.754      0.052   -4.114   0.000        0.659       0.863
OO_work_ex~0 |      0.979      0.004   -5.493   0.000        0.972       0.987
OO_age_own~0 |      1.005      0.003    1.485   0.138        0.998       1.012
OO_race_wh~0 |      0.908      0.087   -1.005   0.315        0.752       1.096
    log_time |      0.819      0.021   -7.939   0.000        0.780       0.860
------------------------------------------------------------------------------

mi xeq 0: svy:  cloglog _d LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0  OO_race_white_owner_0 log_time  if year<2011  ,   eform
nocons cformat(%6.3f)  sformat(%6.3f) nolstretch

Number of strata   =         6                  Number of obs      =     15275
Number of PSUs     =      2872                  Population size    = 342091.18
                                                 Design df          =      2866
                                                 F(  10,   2857)    =    501.33
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
          _d |     exp(b)   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  LnAssets_0 |      0.968      0.009   -3.600   0.000        0.952       0.985
1.Home_Bas~0 |      0.905      0.060   -1.490   0.136        0.794       1.032
1.Sole_Pro~0 |      0.861      0.059   -2.185   0.029        0.753       0.985
1.Comp_adv~0 |      0.984      0.066   -0.232   0.816        0.862       1.124
 1.Have_IP_0 |      0.881      0.074   -1.504   0.133        0.747       1.039
1.OO_D_edu~0 |      0.766      0.050   -4.105   0.000        0.674       0.870
OO_work_ex~0 |      0.980      0.004   -5.480   0.000        0.973       0.987
OO_age_own~0 |      1.005      0.003    1.466   0.143        0.998       1.011
OO_race_wh~0 |      0.911      0.083   -1.022   0.307        0.763       1.089
    log_time |      0.809      0.019   -8.924   0.000        0.773       0.848
------------------------------------------------------------------------------
```

```
*Multiply Imputed (MI)
mi estimate : svy:  logit _d LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0  OO_race_white_owner_0 log_time if year<2011  ,  or   nocons
cformat(%6.3f)  sformat(%6.3f) nolstretch
mi estimate,hr  cformat(%6.3f)  sformat(%6.3f) nolstretch
```

```
Multiple-imputation estimates                   Imputations        =         5
Survey: Logistic regression                     Number of obs      =     16656

Number of strata  =         6                   Population size    = 373603.39
Number of PSUs    =      3140
                                                Average RVI        =    0.0074
                                                Largest FMI        =    0.0246
                                                Complete DF        =      3134
DF adjustment:   Small sample                   DF:     min        =   2106.25
                                                        avg        =   2886.57
                                                        max        =   3127.34
Model F test:       Equal FMI                   F(  10, 3115.9)    =    464.28
Within VCE type:   Linearized                   Prob > F           =    0.0000
------------------------------------------------------------------------------
          _d | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  LnAssets_0 |      0.964      0.009   -3.914   0.000        0.947       0.982
1.Home_Bas~0 |      0.908      0.061   -1.426   0.154        0.796       1.037
1.Sole_Pro~0 |      0.829      0.057   -2.700   0.007        0.724       0.950
1.Comp_adv~0 |      0.993      0.069   -0.101   0.920        0.867       1.137
 1.Have_IP_0 |      0.895      0.076   -1.304   0.192        0.759       1.057
1.OO_D_edu~0 |      0.772      0.051   -3.933   0.000        0.678       0.878
OO_work_ex~0 |      0.980      0.004   -5.571   0.000        0.973       0.987
OO_age_own~0 |      1.006      0.003    1.803   0.072        0.999       1.012
OO_race_wh~0 |      0.894      0.083   -1.216   0.224        0.746       1.071
    log_time |      0.818      0.020   -8.217   0.000        0.780       0.858
------------------------------------------------------------------------------

mi estimate : svy:  cloglog _d LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0  OO_race_white_owner_0 log_time  if year<2011  ,   eform
nocons cformat(%6.3f)  sformat(%6.3f) nolstretch
mi estimate,hr  cformat(%6.3f)  sformat(%6.3f) nolstretch

Multiple-imputation estimates                   Imputations        =         5
Survey: Complementary log-log regression        Number of obs      =     16656

Number of strata  =         6                   Population size    = 373603.39
Number of PSUs    =      3140
                                                Average RVI        =    0.0074
                                                Largest FMI        =    0.0250
                                                Complete DF        =      3134
DF adjustment:   Small sample                   DF:     min        =   2085.25
                                                        avg        =   2887.94
                                                        max        =   3127.51
Model F test:       Equal FMI                   F(  10, 3116.0)    =    541.90
Within VCE type:   Linearized                   Prob > F           =    0.0000

------------------------------------------------------------------------------
          _d | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  LnAssets_0 |      0.966      0.008   -3.944   0.000        0.950       0.983
1.Home_Bas~0 |      0.914      0.058   -1.421   0.156        0.806       1.035
1.Sole_Pro~0 |      0.837      0.055   -2.713   0.007        0.736       0.952
1.Comp_adv~0 |      0.994      0.065   -0.092   0.927        0.875       1.130
 1.Have_IP_0 |      0.899      0.072   -1.319   0.187        0.768       1.053
1.OO_D_edu~0 |      0.783      0.049   -3.920   0.000        0.693       0.885
OO_work_ex~0 |      0.981      0.003   -5.563   0.000        0.974       0.987
OO_age_own~0 |      1.006      0.003    1.781   0.075        0.999       1.012
OO_race_wh~0 |      0.899      0.078   -1.227   0.220        0.758       1.066
    log_time |      0.808      0.019   -9.246   0.000        0.773       0.846
------------------------------------------------------------------------------
```

```
*the estimated coefficients are similar to those for the Weibull model

* we can use glm TOO
*mi estimate : svy: glm _d LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0  OO_race_white_owner_0 log_time  if year<2011,  family( binomial)
link(  logit  )    nocons
*mi estimate,hr
*mi estimate : svy: glm _d LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0  OO_race_white_owner_0 log_time  if year<2011,  family( binomial)
link( cloglog )    nocons
*mi estimate,hr
```

Examples 5.12 Discrete Time Models: Time-Varying Covariates

```
*Multiply Imputed (MI) Longitudinal data in long format:Time-Varying Covariates
use Longitudinal_Long_MI_Survival_Ready,clear
mi xtset mprid year
gen LnAssets=ln( Assets+1)
mi svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
*Describing duration dependence :For a non-parametric baseline, we need to
create year dummy variables
tab year , gen(D)
*For ML estimation of the discrete time logistic model we use logit
mi xeq 0: svy:  logit _d LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner   OO_race_white_owner  i.D1 i.D2 i.D3 i.D4 i.D5 i.D6 i.D7 if
year<2011  ,  or   nocons cformat(%6.3f)  sformat(%6.3f) nolstretch
```

```
Number of strata   =         6                 Number of obs      =     15042
Number of PSUs     =      3091                 Population size    = 336498.71
                                                Design df          =      3085
                                                F(  16,   3070)    =    266.94
                                                Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
          _d | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      0.927      0.009   -7.877   0.000        0.910       0.945
1.Home_Based |      0.816      0.062   -2.689   0.007        0.703       0.946
1.Sole_Pro~p |      0.816      0.064   -2.580   0.010        0.700       0.952
1.Comp_adv~e |      0.786      0.060   -3.176   0.002        0.677       0.912
   1.Have_IP |      0.918      0.091   -0.864   0.388        0.757       1.114
1.OO_D_edu~r |      0.816      0.060   -2.745   0.006        0.706       0.944
OO_work_ex~r |      0.982      0.004   -4.317   0.000        0.974       0.990
OO_age_owner |      1.001      0.004    0.384   0.701        0.994       1.009
OO_race_wh~r |      1.078      0.116    0.693   0.488        0.872       1.332
        1.D1 |      0.369      0.077   -4.794   0.000        0.245       0.554
        1.D2 |      0.467      0.101   -3.534   0.000        0.306       0.712
        1.D3 |      0.307      0.069   -5.253   0.000        0.198       0.477
        1.D4 |      0.254      0.059   -5.890   0.000        0.161       0.401
        1.D5 |      0.242      0.057   -6.029   0.000        0.152       0.384
        1.D6 |      0.221      0.053   -6.302   0.000        0.138       0.353
        1.D7 |      0.231      0.054   -6.294   0.000        0.146       0.364
------------------------------------------------------------------------------
```

```
*For ML estimation of the discrete time complementary log-log model we use
cloglog

mi xeq 0: svy:  cloglog _d LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner   OO_race_white_owner  i.D1 i.D2 i.D3 i.D4 i.D5 i.D6 i.D7 if
year<2011  ,   eform    nocons cformat(%6.3f)  sformat(%6.3f) nolstretch



Number of strata   =         6                 Number of obs      =     15042
Number of PSUs     =      3091                 Population size    = 336498.71
                                               Design df          =      3085
                                               F(  16,   3070)    =    304.20
                                               Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
          _d |     exp(b)   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      0.933      0.008   -7.944   0.000        0.917       0.949
1.Home_Based |      0.827      0.059   -2.659   0.008        0.719       0.951
1.Sole_Pro~p |      0.826      0.061   -2.573   0.010        0.714       0.956
1.Comp_adv~e |      0.797      0.057   -3.151   0.002        0.693       0.918
   1.Have_IP |      0.919      0.086   -0.905   0.366        0.764       1.104
1.OO_D_edu~r |      0.826      0.058   -2.732   0.006        0.719       0.947
OO_work_ex~r |      0.983      0.004   -4.311   0.000        0.975       0.991
OO_age_owner |      1.001      0.003    0.368   0.713        0.994       1.008
OO_race_wh~r |      1.071      0.109    0.675   0.500        0.877       1.309
        1.D1 |      0.323      0.063   -5.819   0.000        0.220       0.472
        1.D2 |      0.403      0.081   -4.536   0.000        0.272       0.597
        1.D3 |      0.272      0.057   -6.182   0.000        0.180       0.411
        1.D4 |      0.226      0.049   -6.793   0.000        0.147       0.347
        1.D5 |      0.216      0.048   -6.912   0.000        0.140       0.334
        1.D6 |      0.199      0.045   -7.158   0.000        0.128       0.309
        1.D7 |      0.207      0.046   -7.165   0.000        0.135       0.319
------------------------------------------------------------------------------
```

```
*Multiply Imputed (MI)

mi estimate :svy: logit _d LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner   OO_race_white_owner  i.D1 i.D2 i.D3 i.D4 i.D5 i.D6 i.D7 if
year<2011  ,  or   nocons cformat(%6.3f)  sformat(%6.3f) nolstretch
mi estimate,hr  cformat(%6.3f)  sformat(%6.3f) nolstretch

Multiple-imputation estimates                   Imputations        =         5
Survey: Logistic regression                     Number of obs      =     16656

Number of strata  =         6                   Population size    = 373603.39
Number of PSUs    =      3140
                                                Average RVI        =    0.0041
                                                Largest FMI        =    0.0214
                                                Complete DF        =      3134
DF adjustment:   Small sample                   DF:     min        =   2283.70
                                                        avg        =   3000.68
                                                        max        =   3130.28
Model F test:       Equal FMI                   F(  16, 3129.1)    =    301.76
Within VCE type:   Linearized                   Prob > F           =    0.0000

------------------------------------------------------------------------------
          _d | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      0.931      0.008   -8.311   0.000        0.915       0.947
1.Home_Based |      0.887      0.059   -1.806   0.071        0.779       1.010
1.Sole_Pro~p |      0.813      0.056   -3.019   0.003        0.711       0.930
1.Comp_adv~e |      0.810      0.054   -3.153   0.002        0.711       0.924
   1.Have_IP |      0.971      0.081   -0.356   0.722        0.823       1.144
1.OO_D_edu~r |      0.806      0.052   -3.328   0.001        0.710       0.915
OO_work_ex~r |      0.982      0.004   -4.895   0.000        0.975       0.989
OO_age_owner |      1.003      0.003    1.022   0.307        0.997       1.010
OO_race_wh~r |      0.945      0.087   -0.611   0.542        0.789       1.132
        1.D1 |      0.360      0.067   -5.528   0.000        0.250       0.517
        1.D2 |      0.449      0.086   -4.184   0.000        0.308       0.653
        1.D3 |      0.333      0.067   -5.495   0.000        0.225       0.493
        1.D4 |      0.383      0.077   -4.770   0.000        0.258       0.568
        1.D5 |      0.307      0.064   -5.669   0.000        0.204       0.462
        1.D6 |      0.284      0.060   -5.942   0.000        0.187       0.430
        1.D7 |      0.269      0.055   -6.425   0.000        0.180       0.401
------------------------------------------------------------------------------




mi estimate : svy:  cloglog _d LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner   OO_race_white_owner  i.D1 i.D2 i.D3 i.D4 i.D5 i.D6 i.D7 if
year<2011  ,   eform    nocons cformat(%6.3f)  sformat(%6.3f) nolstretch
mi estimate,hr  cformat(%6.3f)  sformat(%6.3f) nolstretch
```

```
Multiple-imputation estimates                   Imputations        =         5
Survey: Complementary log-log regression        Number of obs      =     16656

Number of strata  =         6                   Population size    = 373603.39
Number of PSUs    =      3140
                                                Average RVI        =    0.0041
                                                Largest FMI        =    0.0214
                                                Complete DF        =      3134
DF adjustment:   Small sample                   DF:     min        =   2283.34
                                                        avg        =   2999.23
                                                        max        =   3130.21
Model F test:       Equal FMI                   F(  16, 3129.1)    =    352.88
Within VCE type:   Linearized                   Prob > F           =    0.0000

------------------------------------------------------------------------------
          _d | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      0.936      0.007   -8.400   0.000        0.922       0.951
1.Home_Based |      0.896      0.056   -1.764   0.078        0.793       1.012
1.Sole_Pro~p |      0.824      0.053   -3.007   0.003        0.727       0.935
1.Comp_adv~e |      0.822      0.052   -3.126   0.002        0.727       0.929
   1.Have_IP |      0.970      0.077   -0.379   0.705        0.831       1.134
1.OO_D_edu~r |      0.817      0.050   -3.302   0.001        0.725       0.921
OO_work_ex~r |      0.983      0.003   -4.877   0.000        0.977       0.990
OO_age_owner |      1.003      0.003    0.994   0.320        0.997       1.009
OO_race_wh~r |      0.950      0.082   -0.597   0.550        0.802       1.125
        1.D1 |      0.313      0.054   -6.779   0.000        0.223       0.438
        1.D2 |      0.384      0.068   -5.410   0.000        0.271       0.543
        1.D3 |      0.291      0.054   -6.627   0.000        0.202       0.420
        1.D4 |      0.331      0.062   -5.925   0.000        0.230       0.477
        1.D5 |      0.269      0.052   -6.765   0.000        0.184       0.394
        1.D6 |      0.251      0.050   -6.997   0.000        0.170       0.369
        1.D7 |      0.238      0.045   -7.526   0.000        0.164       0.346
------------------------------------------------------------------------------

*Describing duration dependence :we can use log(time)

gen log_time=log(year)
mi xeq 0: svy:  logit _d LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner   OO_race_white_owner  log_time if year<2011  ,  or   nocons
cformat(%6.3f)  sformat(%6.3f) nolstretch
```

```
Number of strata   =         6                  Number of obs      =     15042
Number of PSUs     =      3091                  Population size    = 336498.71
                                                 Design df          =      3085
                                                 F(  10,   3076)    =    402.68
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
          _d | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      0.925      0.009   -8.235   0.000        0.907       0.942
1.Home_Based |      0.819      0.064   -2.561   0.010        0.704       0.954
1.Sole_Pro~p |      0.819      0.066   -2.468   0.014        0.699       0.960
1.Comp_adv~e |      0.839      0.064   -2.312   0.021        0.723       0.974
   1.Have_IP |      0.920      0.092   -0.836   0.403        0.757       1.119
1.OO_D_edu~r |      0.805      0.061   -2.863   0.004        0.694       0.934
OO_work_ex~r |      0.982      0.004   -4.171   0.000        0.974       0.990
OO_age_owner |      0.999      0.004   -0.373   0.709        0.991       1.006
OO_race_wh~r |      1.084      0.120    0.727   0.467        0.872       1.347
    log_time |      0.872      0.024   -4.992   0.000        0.826       0.920
------------------------------------------------------------------------------

mi xeq 0: svy:  cloglog _d LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner   OO_race_white_owner  log_time  if year<2011  ,   eform    nocons
cformat(%6.3f)  sformat(%6.3f) nolstretch

Survey: Complementary log-log regression

Number of strata   =         6                  Number of obs      =     15042
Number of PSUs     =      3091                  Population size    = 336498.71
                                                 Design df          =      3085
                                                 F(  10,   3076)    =    458.78
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
          _d |     exp(b)   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      0.930      0.008   -8.357   0.000        0.914       0.946
1.Home_Based |      0.830      0.061   -2.532   0.011        0.718       0.959
1.Sole_Pro~p |      0.829      0.063   -2.445   0.015        0.714       0.964
1.Comp_adv~e |      0.847      0.061   -2.310   0.021        0.735       0.975
   1.Have_IP |      0.923      0.088   -0.844   0.399        0.766       1.112
1.OO_D_edu~r |      0.815      0.059   -2.841   0.005        0.708       0.939
OO_work_ex~r |      0.983      0.004   -4.178   0.000        0.975       0.991
OO_age_owner |      0.999      0.004   -0.382   0.702        0.992       1.006
OO_race_wh~r |      1.079      0.113    0.724   0.469        0.878       1.326
    log_time |      0.858      0.022   -5.958   0.000        0.815       0.902
------------------------------------------------------------------------------

*Multiply Imputed (MI)
mi estimate : svy:  logit _d LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner   OO_race_white_owner  log_time if year<2011  ,  or   nocons
cformat(%6.3f)  sformat(%6.3f) nolstretch
mi estimate,hr  cformat(%6.3f)  sformat(%6.3f) nolstretch
```

```
Multiple-imputation estimates                   Imputations        =         5
Survey: Logistic regression                     Number of obs      =     16656
Number of strata  =         6                   Population size    = 373603.39
Number of PSUs    =      3140
                                                Average RVI        =    0.0062
                                                Largest FMI        =    0.0205
                                                Complete DF        =      3134
DF adjustment:   Small sample                   DF:     min        =   2333.24
                                                        avg        =   2942.10
                                                        max        =   3130.72
Model F test:       Equal FMI                   F(  10, 3120.4)    =    459.47
Within VCE type:   Linearized                   Prob > F           =    0.0000
------------------------------------------------------------------------------
          _d | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      0.930      0.008   -8.420   0.000        0.914       0.946
1.Home_Based |      0.891      0.060   -1.705   0.088        0.781       1.017
1.Sole_Pro~p |      0.815      0.057   -2.929   0.003        0.711       0.935
1.Comp_adv~e |      0.845      0.056   -2.537   0.011        0.741       0.962
   1.Have_IP |      0.969      0.082   -0.370   0.712        0.821       1.144
1.OO_D_edu~r |      0.799      0.053   -3.403   0.001        0.703       0.909
OO_work_ex~r |      0.982      0.004   -4.812   0.000        0.975       0.989
OO_age_owner |      1.002      0.003    0.689   0.491        0.996       1.009
OO_race_wh~r |      0.946      0.089   -0.591   0.554        0.787       1.137
    log_time |      0.875      0.021   -5.570   0.000        0.835       0.917
------------------------------------------------------------------------------


mi estimate : svy:  cloglog _d LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner   OO_race_white_owner  log_time  if year<2011  ,   eform    nocons
cformat(%6.3f)  sformat(%6.3f) nolstretch
mi estimate,hr  cformat(%6.3f)  sformat(%6.3f) nolstretch

Multiple-imputation estimates                   Imputations        =         5
Survey: Complementary log-log regression        Number of obs      =     16656

Number of strata  =         6                   Population size    = 373603.39
Number of PSUs    =      3140
                                                Average RVI        =    0.0062
                                                Largest FMI        =    0.0204
                                                Complete DF        =      3134
DF adjustment:   Small sample                   DF:     min        =   2336.32
                                                        avg        =   2942.06
                                                        max        =   3130.71
Model F test:       Equal FMI                   F(  10, 3120.5)    =    537.67
Within VCE type:   Linearized                   Prob > F           =    0.0000
------------------------------------------------------------------------------
          _d | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      0.935      0.007   -8.546   0.000        0.921       0.950
1.Home_Based |      0.900      0.057   -1.656   0.098        0.795       1.020
1.Sole_Pro~p |      0.827      0.054   -2.904   0.004        0.727       0.940
1.Comp_adv~e |      0.853      0.054   -2.527   0.012        0.755       0.965
   1.Have_IP |      0.970      0.078   -0.380   0.704        0.829       1.135
1.OO_D_edu~r |      0.811      0.050   -3.377   0.001        0.718       0.916
OO_work_ex~r |      0.983      0.003   -4.811   0.000        0.976       0.990
OO_age_owner |      1.002      0.003    0.682   0.496        0.996       1.008
OO_race_wh~r |      0.951      0.084   -0.577   0.564        0.800       1.130
    log_time |      0.859      0.019   -6.826   0.000        0.822       0.897
------------------------------------------------------------------------------
```

```
* we can use glm TOO
*mi estimate : svy: glm _d LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0  OO_race_white_owner_0 log_time  if year<2011,  family( binomial)
link(  logit  )    nocons
*mi estimate,hr
*mi estimate : svy: glm _d LnAssets_0 i.Home_Based_0 i.Sole_Proprietorship_0
i.Comp_advantage_0 i.Have_IP_0 i.OO_D_education_owner_0  OO_work_exp_owner_0
OO_age_owner_0  OO_race_white_owner_0 log_time  if year<2011,  family( binomial)
link( cloglog )    nocons
*mi estimate,hr
```

5.7 Multinomial Logit Response Models Approach to Competing Risks:

The binary response models can be extended to accommodate competing risks. Assume there are *m* distinct types of events (outcomes) of interest (reasons for exit) indexed by $j \in \{1,2,\ldots,m\}$; let $\mathbf{x}$ be a vector of covariates, $f(t)$ the probability density function, and $S(t)$ the survival function. The maximum likelihood function for the full sample can be written as:

$$L = \prod_{i=1}^{n}\prod_{j=1}^{m} f_j\left(t_i \middle| x_{ij}, \beta_j\right) S_j\left(t_i \middle| x_{ij}, \beta_j\right)$$

where *n* is the number of observations at risk of facing *m* distinct types of events of interest.

Let d_{ij} be an indicator variable that takes a value of 1 if the businesses exited due to event of interest type *j*, and 0 otherwise (when d_{ij} = 0, the observation is right censored). Integrating d_{ij} into the likelihood function yields:

$$L = \prod_{i=1}^{n}\prod_{j=1}^{m} f_j\left(t_i \middle| x_{ij}, \beta_j\right)^{d_{ij}} S_j\left(t_i \middle| x_{ij}, \beta_j\right)^{1-d_{ij}}$$

The above likelihood function for the full sample is partitioned into m sub-contributions, where failures due to risks other than m are treated as right-censored. Therefore, the likelihood function indicates that we need to estimate m binary response models where all events other than m are treated as randomly censored. While estimating separate binary response models for each type of event yields unbiased estimators, it could result in a loss of efficiency. A natural extension of the logit model that accommodates competing risks is the multinomial logit model. For m possible events, the multinomial logit estimates $m-1$ logit models to obtain parameter estimates on the cause-specific hazards. Under the multinomial logit model, the cause-specific hazard λ_j is

$$\lambda_j = P_{tj} = \frac{e^{\mathbf{x}'\boldsymbol{\beta}_j}}{\sum_{j=1}^{k} e^{\mathbf{x}'\boldsymbol{\beta}_k}}$$

where P_{tj} is the conditional probability that an event of type j occurs to business *i* at time *t*, given that the business didn't have any type of events prior to *t*.

Examples 5.13 Competing Risks: Time-Varying Covariates

```
*Multiply Imputed (MI) Longitudinal data in long format:Time-Varying Covariates
use Longitudinal_Long_MI_Survival_Ready,clear
mi xtset mprid year
gen LnAssets=ln( Assets+1)
mi svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
*Describing duration dependence :For a non-parametric baseline, we need to
create year dummy variables
tab year , gen(D)
*For ML estimation of the discrete time logistic model we use logit
*Estimating separate binary response models for each type of event, by treating
all other events as censored observations
* Studying Sold Businesses only.
gen outcome=_d
replace outcome=0 if Competing!=1
mi xeq 0: svy:  logit outcome LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner   OO_race_white_owner  i.D1 i.D2 i.D3 i.D4 i.D5 i.D6 i.D7 if
year<2011  ,  or   nocons cformat(%6.3f)  sformat(%6.3f) nolstretch

Number of strata   =         6                 Number of obs      =     15042
Number of PSUs     =      3091                 Population size    = 336498.71
                                                Design df          =      3085
                                                F(  16,   3070)    =    122.75
                                                Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
     outcome | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      1.069      0.037    1.908   0.056        0.998       1.144
1.Home_Based |      0.422      0.105   -3.465   0.001        0.259       0.687
1.Sole_Pro~p |      0.903      0.227   -0.406   0.684        0.552       1.477
1.Comp_adv~e |      1.184      0.265    0.753   0.451        0.763       1.835
   1.Have_IP |      1.057      0.264    0.224   0.823        0.649       1.724
1.OO_D_edu~r |      1.092      0.234    0.412   0.681        0.718       1.662
OO_work_ex~r |      0.966      0.014   -2.381   0.017        0.938       0.994
OO_age_owner |      1.010      0.012    0.853   0.394        0.987       1.035
OO_race_wh~r |      0.872      0.273   -0.437   0.662        0.472       1.611
        1.D1 |      0.008      0.005   -7.196   0.000        0.002       0.029
        1.D2 |      0.008      0.006   -6.957   0.000        0.002       0.032
        1.D3 |      0.005      0.004   -7.370   0.000        0.001       0.020
        1.D4 |      0.003      0.002   -8.512   0.000        0.001       0.010
        1.D5 |      0.006      0.004   -7.786   0.000        0.002       0.022
        1.D6 |      0.005      0.004   -7.016   0.000        0.001       0.023
        1.D7 |      0.003      0.002   -7.708   0.000        0.001       0.014
------------------------------------------------------------------------------
```

```
*Multinomial logit estimates, all events
mi xeq 0: svy:  mlogit Competing LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner   OO_race_white_owner  i.D1 i.D2 i.D3 i.D4 i.D5 i.D6 i.D7 if
year<2011  ,   nocons cformat(%6.3f)  sformat(%6.3f) nolstretch
mi xeq 0: svy:  mlogit, rrr   cformat(%6.3f)  sformat(%6.3f) nolstretch



Survey: Multinomial logistic regression

Number of strata   =         6                  Number of obs      =     15042
Number of PSUs     =      3091                  Population size    = 336498.71
                                                 Design df          =      3085
                                                 F(  64,   3022)    =   7949.07
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
   Competing |        RRR   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
0__No_Event  |  (base outcome)
-------------+----------------------------------------------------------------
1__Sold      |
    LnAssets |      1.060      0.037    1.648   0.099        0.989       1.135
1.Home_Based |      0.417      0.104   -3.507   0.000        0.256       0.680
1.Sole_Pro~p |      0.887      0.223   -0.479   0.632        0.542       1.451
1.Comp_adv~e |      1.156      0.258    0.647   0.518        0.745       1.791
   1.Have_IP |      1.050      0.262    0.195   0.845        0.644       1.711
1.OO_D_edu~r |      1.071      0.229    0.321   0.749        0.704       1.630
OO_work_ex~r |      0.965      0.014   -2.458   0.014        0.937       0.993
OO_age_owner |      1.010      0.012    0.855   0.393        0.987       1.035
OO_race_wh~r |      0.878      0.275   -0.415   0.678        0.475       1.622
        1.D1 |      0.009      0.006   -6.835   0.000        0.002       0.036
        1.D2 |      0.011      0.007   -6.558   0.000        0.003       0.042
        1.D3 |      0.006      0.004   -7.046   0.000        0.001       0.025
        1.D4 |      0.003      0.002   -8.190   0.000        0.001       0.013
        1.D5 |      0.007      0.005   -7.463   0.000        0.002       0.026
        1.D6 |      0.007      0.005   -6.743   0.000        0.002       0.028
        1.D7 |      0.004      0.003   -7.412   0.000        0.001       0.017
-------------+----------------------------------------------------------------
2__Merged    |
    LnAssets |      0.956      0.047   -0.924   0.356        0.869       1.052
1.Home_Based |      0.475      0.168   -2.106   0.035        0.237       0.950
1.Sole_Pro~p |      0.517      0.213   -1.602   0.109        0.231       1.159
1.Comp_adv~e |      1.229      0.470    0.539   0.590        0.581       2.600
   1.Have_IP |      0.880      0.372   -0.301   0.763        0.384       2.018
1.OO_D_edu~r |      1.047      0.361    0.132   0.895        0.533       2.057
OO_work_ex~r |      1.014      0.018    0.758   0.448        0.979       1.050
OO_age_owner |      0.956      0.019   -2.313   0.021        0.920       0.993
OO_race_wh~r |      1.458      0.805    0.683   0.494        0.494       4.306
        1.D1 |      0.042      0.034   -3.893   0.000        0.009       0.208
        1.D2 |      0.041      0.036   -3.664   0.000        0.007       0.227
        1.D3 |      0.045      0.039   -3.600   0.000        0.008       0.244
        1.D4 |      0.068      0.058   -3.143   0.002        0.013       0.363
        1.D5 |      0.032      0.036   -3.055   0.002        0.004       0.293
        1.D6 |      0.031      0.032   -3.301   0.001        0.004       0.243
        1.D7 |      0.033      0.029   -3.938   0.000        0.006       0.181
-------------+----------------------------------------------------------------
```

```
3__Tempora~d |
    LnAssets |      0.888      0.061   -1.726   0.084        0.777       1.016
1.Home_Based |      3.976      4.328    1.268   0.205        0.470      33.614
1.Sole_Pro~p |      2.758      2.087    1.340   0.180        0.625      12.165
1.Comp_adv~e |      0.252      0.258   -1.347   0.178        0.034       1.874
   1.Have_IP |      0.000      0.000  -43.976   0.000        0.000       0.000
1.OO_D_edu~r |      0.672      0.434   -0.616   0.538        0.189       2.382
OO_work_ex~r |      1.018      0.035    0.528   0.597        0.952       1.089
OO_age_owner |      0.970      0.023   -1.297   0.195        0.927       1.016
OO_race_wh~r |      0.672      0.507   -0.526   0.599        0.153       2.952
        1.D1 |      0.000      0.000  -14.094   0.000        0.000       0.000
        1.D2 |      0.000      0.000  -13.274   0.000        0.000       0.000
        1.D3 |      0.000      0.000  -13.378   0.000        0.000       0.000
        1.D4 |      0.000      0.000  -13.674   0.000        0.000       0.000
        1.D5 |      0.000      0.000  -13.383   0.000        0.000       0.000
        1.D6 |      0.000      0.000  -13.500   0.000        0.000       0.000
        1.D7 |      0.039      0.069   -1.823   0.068        0.001       1.279
-------------+------------------------------------------------------------------
4__Out_of_~s |
    LnAssets |      0.916      0.009   -8.832   0.000        0.899       0.934
1.Home_Based |      0.911      0.075   -1.126   0.260        0.776       1.071
1.Sole_Pro~p |      0.831      0.070   -2.204   0.028        0.705       0.980
1.Comp_adv~e |      0.734      0.060   -3.781   0.000        0.625       0.861
   1.Have_IP |      0.898      0.098   -0.982   0.326        0.724       1.113
1.OO_D_edu~r |      0.777      0.062   -3.149   0.002        0.664       0.909
OO_work_ex~r |      0.983      0.004   -3.845   0.000        0.974       0.992
OO_age_owner |      1.002      0.004    0.574   0.566        0.995       1.010
OO_race_wh~r |      1.087      0.125    0.726   0.468        0.868       1.362
        1.D1 |      0.326      0.073   -5.014   0.000        0.210       0.505
        1.D2 |      0.423      0.098   -3.702   0.000        0.268       0.667
        1.D3 |      0.277      0.067   -5.285   0.000        0.172       0.446
        1.D4 |      0.228      0.058   -5.846   0.000        0.138       0.374
        1.D5 |      0.206      0.053   -6.147   0.000        0.125       0.341
        1.D6 |      0.188      0.049   -6.443   0.000        0.113       0.313
        1.D7 |      0.212      0.053   -6.178   0.000        0.130       0.347
------------------------------------------------------------------------------
```

```
*Multiply Imputed (MI)
* Studying Sold Businesses only.
mi estimate: svy:  logit outcome LnAssets  i.Home_Based  i.Sole_Proprietorship
i.Comp_advantage  i.Have_IP  i.OO_D_education_owner   OO_work_exp_owner
OO_age_owner   OO_race_white_owner  i.D1 i.D2 i.D3 i.D4 i.D5 i.D6 i.D7 if
year<2011  ,  or   nocons cformat(%6.3f)  sformat(%6.3f) nolstretch
mi estimate,hr  cformat(%6.3f)  sformat(%6.3f) nolstretch



Multiple-imputation estimates                   Imputations        =         5
Survey: Logistic regression                     Number of obs      =     16656

Number of strata  =         6                   Population size    = 373603.39
Number of PSUs    =      3140
                                                Average RVI        =    0.0061
                                                Largest FMI        =    0.0302
                                                Complete DF        =      3134
DF adjustment:   Small sample                   DF:     min        =   1818.40
                                                        avg        =   2852.63
                                                        max        =   3131.33
Model F test:       Equal FMI                   F(  16, 3125.6)    =    133.24
Within VCE type:   Linearized                   Prob > F           =    0.0000

------------------------------------------------------------------------------
     outcome | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      1.067      0.035    1.961   0.050        1.000       1.139
1.Home_Based |      0.394      0.095   -3.866   0.000        0.245       0.632
1.Sole_Pro~p |      0.823      0.203   -0.790   0.430        0.508       1.335
1.Comp_adv~e |      1.211      0.267    0.868   0.386        0.786       1.866
   1.Have_IP |      1.027      0.248    0.109   0.914        0.640       1.648
1.OO_D_edu~r |      1.135      0.237    0.605   0.545        0.753       1.710
OO_work_ex~r |      0.969      0.013   -2.307   0.021        0.943       0.995
OO_age_owner |      1.009      0.012    0.813   0.416        0.987       1.033
OO_race_wh~r |      0.831      0.246   -0.627   0.531        0.465       1.484
        1.D1 |      0.008      0.005   -7.564   0.000        0.002       0.028
        1.D2 |      0.009      0.006   -7.194   0.000        0.003       0.033
        1.D3 |      0.005      0.004   -7.700   0.000        0.001       0.020
        1.D4 |      0.003      0.002   -8.464   0.000        0.001       0.012
        1.D5 |      0.006      0.004   -8.161   0.000        0.002       0.020
        1.D6 |      0.006      0.004   -7.415   0.000        0.002       0.023
        1.D7 |      0.003      0.002   -8.042   0.000        0.001       0.013
------------------------------------------------------------------------------
```

```
*Multinomial logit estimates, all events
mi estimate: svy:  mlogit Competing LnAssets  i.Home_Based
i.Sole_Proprietorship  i.Comp_advantage  i.Have_IP  i.OO_D_education_owner
OO_work_exp_owner    OO_age_owner   OO_race_white_owner  i.D1 i.D2 i.D3 i.D4
i.D5 i.D6 i.D7 if year<2011  ,   nocons cformat(%6.3f)  sformat(%6.3f)
nolstretch
mi estimate, rrr   cformat(%6.3f)  sformat(%6.3f) nolstretch

Multiple-imputation estimates                   Imputations        =         5
Survey: Multinomial logistic regression         Number of obs      =     16656

Number of strata  =         6                   Population size    = 373603.39
Number of PSUs    =      3140
                                                Average RVI        =    0.0029
                                                Largest FMI        =    0.0308
                                                Complete DF        =      3134
DF adjustment:   Small sample                   DF:     min        =   1787.50
                                                        avg        =   3037.38
                                                        max        =   3132.00
Model F test:       Equal FMI                   F(  64, 3131.7)    =   9046.52
Within VCE type:   Linearized                   Prob > F           =    0.0000

------------------------------------------------------------------------------
   Competing |        RRR   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
0__No_Event  |  (base outcome)
-------------+----------------------------------------------------------------
1__Sold      |
    LnAssets |      1.058      0.036     1.675   0.094        0.990       1.130
1.Home_Based |      0.392      0.094    -3.885   0.000        0.244       0.629
1.Sole_Pro~p |      0.807      0.199    -0.870   0.384        0.497       1.309
1.Comp_adv~e |      1.183      0.261     0.762   0.446        0.768       1.822
   1.Have_IP |      1.024      0.247     0.098   0.922        0.638       1.643
1.OO_D_edu~r |      1.109      0.232     0.495   0.621        0.736       1.671
OO_work_ex~r |      0.967      0.013    -2.403   0.016        0.941       0.994
OO_age_owner |      1.010      0.012     0.834   0.404        0.987       1.033
OO_race_wh~r |      0.826      0.245    -0.643   0.520        0.462       1.478
        1.D1 |      0.010      0.006    -7.171   0.000        0.003       0.035
        1.D2 |      0.012      0.008    -6.765   0.000        0.003       0.043
        1.D3 |      0.007      0.005    -7.333   0.000        0.002       0.025
        1.D4 |      0.004      0.003    -8.065   0.000        0.001       0.015
        1.D5 |      0.007      0.005    -7.775   0.000        0.002       0.025
        1.D6 |      0.007      0.005    -7.082   0.000        0.002       0.028
        1.D7 |      0.004      0.003    -7.700   0.000        0.001       0.017
-------------+----------------------------------------------------------------
2__Merged    |
    LnAssets |      0.980      0.060    -0.334   0.738        0.870       1.104
1.Home_Based |      0.464      0.156    -2.286   0.022        0.240       0.897
1.Sole_Pro~p |      0.557      0.220    -1.480   0.139        0.257       1.209
1.Comp_adv~e |      1.355      0.492     0.838   0.402        0.665       2.761
   1.Have_IP |      0.979      0.379    -0.054   0.957        0.459       2.091
1.OO_D_edu~r |      1.115      0.360     0.339   0.735        0.592       2.100
OO_work_ex~r |      1.010      0.018     0.563   0.574        0.976       1.045
OO_age_owner |      0.962      0.019    -1.920   0.055        0.925       1.001
OO_race_wh~r |      1.529      0.807     0.805   0.421        0.543       4.304
        1.D1 |      0.021      0.022    -3.576   0.000        0.002       0.173
        1.D2 |      0.019      0.022    -3.437   0.001        0.002       0.184
        1.D3 |      0.025      0.028    -3.270   0.001        0.003       0.227
        1.D4 |      0.033      0.038    -2.948   0.003        0.003       0.320
        1.D5 |      0.020      0.026    -2.980   0.003        0.001       0.261
        1.D6 |      0.015      0.019    -3.237   0.001        0.001       0.189
        1.D7 |      0.025      0.023    -4.070   0.000        0.004       0.149
```

```
-------------+----------------------------------------------------------------
3__Tempora~d |
     LnAssets |      0.925      0.065   -1.103   0.270        0.806       1.062
1.Home_Based |      5.957      6.407    1.659   0.097        0.723      49.076
1.Sole_Pro~p |      3.728      2.531    1.939   0.053        0.985      14.109
1.Comp_adv~e |      0.493      0.384   -0.909   0.363        0.107       2.266
   1.Have_IP |      0.000      0.000  -45.297   0.000        0.000       0.000
1.OO_D_edu~r |      0.676      0.387   -0.684   0.494        0.220       2.078
OO_work_ex~r |      1.005      0.029    0.190   0.849        0.951       1.064
OO_age_owner |      0.972      0.017   -1.633   0.103        0.939       1.006
OO_race_wh~r |      0.513      0.327   -1.046   0.296        0.147       1.793
        1.D1 |      0.000      0.000  -15.018   0.000        0.000       0.000
        1.D2 |      0.000      0.000  -14.294   0.000        0.000       0.000
        1.D3 |      0.000      0.000  -14.430   0.000        0.000       0.000
        1.D4 |      0.000      0.000  -14.521   0.000        0.000       0.000
        1.D5 |      0.000      0.000  -14.395   0.000        0.000       0.000
        1.D6 |      0.000      0.000  -14.660   0.000        0.000       0.000
        1.D7 |      0.023      0.038   -2.315   0.021        0.001       0.563
-------------+----------------------------------------------------------------
4__Out_of_~s |
     LnAssets |      0.921      0.008   -9.260   0.000        0.906       0.938
1.Home_Based |      0.989      0.070   -0.155   0.877        0.860       1.137
1.Sole_Pro~p |      0.830      0.060   -2.583   0.010        0.721       0.956
1.Comp_adv~e |      0.765      0.054   -3.767   0.000        0.666       0.880
   1.Have_IP |      0.958      0.087   -0.472   0.637        0.801       1.145
1.OO_D_edu~r |      0.767      0.053   -3.830   0.000        0.670       0.879
OO_work_ex~r |      0.983      0.004   -4.475   0.000        0.975       0.990
OO_age_owner |      1.004      0.003    1.209   0.227        0.997       1.011
OO_race_wh~r |      0.939      0.090   -0.655   0.513        0.778       1.134
        1.D1 |      0.321      0.063   -5.824   0.000        0.219       0.471
        1.D2 |      0.405      0.082   -4.455   0.000        0.272       0.603
        1.D3 |      0.304      0.065   -5.605   0.000        0.201       0.461
        1.D4 |      0.362      0.077   -4.762   0.000        0.238       0.550
        1.D5 |      0.277      0.061   -5.800   0.000        0.180       0.428
        1.D6 |      0.255      0.057   -6.091   0.000        0.164       0.396
        1.D7 |      0.249      0.054   -6.390   0.000        0.163       0.382
------------------------------------------------------------------------------
```

6.1 Longitudinal Data Analysis

The KFS is a true longitudinal study with a very special feature—it is a single-cohort panel (a type of single indefinite life panels) that tracks the same group of businesses from a common starting point (birth) and records a wide range of information about them over time.[1] Like most longitudinal panel data, the KFS provides the researcher with an opportunity to analyze individual-level change, and it allows for the aggregation of data for businesses over time by examining the occurrence of special events, frequency, timing, and duration, controlling for omitted variables and heterogeneity, and utilizing dynamic panel models. Unlike most longitudinal panel data, the longitudinal nature of the KFS has greater analytical potential to analyze change over time because it remains a single-cohort panel and, thus, can avoid any problems of population composition changes.

The measure issue of using conventional statistical analysis methods with longitudinal data is that those methods assume that observations are independent. Meanwhile, longitudinal data observations by nature are dependent, which is simply because measurements taken from the same observation (business) tend to be similar (highly correlated) over time. Thus, any statistical analysis methods used with longitudinal data should take into account the covariance structure to draw valid statistical inferences.

If we use conventional statistical analysis methods with longitudinal data, then the variance of the nontime-varying predictor variables will be underestimated and the variance of the time-varying predictor variables will be overestimated.

The second layer of complexity comes from the fact that we deal with longitudinal data from a complex sample design. With survey data, the assumptions that observations are independent of each other are violated. Our previous analysis shows that stratification and finite population corrections affect the standard errors, confidence intervals, and significance tests, but they do not affect the points and coefficient estimates. On the other hand, weighting affects population estimates, standard errors, confidence intervals, and significance tests.

6.2 Regression Commands in Stata

Stata can be used to analyze survey data by using the svy commands or non-svy commands that allow pweights. Non-svy commands ignore any observations with zero weights (zero weights affect only the variance computation), do not count for stratification and finite population corrections, do not allow for subpopulation analysis,

[1] However, the "unit of analysis for the KFS design is the sampled business so that if the same business changed ownership from one reporting period to another, it would remain in the sample" (Kauffman Firm Survey Fifth Follow-up Methodology Report); data for businesses that sold or merged were not collected.

and non-svy commands post-estimation commands do not adjust the test statistics correctly for the sample design.

In the KFS longitudinal data, we do not have zero weights, but in KFS cross-sectional data, we have zero weights; thus, we should set those zero to missing. Theoretically speaking, zero sampling weights are not possible. By setting zero weights to missing, the difference in standard errors will be due to ignoring stratification and finite population corrections. In return, this makes standard errors smaller; ignoring them gives us more conservative estimates of standard errors.

If you have to use non-svy commands, make sure to request a clustered sandwich estimator for the variance (vce (cluster clustvar)). The vce (cluster clustvar) option relaxes the usual requirement that the observations be independent. If the vce (cluster clustvar) option is not allowed, then use the vce (robust) option. We highly recommend using the survey commands when available.

All svy commands imply robust standard errors; thus, svy prefix commands will not accept the vce() option. By default, svy commands allow pweights.

The following table shows the Stata regressions command that supports svy commands or non-svy commands with pweight options, as well as if they support imputed data estimation command.

	Estimation commands support					
Description	command	svy	mi estimate & svy	[pweights] Option	mi estimate & [pweight]	Type of weight L: longitudinal CS: cross sectional
Random-effects and population-averaged cloglog models	xtcloglog	No	No	Yes: with pa	Yes	L
Population-averaged panel-data models using GEE	xtgee	No	No	Yes	Yes	L
Fixed-effects, random-effects, & population-averaged logit models	xtlogit	No	No	Yes: with pa	Yes	L
Multilevel mixed-effects linear regression	xtmixed	No	No	Yes: with re	Yes	L
Fixed-effects, random-effects, & population-averaged negative binomial models	xtnbreg	No	No	Yes: with pa	Yes	L
Fixed-effects, random-effects, & population-averaged Poisson models	xtpoisson	No	No	Yes: with pa	Yes	L
Random-effects and population-averaged probit models	xtprobit	No	No	Yes: with pa	Yes	L
Fixed-, between- and random-effects, and population-averaged linear models	xtreg	No	No	Yes : with fe,pa	Yes	L
Linear regression with a large dummy-variable set	areg	No	No	Yes	No	CS / L
Alternative-specific conditional logit (McFadden's choice)	asclogit	No	No	Yes	No	CS / L
Alternative-specific multinomial probit regression	asmprobit	No	No	Yes	No	CS / L
Bivariate probit regression	biprobit	Yes	No	Yes	No	CS / L
Conditional (fixed-effects) logistic regression	clogit	Yes	Yes	Yes	Yes	CS / L
Complementary log-log regression	cloglog	Yes	Yes	Yes	Yes	CS / L
Constrained linear regression	cnsreg	Yes	Yes	Yes	Yes	CS / L
Stochastic frontier models	frontier	No	No	Yes	No	CS / L
Generalized linear models	glm	Yes	Yes	Yes	Yes	CS / L
Generalized method of moments estimation	gmm	No	No	Yes	No	CS / L
Generalized negative binomial regression	gnbreg	Yes	Yes	Yes	Yes	CS / L
Heckman selection model	heckman	Yes	No	Yes	No	CS / L
Probit model with sample selection	heckprob	Yes	No	Yes	No	CS / L
Heteroskedastic probit regression	hetprob	Yes	No	Yes	No	CS / L

Estimation commands support						
Description	command	svy	mi estimate & svy	[pweights] Option	mi estimate & [pweight]	Type of weight L: longitudinal CS: cross sectional
Interval regression	intreg	Yes	No	Yes	No	CS / L
Probit model with endogenous regressors	ivprobit	Yes	No	Yes	No	CS / L
Single-equation instrumental-variables regression	ivregress	Yes	No	Yes	No	CS / L
Tobit model with continuous endogenous regressors	ivtobit	Yes	No	Yes	No	CS / L
Logistic regression, reporting odds ratios	logistic	Yes	Yes	Yes	Yes	CS / L
Logistic regression, reporting coefficients	logit	Yes	Yes	Yes	Yes	CS / L
Maximum likelihood estimation	ml	No	No	Yes	No	CS / L
Multinomial (polytomous) logistic regression	mlogit	Yes	Yes	Yes	Yes	CS / L
Multinomial probit regression	mprobit	Yes	Yes	Yes	Yes	CS / L
Negative binomial regression	nbreg	Yes	Yes	Yes	Yes	CS / L
Nonlinear least-squares estimation	nl	Yes	No	Yes	No	CS / L
Nested logit regression	nlogit	No	No	Yes	No	CS / L
Estimation of nonlinear systems of equations	nlsur	No	No	Yes	No	CS / L
Ordered logistic regression	ologit	Yes	Yes	Yes	Yes	CS / L
Ordered probit regression	oprobit	Yes	Yes	Yes	Yes	CS / L
Poisson regression	poisson	Yes	Yes	Yes	Yes	CS / L
Probit regression	probit	Yes	Yes	Yes	Yes	CS / L
Linear regression	regress	Yes	Yes	Yes	Yes	CS / L
Receiver operating characteristic (ROC) regression	rocreg	No	No	Yes	No	CS / L
Rank-ordered logistic regression	rologit	No	No	Yes	No	CS / L
Skewed logistic regression	scobit	Yes	No	Yes	No	CS / L
Structural equation models	sem	Yes	No	Yes	No	CS / L
Stereotype logistic regression	slogit	Yes	No	Yes	No	CS / L
Cox proportional hazards model	stcox	Yes	Yes	Yes	Yes	CS / L
Parametric survival models	streg	Yes	Yes	Yes	Yes	CS / L
Truncated negative binomial regression	tnbreg	Yes	No	Yes	No	CS / L
Tobit regression	tobit	Yes	No	Yes	No	CS / L
Truncated Poisson regression	tpoisson	Yes	No	Yes	No	CS / L
Treatment-effects model	treatreg	Yes	No	Yes	No	CS / L

	Estimation commands support					
Description	command	svy	mi estimate & svy	[pweights] Option	mi estimate & [pweight]	Type of weight L: longitudinal CS: cross sectional
Truncated regression	truncreg	Yes	Yes	Yes	Yes	CS / L
Zero-inflated negative binomial regression	zinb	Yes	No	Yes	No	CS / L
Zero-inflated Poisson regression	zip	Yes	No	Yes	No	CS / L

6.3 XT Commands in Stata

Stata has a series of commands (xt commands) that provide tools for analyzing panel data. Because we are working with survey data, the following commands are important:

1. xtset: declare data to be panel data
2. xtdescribe: describe pattern of xt data
3. xtdata: faster specification searches with xt data

xtdata produces a transformed dataset of the variables specified in varlist to work with the regress command to produce fixed effects, between and random effect estimators.

The xtsum that summarizes xt data is not useful with a survey data because it will only work with unweighted data. We created our own non-survey xtsum command to work with the KFS data. The command name is FR_xtsum; it has the same format as xtsum command, but we need to add the pweight option.

```
use Longitudinal_Long_MI_Long_L2,clear
merge m:1 mprid   using Kfs8_enclave_14oct13, keepusing(wgt_ini_0 )
keep if _merge==3
*Declare data to be panel data
mi xtset mprid year
gen LnAssets=ln( Assets+1)
*Declare survey design for dataset
mi svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
* Debt injections  to capital injections
* Farhat, Cotei 2014, draft
egen capital_injection=rowtotal(Debt  Equity )
gen Debt_inj= Debt / capital_injection
gen Equity_inj= Equity / capital_injection
recode Debt_inj Equity_inj (.=0) if master!=0 & capital_injection==0

egen capital=rowtotal(Equity_AllYrs Debt_Owed )
gen tdca= Debt_Owed / capital
recode tdca (.=.a) if  Debt_Owed==.a & Equity_AllYrs==.a
recode tdca (.=0) if master!=0

* Replace few missing in PO by OO
bysort master mprid (year):replace             PO_hours      =
       OO_hours_owner[1]   if     PO_hours       ==.   &      OO_hours_owner[1]
       <.
bysort master mprid (year):replace             PO_work_exp   =
       OO_work_exp_owner[1]       if     PO_work_exp    ==.   &
       OO_work_exp_owner   <.
bysort master mprid (year):replace             PO_age_owner  =
       OO_age_owner[1]     if     PO_age_owner   ==.   &      OO_age_owner <.
bysort master mprid (year):replace             PO_emp        =
       round(OO_emp_owner[1] ,1)  if     PO_emp         ==.   &
       round(OO_emp_owner[1] ,1)  <.
bysort master mprid (year):replace             PO_oth_bus_owner    =
       round(OO_oth_bus_owner[1] ,1)     if     PO_oth_bus_owner     ==.   &
       round(OO_oth_bus_owner[1] ,1)     <.
bysort master mprid (year):replace             PO_hisp_origin      =
       round(OO_hisp_origin_owner ,1)    if     PO_hisp_origin       ==.   &
       round(OO_hisp_origin_owner[1] ,1)<.
bysort master mprid (year):replace             PO_race_amind_owner        =
       round(OO_race_amind_owner[1] ,1) if     PO_race_amind_owner         ==.
       &      round(OO_race_amind_owner[1] ,1) <.
```

```
bysort master mprid (year):replace             PO_race_asian_owner        =
      round(OO_race_asian_owner[1] ,1) if      PO_race_asian_owner         ==.
      &      round(OO_race_asian_owner[1] ,1) <.
bysort master mprid (year):replace             PO_race_black_owner        =
      round(OO_race_black_owner[1] ,1) if      PO_race_black_owner         ==.
      &      round(OO_race_black_owner[1] ,1) <.
bysort master mprid (year):replace             PO_race_nathaw_owner       =
      round(OO_race_nathaw_owner[1] ,1)if      PO_race_nathaw_owner        ==.
      &      round(OO_race_nathaw_owner[1] ,1)<.
bysort master mprid (year):replace             PO_race_other_owner        =
      round(OO_race_other_owner[1] ,1) if      PO_race_other_owner         ==.
      &      round(OO_race_other_owner[1] ,1) <.
bysort master mprid (year):replace             PO_race_white_owner        =
      round(OO_race_white_owner[1] ,1) if      PO_race_white_owner         ==.
      &      round(OO_race_white_owner[1] ,1) <.
bysort master mprid (year):replace             PO_native_born      =
      round(OO_native_born_owner[1] ,1)if      PO_native_born       ==.   &
      round(OO_native_born_owner[1] ,1)<.
bysort master mprid (year):replace             PO_us_cit    =
      round(OO_us_cit_owner[1] ,1)     if      PO_us_cit     ==.   &
      round(OO_us_cit_owner[1] ,1)     <.
bysort master mprid (year):replace             PO_education =
      round(OO_education_owner[1] ,1)  if      PO_education  ==.   &
      round(OO_education_owner[1] ,1)  <.
bysort master mprid (year):replace             PO_gender    =
      round(OO_gender_owner[1],1)      if      PO_gender     ==.   &
      round(OO_gender_owner[1],1)      <.



*Describe pattern of xt data
mi xeq 0: xtdescribe

     Freq.  Percent    Cum. |  Pattern
 -------------------------+----------
     1630     51.91   51.91 |  11111111
      303      9.65   61.56 |  1.......
      283      9.01   70.57 |  11......
      238      7.58   78.15 |  1111....
      224      7.13   85.29 |  111.....
      164      5.22   90.51 |  11111...
      153      4.87   95.38 |  111111..
      145      4.62  100.00 |  1111111.
 -------------------------+----------
     3140    100.00         |  XXXXXXXX
```

```
*Summarize xt data
mi xeq 1:FR_xtsum tdca Debt_inj Equity_inj LnAssets Net_Profit Home_Based  Have_IP
OO_D_education_owner      OO_work_exp_owner      OO_race_white_owner   PO_gender
f13_trade_fin f19_res_dev [pweight=wgt_7_long]

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
tdca     overall |  .2486094   .3364657          0          1 |     N =   18286
         between |             .2616562          0          1 |     n =    3140
         within  |             .2416033  -.6063906   1.123609 | T-bar = 5.82357
                 |                                            |
Debt_inj overall |  .4110795   .4435556          0          1 |     N =   18286
         between |             .3197957          0          1 |     n =    3140
         within  |             .3173692  -.4639205    1.28608 | T-bar = 5.82357
                 |                                            |
Equity~j overall |  .3001113   .4058216          0          1 |     N =   18286
         between |             .2909057          0          1 |     n =    3140
         within  |             .3209941  -.5748887   1.175111 | T-bar = 5.82357
                 |                                            |
LnAssets overall |  9.835462   3.378436          0    20.8632 |     N =   18286
         between |             2.916259          0   18.28905 |     n =    3140
         within  |             2.029399  -3.796146    20.3238 | T-bar = 5.82357
                 |                                            |
Net_Pr~t overall |  73570.36    4774561  -1.50e+07   5.00e+08 |     N =   18286
         between |              1435768   -6250000   6.26e+07 |     n =    3140
         within  |              4456708  -6.25e+07   4.37e+08 | T-bar = 5.82357
                 |                                            |
Home_B~d overall |  .5021481   .5000091          0          1 |     N =   18286
         between |             .4833637          0          1 |     n =    3140
         within  |             .1431235  -.3728519   1.377148 | T-bar = 5.82357
                 |                                            |
Have_IP  overall |  .1908688   .3929967          0          1 |     N =   18286
         between |             .3242967          0          1 |     n =    3140
         within  |             .2336701  -.6841312   1.065869 | T-bar = 5.82357
                 |                                            |
OO_D_e~r overall |   .469045   .4990545          0          1 |     N =   18286
         between |             .4798692          0          1 |     n =    3140
         within  |             .1434693   -.405955   1.344045 | T-bar = 5.82357
                 |                                            |
OO_wor~r overall |   11.9013   9.900631          0         60 |     N =   18286
         between |             9.837195          0         60 |     n =    3140
         within  |             1.561646   -11.8487    29.8263 | T-bar = 5.82357
                 |                                            |
O~whit~r overall |   .837747   .3581886          0          1 |     N =   18286
         between |             .3619703          0          1 |     n =    3140
         within  |             .0533101   -.037253   1.712747 | T-bar = 5.82357
                 |                                            |
PO_gen~r overall |  .7072223   .4550497          0          1 |     N =   18286
         between |             .4597272          0          1 |     n =    3140
         within  |                   0   .7072223   .7072223 | T-bar = 5.82357
                 |                                            |
f13_tr~n overall |  .2414841   .4279948          0          1 |     N =   18286
         between |             .3437899          0          1 |     n =    3140
         within  |             .2688988  -.6335159   1.116484 | T-bar = 5.82357
                 |                                            |
f19_re~v overall |  .1549057   .3618246          0          1 |     N =   18286
         between |             .2784021          0          1 |     n =    3140
         within  |              .252498  -.7200943   1.029906 | T-bar = 5.82357
```

FR_xtsum decomposes the variable x_{it} into a between (x_i) and within ($x_{it} - \bar{x}_i + \bar{\bar{x}}$), the global mean $\bar{\bar{x}}$ being added back in makes results comparable. The overall and within are calculated over 18,286 firm-years of data. The between is calculated over 3,140 firms. The average number of years a firm was observed is 5.82.

The reported standard deviation shows that the variation in tdeq across firms is nearly equal to that observed within a firm over time. Also, for the variables that do not vary over time, the within standard deviation will be zero

6.4 Linear Panel Models

In general, the assumption that observations are independent is not appropriate when dealing with longitudinal data, clustered data, and multilevel data. As a result of dependence among observations within and across groups, a special method of estimations was needed. Those methods include the following approaches:

1. Cluster-robust Standard Errors
2. Generalized Estimating Equations (FGLS) [population averaged models]
3. Fixed Effects Model
4. Random Effects Models

Let us introduce some notation for the analysis of panel (repeated) measures data. Our response variable is y_{it} for firm i at time t. x_{it} is a column vector of variables that vary both over firms (i) and over time (t).

The model to be fit is

$$y_{it} = \alpha + \beta x_{it} + \epsilon_{it} \text{ , for } i = 1,2,..n \text{ and } t = 1,2,..T.$$

As it stands, this model is a classical regression model. Panel models mainly differ in their assumptions on ϵ_{it}.

6.4.1 Pooled Regression

In a pooled OLS regression, the model to be fit is

$$y_{it} = \alpha + \beta x_{it} + \epsilon_{it} \text{ , for } i = 1,2,..n \text{ and } t = 1,2,..T.$$

$$\beta = (X'X)^{-1}X'y$$

Standard OLS assume that: $E(\epsilon_{it}) = 0$, $Cov(x_{it}, \epsilon_{is}) = 0$ for all t and s (Strict exogeneity), $Var(\epsilon_{it}) = \sigma^2$ (Homoscedasticity) and $Cov(\epsilon_{it}, \epsilon_{jt}) = 0 \; for \; i \neq j$, this assumption may be violated in the context of panel data, cluster samples, hierarchical data, repeated measures data, and other data with dependencies. As a result, standard errors will be too low.

In the context of panel data, firms are unlike one another—that is, they are heterogeneous ($\alpha_i \neq \alpha_j$); meanwhile, our model is homogeneous model ($\alpha_i = \alpha$) because ignoring heterogeneity may introduce bias into the model estimators.

While estimators may be unbiased, it will not be efficient to control for the dependency among observations; we can view the panel data as a special case of

clustered data where there is within-firm clustering so that errors are correlated over time for a given firm. Thus, we can overcome the problem of biased standard errors by using cluster-robust standard errors (Huber-White sandwich estimators), which allows for relaxing the assumption of independence within groups.

Examples 6.1 Cluster-Robust Standard Errors

```
*svy commands
*Robust Standard Errors
mi  xeq  0:svy:  reg  tdca      LnAssets  Net_Profit  i.Home_Based    i.Have_IP
OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 i.f13_trade_fin i.f19_res_dev, cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store rols

Number of strata   =         6                  Number of obs      =     12586
Number of PSUs     =      2951                  Population size    = 281145.73
                                                 Design df          =      2945
                                                 F(   9,   2937)    =         .
                                                 Prob > F           =         .
                                                 R-squared          =    0.0363

------------------------------------------------------------------------------
             |             Linearized
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |       0.015      0.002     9.242   0.000        0.011       0.018
  Net_Profit |      -0.000      0.000    -4.393   0.000       -0.000      -0.000
1.Home_Based |      -0.024      0.012    -2.060   0.040       -0.047      -0.001
   1.Have_IP |      -0.023      0.012    -1.964   0.050       -0.045      -0.000
OO_D_educa~r |      -0.040      0.011    -3.605   0.000       -0.062      -0.018
OO_work_ex~r |      -0.002      0.001    -3.324   0.001       -0.003      -0.001
OO_race_wh~r |      -0.010      0.016    -0.635   0.525       -0.042       0.021
   PO_gender |       0.005      0.013     0.384   0.701       -0.021       0.031
1.f13_trad~n |       0.048      0.012     3.961   0.000        0.024       0.072
1.f19_res_~v |       0.008      0.012     0.643   0.520       -0.016       0.032
       _cons |       0.178      0.023     7.624   0.000        0.132       0.224
------------------------------------------------------------------------------


* Check if residuals are correlated within - firms
preserve
mi xtset,clear
mi unset
keep if master==0
svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
svy: reg tdca   LnAssets Net_Profit i.Home_Based  i.Have_IP OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 i.f13_trade_fin i.f19_res_dev, cformat(%6.3f)  sformat(%6.3f)   nolstretch
 predict res, residuals
 keep mprid res year
 reshape wide res, i(mprid) j(year)
 pwcorr  res*,   sig
restore
```

```
             |  res2004  res2005  res2006  res2007  res2008  res2009  res2010
-------------+---------------------------------------------------------------
     res2004 |   1.0000
             |
     res2005 |   0.3570   1.0000
             |   0.0000
     res2006 |   0.3497   0.4278   1.0000
             |   0.0000   0.0000
     res2007 |   0.2750   0.3231   0.4567   1.0000
             |   0.0000   0.0000   0.0000
     res2008 |   0.3064   0.3428   0.4386   0.4632   1.0000
             |   0.0000   0.0000   0.0000   0.0000
     res2009 |   0.1876   0.3483   0.3880   0.4644   0.5210   1.0000
             |   0.0000   0.0000   0.0000   0.0000   0.0000
     res2010 |   0.1762   0.2729   0.3134   0.3462   0.4572   0.4388   1.0000
             |   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000
     res2011 |   0.1760   0.1958   0.3298   0.3248   0.3831   0.4595   0.4774
             |   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000

* Correlation of observations closer together in time is larger than that of
observations farther apart in time

mi  estimate  :svy:reg  tdca      LnAssets  Net_Profit  i.Home_Based    i.Have_IP
OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 i.f13_trade_fin i.f19_res_dev
mi estimate,    cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                   Imputations        =         5
Survey: Linear regression                       Number of obs      =     18286

Number of strata  =         6                   Population size    = 408495.43
Number of PSUs    =      3140
                                                Average RVI        =    0.0113
                                                Largest FMI        =    0.0225
                                                Complete DF        =      3134
DF adjustment:   Small sample                   DF:     min        =   2220.92
                                                        avg        =   2773.90
                                                        max        =   3129.52
Model F test:       Equal FMI                   F(   9, 3092.7)    =     28.02
Within VCE type:   Linearized                   Prob > F           =    0.0000

------------------------------------------------------------------------------
         tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
     LnAssets |       0.014      0.001   11.264   0.000        0.012       0.017
   Net_Profit |      -0.000      0.000   -3.875   0.000       -0.000      -0.000
1.Home_Based |      -0.025      0.010   -2.435   0.015       -0.045      -0.005
    1.Have_IP |      -0.014      0.010   -1.358   0.175       -0.034       0.006
OO_D_educa~r |      -0.038      0.010   -3.884   0.000       -0.056      -0.019
OO_work_ex~r |      -0.002      0.000   -4.769   0.000       -0.003      -0.001
OO_race_wh~r |      -0.005      0.014   -0.336   0.737       -0.032       0.023
    PO_gender |       0.007      0.011    0.622   0.534       -0.015       0.030
1.f13_trad~n |       0.051      0.011    4.743   0.000        0.030       0.072
1.f19_res_~v |       0.012      0.011    1.108   0.268       -0.009       0.033
       _cons |       0.155      0.019    7.947   0.000        0.117       0.193
------------------------------------------------------------------------------

*Non-svy commands
*Cluster-robust Standard Errors
mi   xeq   0:     reg   tdca       LnAssets   Net_Profit   i.Home_Based    i.Have_IP
OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner PO_gender   ///
```

```
 i.f13_trade_fin   i.f19_res_dev     [pweight=wgt_7_long],vce(   cluster   mprid)
cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store clusterols

Linear regression                                      Number of obs =   12586
                                                       F(  9,  2950) =       .
                                                       Prob > F      =       .
                                                       R-squared     =  0.0363
                                                       Root MSE      =  .33633

                             (Std. Err. adjusted for 2951 clusters in mprid)
------------------------------------------------------------------------------
             |               Robust
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |       0.015      0.002     9.242   0.000        0.011       0.018
  Net_Profit |      -0.000      0.000    -4.390   0.000       -0.000      -0.000
1.Home_Based |      -0.024      0.012    -2.058   0.040       -0.047      -0.001
   1.Have_IP |      -0.023      0.012    -1.964   0.050       -0.045      -0.000
OO_D_educa~r |      -0.040      0.011    -3.604   0.000       -0.062      -0.018
OO_work_ex~r |      -0.002      0.001    -3.324   0.001       -0.003      -0.001
OO_race_wh~r |      -0.010      0.016    -0.635   0.525       -0.042       0.021
   PO_gender |       0.005      0.013     0.384   0.701       -0.021       0.031
1.f13_trad~n |       0.048      0.012     3.961   0.000        0.024       0.072
1.f19_res_~v |       0.008      0.012     0.643   0.520       -0.016       0.032
       _cons |       0.178      0.023     7.623   0.000        0.132       0.224
------------------------------------------------------------------------------

mi  estimate  :  reg  tdca      LnAssets  Net_Profit  i.Home_Based    i.Have_IP
OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 i.f13_trade_fin i.f19_res_dev [pweight=wgt_7_long],vce( cluster mprid)
mi estimate,    cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                      Imputations     =          5
Linear regression                                  Number of obs   =      18286
                                                   Average RVI     =     0.0113
                                                   Largest FMI     =     0.0225
                                                   Complete DF     =       3139
DF adjustment:   Small sample                      DF:     min     =    2222.82
                                                           avg     =    2777.96
                                                           max     =    3134.52
Model F test:       Equal FMI                      F(   9, 3097.6) =      28.02
Within VCE type:       Robust                      Prob > F        =     0.0000

                            (Within VCE adjusted for 3140 clusters in mprid)
------------------------------------------------------------------------------
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |       0.014      0.001    11.264   0.000        0.012       0.017
  Net_Profit |      -0.000      0.000    -3.873   0.000       -0.000      -0.000
1.Home_Based |      -0.025      0.010    -2.434   0.015       -0.045      -0.005
   1.Have_IP |      -0.014      0.010    -1.358   0.175       -0.034       0.006
OO_D_educa~r |      -0.038      0.010    -3.883   0.000       -0.056      -0.019
OO_work_ex~r |      -0.002      0.000    -4.769   0.000       -0.003      -0.001
OO_race_wh~r |      -0.005      0.014    -0.336   0.737       -0.032       0.023
   PO_gender |       0.007      0.011     0.622   0.534       -0.015       0.030
1.f13_trad~n |       0.051      0.011     4.744   0.000        0.030       0.072
1.f19_res_~v |       0.012      0.011     1.107   0.268       -0.009       0.033
       _cons |       0.155      0.019     7.946   0.000        0.117       0.193
------------------------------------------------------------------------------

estimates table clusterols rols,stats(se  ) se(%8.5f)
```

```
-------------------------------------
    Variable | clusterols      rols
-------------+-----------------------
    LnAssets |  .01451436    .01451436
             |    0.00157      0.00157
  Net_Profit | -6.887e-10   -6.887e-10
             |    0.00000      0.00000
             |
  Home_Based |
           1 | -.02424888   -.02424888
             |    0.01178      0.01177
             |
     Have_IP |
           1 | -.02265174   -.02265174
             |    0.01153      0.01153
             |
OO_D_educa~r | -.04013449   -.04013449
             |    0.01114      0.01113
OO_work_ex~r | -.00184825   -.00184825
             |    0.00056      0.00056
OO_race_wh~r | -.01028954   -.01028954
             |    0.01619      0.01619
   PO_gender |  .00504712    .00504712
             |    0.01314      0.01314
             |
f13_trade_~n |
           1 |  .04840794    .04840794
             |    0.01222      0.01222
             |
 f19_res_dev |
           1 |  .00781436    .00781436
             |    0.01215      0.01215
             |
       _cons |  .17818248    .17818248
             |    0.02337      0.02337
-------------+-----------------------
          se |
-------------------------------------
                           legend: b/se
```

The svy commands and non-svy commands that allow pweights produce almost the same standard errors, yet we cannot generalize these results. In the above example, stratification has little impact on the standard errors; in other models, stratification could have a significant impact on the standard errors.

6.4.2 Generalized Estimating Equations (FGLS)

In generalized linear models, the model to be fit is

$$g\{E(y_{it})\} = \beta x_{it} \quad , y \sim F$$

where g( ) is called the link function and F is the distributional family. Substituting various definitions for g( ) and F results in an array of models.

The distributional family (F) can be gaussian, igaussian, bernoulli/binomial, Poisson, nbinomial, or gamma. The possible link functions (g) are identity(y=y), log (ln(y)) , logit

(ln{y/(1-y)}), probit, cloglog (ln{-ln(1-y)}),power(y^k), odds power ([{y/(1-y)}^k - 1]/k), negative binomial or reciprocal (1/y).

GEE models allow us to specify the assumed within-group correlation structure. Correlation structure can be exchangeable, independent, unstructured, autoregressive of order #, stationary of order #, or nonstationary of order #.

It is efficient to estimate the parameters using FGLS, but only when you know the true model of the variance-covariance matrix.

Examples 6.2 Population-Averaged Model

```
*Generalized Estimating Equations (GEE)
*GEE population-averaged model
mi   xeq   0:   xtgee   tdca       LnAssets   Net_Profit   i.Home_Based     i.Have_IP
OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 i.f13_trade_fin i.f19_res_dev [pweight=wgt_7_long] ,vce(robust)  cformat(%6.3f)
sformat(%6.3f)   nolstretch

GEE population-averaged model                   Number of obs      =     12586
Group variable:                       mprid     Number of groups   =      2951
Link:                                identity     Obs per group: min =         1
Family:                              Gaussian                    avg =       4.1
Correlation:                     exchangeable                    max =         8
                                                 Wald chi2(9)       =         .
Scale parameter:                     .1135563    Prob > chi2        =         .

                                  (Std. Err. adjusted for clustering on mprid)
------------------------------------------------------------------------------
             |               Robust
        tdca |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |       0.010      0.001     7.272   0.000        0.007       0.012
  Net_Profit |      -0.000      0.000    -3.486   0.000       -0.000      -0.000
1.Home_Based |      -0.031      0.011    -2.821   0.005       -0.053      -0.009
   1.Have_IP |      -0.007      0.010    -0.666   0.505       -0.027       0.013
OO_D_educa~r |      -0.036      0.011    -3.325   0.001       -0.057      -0.015
OO_work_ex~r |      -0.002      0.001    -3.628   0.000       -0.003      -0.001
OO_race_wh~r |      -0.004      0.016    -0.261   0.794       -0.036       0.027
   PO_gender |       0.005      0.013     0.377   0.706       -0.021       0.031
1.f13_trad~n |       0.026      0.010     2.665   0.008        0.007       0.045
1.f19_res_~v |       0.019      0.010     1.922   0.055       -0.000       0.039
       _cons |       0.235      0.023    10.206   0.000        0.190       0.280
------------------------------------------------------------------------------

*estat wcor      estimated matrix of the within-group correlations
estat wcor  , format(%6.3f)
Estimated within-mprid correlation matrix R:

      |     c1      c2      c3      c4      c5      c6      c7      c8
------+------------------------------------------------------------------
   r1 |  1.000
   r2 |  0.350   1.000
   r3 |  0.350   0.350   1.000
   r4 |  0.350   0.350   0.350   1.000
   r5 |  0.350   0.350   0.350   0.350   1.000
   r6 |  0.350   0.350   0.350   0.350   0.350   1.000
   r7 |  0.350   0.350   0.350   0.350   0.350   0.350   1.000
   r8 |  0.350   0.350   0.350   0.350   0.350   0.350   0.350   1.000
```

```
mi  estimate  :    xtgee  tdca      LnAssets  Net_Profit  i.Home_Based    i.Have_IP
OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 i.f13_trade_fin i.f19_res_dev [pweight=wgt_7_long],vce(robust)
mi estimate,    cformat(%6.3f)  sformat(%6.3f)   nolstretch



Multiple-imputation estimates                   Imputations        =         5
GEE population-averaged model                   Number of obs      =     18286

Group variable:                       mprid     Number of groups   =      3140
Link:                              identity     Obs per group: min =         1
Family:                            Gaussian                    avg =       5.6
Correlation:                   exchangeable                    max =         8
Scale parameter:                         x2
                                                Average RVI        =    0.0223
                                                Largest FMI        =    0.0587
DF adjustment:                 Large sample     DF:     min        =   1225.17
                                                        avg        =  18816.78
                                                        max        =  47072.78
Model F test:                     Equal FMI     F(   9,56753.0)    =     18.93
Within VCE type:                     Robust     Prob > F           =    0.0000

                                (Within VCE adjusted for clustering on mprid)
------------------------------------------------------------------------------
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |       0.009      0.001     8.306   0.000        0.007       0.012
  Net_Profit |      -0.000      0.000    -2.585   0.010       -0.000      -0.000
1.Home_Based |      -0.033      0.010    -3.376   0.001       -0.053      -0.014
   1.Have_IP |      -0.002      0.009    -0.219   0.826       -0.020       0.016
OO_D_educa~r |      -0.031      0.009    -3.318   0.001       -0.050      -0.013
OO_work_ex~r |      -0.002      0.000    -4.896   0.000       -0.003      -0.001
OO_race_wh~r |      -0.001      0.014    -0.085   0.932       -0.029       0.027
   PO_gender |       0.006      0.012     0.535   0.593       -0.017       0.029
1.f13_trad~n |       0.029      0.009     3.381   0.001        0.012       0.046
1.f19_res_~v |       0.023      0.009     2.485   0.013        0.005       0.040
       _cons |       0.213      0.020    10.630   0.000        0.173       0.252
------------------------------------------------------------------------------
```

```
*Equivalently
mi   xeq   0:   xtreg   tdca       LnAssets   Net_Profit   i.Home_Based     i.Have_IP
OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 i.f13_trade_fin    i.f19_res_dev    [pweight=wgt_7_long],pa           vce(robust)
cformat(%6.3f)  sformat(%6.3f)   nolstretch
estimates store xtregpa
estimates table rgee xtregpa,stats(se  ) se(%8.5f)
```

```
GEE population-averaged model                   Number of obs      =     12586
Group variable:                       mprid     Number of groups   =      2951
Link:                              identity     Obs per group: min =         1
Family:                            Gaussian                    avg =       4.1
Correlation:                   exchangeable                    max =         8
                                                Wald chi2(9)       =         .
Scale parameter:                   .1135563     Prob > chi2        =         .

                                (Std. Err. adjusted for clustering on mprid)
------------------------------------------------------------------------------
             |               Robust
        tdca |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |       0.010      0.001     7.272   0.000        0.007       0.012
  Net_Profit |      -0.000      0.000    -3.486   0.000       -0.000      -0.000
1.Home_Based |      -0.031      0.011    -2.821   0.005       -0.053      -0.009
   1.Have_IP |      -0.007      0.010    -0.666   0.505       -0.027       0.013
OO_D_educa~r |      -0.036      0.011    -3.325   0.001       -0.057      -0.015
OO_work_ex~r |      -0.002      0.001    -3.628   0.000       -0.003      -0.001
OO_race_wh~r |      -0.004      0.016    -0.261   0.794       -0.036       0.027
   PO_gender |       0.005      0.013     0.377   0.706       -0.021       0.031
1.f13_trad~n |       0.026      0.010     2.665   0.008        0.007       0.045
1.f19_res_~v |       0.019      0.010     1.922   0.055       -0.000       0.039
       _cons |       0.235      0.023    10.206   0.000        0.190       0.280
------------------------------------------------------------------------------
mi  estimate  :  xtreg  tdca      LnAssets  Net_Profit  i.Home_Based    i.Have_IP
OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 i.f13_trade_fin i.f19_res_dev [pweight=wgt_7_long],pa   vce(robust)
mi estimate,    cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                   Imputations        =         5
Population-averaged linear regression           Number of obs      =     18286

Group variable:                       mprid     Number of groups   =      3140
Link:                              identity     Obs per group: min =         1
Family:                            Gaussian                    avg =       5.6
Correlation:                   exchangeable                    max =         8
Scale parameter:                         x2
                                                Average RVI        =    0.0223
                                                Largest FMI        =    0.0587
DF adjustment:                 Large sample     DF:     min        =   1225.17
                                                        avg        =  18816.78
                                                        max        =  47072.78
Model F test:                     Equal FMI     F(   9,56753.0)    =     18.93
Within VCE type:                     Robust     Prob > F           =    0.0000

                                (Within VCE adjusted for clustering on mprid)
------------------------------------------------------------------------------
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |       0.009      0.001     8.306   0.000        0.007       0.012
  Net_Profit |      -0.000      0.000    -2.585   0.010       -0.000      -0.000
1.Home_Based |      -0.033      0.010    -3.376   0.001       -0.053      -0.014
   1.Have_IP |      -0.002      0.009    -0.219   0.826       -0.020       0.016
OO_D_educa~r |      -0.031      0.009    -3.318   0.001       -0.050      -0.013
OO_work_ex~r |      -0.002      0.000    -4.896   0.000       -0.003      -0.001
OO_race_wh~r |      -0.001      0.014    -0.085   0.932       -0.029       0.027
   PO_gender |       0.006      0.012     0.535   0.593       -0.017       0.029
1.f13_trad~n |       0.029      0.009     3.381   0.001        0.012       0.046
1.f19_res_~v |       0.023      0.009     2.485   0.013        0.005       0.040
       _cons |       0.213      0.020    10.630   0.000        0.173       0.252
------------------------------------------------------------------------------
```

Default correlation structure is "exchangeable," which means that the correlations between the dependent variables at different points in time are all the same. Gee gives different coefficients relative to pooled OLS.

Using the option vce(robust) gives us standard errors that are not sensitive to the correlation structure.

6.4.3 Fixed Effects Model

In fixed effects models, the model to be fit is

$$y_{it} = \beta x_{it} + \epsilon_{it} \quad \text{, for } i = 1,2,..n \text{ and } t = 1,2,..T.$$

$$\epsilon_{it} = \mu_i + \nu_{it}$$

$$y_{it} = \mu_i + \beta x_{it} + \nu_{it}$$

The idiosyncratic error (ν_{it}) varies over firms and time, where $E(\nu_{it}) = 0$, $Var(\nu_{it}) = \sigma_\nu^2$, $(\nu_{it}, \nu_{jt}) = 0 \; for \; i \neq j$, $Cov(x_{it}, \nu_{is}) = 0$ for all t and s and it is allowed to have $Cov(\mu_i, \nu_{it}) \neq 0$ or $Cov(\mu_i, x_{it}) \neq 0$.

This fixed effects approach takes μ_i to be a firm-specific error term in the regression model. The firm-specific error does not change over time and every firm has a fixed value on this variable (thus, the name "fixed effects"). To identify the model parameters, we assume that $\sum_{i=1}^{n} \mu_i = 0$.

The fixed effects model controls for all time-invariant differences between the firms. Thus, we are controlling for heterogeneity among firms and any other time-invariant unobservable variables, and as a result, we cannot investigate time-invariant variables. In addition, fixed effects will not work well with data for which within-firm variation is minimal. This is because fixed effects estimates use only within-firm differences and discard any information about differences between firms.

In a two-way fixed effects model, in addition to the firm-specific unobserved constants, we will have a time-specific constants (time effects).

$$\epsilon_{it} = \mu_i + \lambda_t + \nu_{it}$$

$$y_{it} = \mu_i + \lambda_t + \beta x_{it} + \nu_{it}$$

The time-specific error (λ_t) does vary over time but not among firms. We assume that $\sum_{i=1}^{n} \mu_i = \sum_{t=1}^{T} \lambda_t = 0$

We need to be careful about the meaning of time effects in this model. The KFS is single cohort panel data, and thus it is impossible to separate age, cohort, and period effects (age = year (period) - year of birth (cohort)).

Examples 6.3 One-Way Fixed Effects

```
mi   xeq   0:xtreg   tdca       LnAssets   Net_Profit   i.Home_Based     i.Have_IP
OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 i.f13_trade_fin  i.f19_res_dev   [pweight=wgt_7_long],  fe  i(mprid)  vce(robust)
cformat(%6.3f)  sformat(%6.3f)   nolstretch

note: PO_gender omitted because of collinearity

Fixed-effects (within) regression               Number of obs      =     12586
Group variable: mprid                           Number of groups   =      2951

R-sq:  within  = 0.0032                         Obs per group: min =         1
       between = 0.0143                                        avg =       4.3
       overall = 0.0119                                        max =         8

                                                F(8,2950)          =         .
corr(u_i, Xb)  = 0.0103                         Prob > F           =         .

                              (Std. Err. adjusted for 2951 clusters in mprid)
------------------------------------------------------------------------------
             |               Robust
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |       0.005      0.002     2.878   0.004        0.001       0.008
  Net_Profit |      -0.000      0.000    -1.398   0.162       -0.000       0.000
1.Home_Based |      -0.014      0.023    -0.606   0.545       -0.060       0.032
   1.Have_IP |       0.011      0.013     0.862   0.389       -0.014       0.036
OO_D_educa~r |      -0.008      0.027    -0.311   0.756       -0.060       0.044
OO_work_ex~r |      -0.001      0.002    -0.433   0.665       -0.005       0.003
OO_race_wh~r |       0.069      0.079     0.878   0.380       -0.085       0.223
   PO_gender |       0.000  (omitted)
1.f13_trad~n |      -0.001      0.011    -0.065   0.948       -0.023       0.021
1.f19_res_~v |       0.034      0.011     2.991   0.003        0.012       0.056
       _cons |       0.183      0.073     2.509   0.012        0.040       0.327
-------------+----------------------------------------------------------------
     sigma_u |       0.274
     sigma_e |       0.274
         rho |       0.516   (fraction of variance due to u_i)
------------------------------------------------------------------------------

mi  estimate  :xtreg  tdca      LnAssets  Net_Profit  i.Home_Based    i.Have_IP
OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 i.f13_trade_fin i.f19_res_dev  [pweight=wgt_7_long], fe i(mprid) vce(robust)
mi estimate,    cformat(%6.3f)  sformat(%6.3f)   nolstretch
```

```
Multiple-imputation estimates                   Imputations       =          5
Fixed-effects (within) regression               Number of obs     =      18286

Group variable: mprid                           Number of groups  =       3140
                                                Obs per group: min =          1
                                                               avg =        5.8
                                                               max =          8

                                                Average RVI       =   342.6428
                                                Largest FMI       =     0.1930
                                                Complete DF       =       3139
DF adjustment:   Small sample                   DF:     min       =     117.86
                                                        avg       =    1229.03
                                                        max       =    2927.91
Model F test:       Equal FMI                   F(   8, 2109.8)   =       3.90
Within VCE type:       Robust                   Prob > F          =     0.0001

                              (Within VCE adjusted for 3140 clusters in mprid)
------------------------------------------------------------------------------
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |       0.005      0.001     3.576   0.000        0.002       0.008
  Net_Profit |      -0.000      0.000    -1.108   0.268       -0.000       0.000
1.Home_Based |      -0.028      0.020    -1.357   0.175       -0.067       0.012
   1.Have_IP |       0.011      0.011     0.980   0.327       -0.011       0.032
OO_D_educa~r |       0.001      0.020     0.059   0.953       -0.039       0.041
OO_work_ex~r |       0.001      0.002     0.272   0.786       -0.003       0.004
OO_race_wh~r |      -0.016      0.053    -0.295   0.768       -0.121       0.089
   PO_gender |       0.000  (omitted)
1.f13_trad~n |       0.003      0.010     0.341   0.733       -0.016       0.023
1.f19_res_~v |       0.036      0.010     3.640   0.000        0.016       0.055
       _cons |       0.212      0.055     3.877   0.000        0.104       0.320
-------------+----------------------------------------------------------------
     sigma_u |       0.253
     sigma_e |       0.266
         rho |       0.475   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

```
mi   xeq   1:FR_xtsum   tdca        LnAssets   Net_Profit   Home_Based     Have_IP
OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 f13_trade_fin f19_res_dev [pweight=wgt_7_long]

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
tdca     overall |  .2486094   .3364657          0          1 |     N =   18286
         between |             .2616562          0          1 |     n =    3140
         within  |             .2416033  -.6063906   1.123609 | T-bar = 5.82357
                 |                                            |
LnAssets overall |  9.835462   3.378436          0    20.8632 |     N =   18286
         between |             2.916259          0   18.28905 |     n =    3140
         within  |             2.029399  -3.796146    20.3238 | T-bar = 5.82357
                 |                                            |
Net_Pr~t overall |  73570.36    4774561  -1.50e+07   5.00e+08 |     N =   18286
         between |              1435768   -6250000   6.26e+07 |     n =    3140
         within  |              4456708  -6.25e+07   4.37e+08 | T-bar = 5.82357
                 |                                            |
Home_B~d overall |  .5021481   .5000091          0          1 |     N =   18286
         between |             .4833637          0          1 |     n =    3140
         within  |             .1431235  -.3728519   1.377148 | T-bar = 5.82357
                 |                                            |
Have_IP  overall |  .1908688   .3929967          0          1 |     N =   18286
         between |             .3242967          0          1 |     n =    3140
         within  |             .2336701  -.6841312   1.065869 | T-bar = 5.82357
                 |                                            |
OO_D_e~r overall |   .469045   .4990545          0          1 |     N =   18286
         between |             .4798692          0          1 |     n =    3140
         within  |             .1434693   -.405955   1.344045 | T-bar = 5.82357
                 |                                            |
OO_wor~r overall |   11.9013   9.900631          0         60 |     N =   18286
         between |             9.837195          0         60 |     n =    3140
         within  |             1.561646   -11.8487    29.8263 | T-bar = 5.82357
                 |                                            |
O~whit~r overall |   .837747   .3581886          0          1 |     N =   18286
         between |             .3619703          0          1 |     n =    3140
         within  |             .0533101   -.037253   1.712747 | T-bar = 5.82357
                 |                                            |
PO_gen~r overall |  .7072223   .4550497          0          1 |     N =   18286
         between |             .4597272          0          1 |     n =    3140
         within  |                   0   .7072223   .7072223 | T-bar = 5.82357
                 |                                            |
f13_tr~n overall |  .2414841   .4279948          0          1 |     N =   18286
         between |             .3437899          0          1 |     n =    3140
         within  |             .2688988  -.6335159   1.116484 | T-bar = 5.82357
                 |                                            |
f19_re~v overall |  .1549057   .3618246          0          1 |     N =   18286
         between |             .2784021          0          1 |     n =    3140
         within  |              .252498  -.7200943   1.029906 | T-bar = 5.82357
```

Time-invariant variables will drop from the analysis. Meanwhile, variables with small within-firm variation relative to the between-firm variation are going to have very large standard errors. As a result, we do not have reliable estimates.

While the fixed effect method reduces bias, it does that at the expense of increased standard errors.

Examples 6.4 Two-Way Fixed Effects

```
mi   xeq   0:xtreg   tdca        LnAssets   Net_Profit   i.Home_Based     i.Have_IP
OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 i.f13_trade_fin   i.f19_res_dev   i.year   [pweight=wgt_7_long],   fe   i(mprid)
vce(robust)  cformat(%6.3f)  sformat(%6.3f)   nolstretch

Fixed-effects (within) regression               Number of obs      =     12586
Group variable: mprid                           Number of groups   =      2951

R-sq:  within  = 0.0415                         Obs per group: min =         1
       between = 0.0560                                        avg =       4.3
       overall = 0.0477                                        max =         8

                                                F(15,2950)         =         .
corr(u_i, Xb)  = 0.0514                         Prob > F           =         .

                              (Std. Err. adjusted for 2951 clusters in mprid)
------------------------------------------------------------------------------
             |               Robust
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |       0.007      0.002     4.285   0.000        0.004       0.010
  Net_Profit |       0.000      0.000     4.699   0.000        0.000       0.000
1.Home_Based |      -0.024      0.023    -1.066   0.286       -0.069       0.020
   1.Have_IP |       0.004      0.013     0.340   0.734       -0.020       0.029
OO_D_educa~r |      -0.013      0.026    -0.490   0.624       -0.064       0.038
OO_work_ex~r |      -0.001      0.002    -0.485   0.628       -0.005       0.003
OO_race_wh~r |       0.046      0.075     0.610   0.542       -0.102       0.193
   PO_gender |       0.000  (omitted)
1.f13_trad~n |      -0.002      0.011    -0.160   0.873       -0.023       0.020
1.f19_res_~v |       0.019      0.011     1.716   0.086       -0.003       0.041
             |
        year |
        2005 |      -0.105      0.011    -9.444   0.000       -0.127      -0.083
        2006 |      -0.085      0.011    -7.434   0.000       -0.108      -0.063
        2007 |      -0.093      0.013    -7.415   0.000       -0.118      -0.069
        2008 |      -0.107      0.013    -8.469   0.000       -0.131      -0.082
        2009 |      -0.130      0.013    -9.957   0.000       -0.156      -0.105
        2010 |      -0.161      0.014   -11.602   0.000       -0.188      -0.134
        2011 |      -0.174      0.014   -12.582   0.000       -0.201      -0.147
             |
       _cons |       0.289      0.071     4.086   0.000        0.150       0.428
-------------+----------------------------------------------------------------
     sigma_u |       0.268
     sigma_e |       0.268
         rho |       0.516   (fraction of variance due to u_i)
------------------------------------------------------------------------------

*Testing for time-fixed effects
 testparm i.year

 ( 1)  2005.year = 0
 ( 2)  2006.year = 0
 ( 3)  2007.year = 0
 ( 4)  2008.year = 0
 ( 5)  2009.year = 0
 ( 6)  2010.year = 0
 ( 7)  2011.year = 0

       F(  7,  2950) =   27.49
            Prob > F =    0.0000
```

```
mi    estimate    :xtreg    LnAssets    Net_Profit    i.Home_Based       i.Have_IP
OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 i.f13_trade_fin  i.f19_res_dev    i.year    [pweight=wgt_7_long],  fe  i(mprid)
vce(robust)
mi estimate,    cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                   Imputations        =         5
Fixed-effects (within) regression               Number of obs      =     18286

Group variable: mprid                           Number of groups   =      3140
                                                Obs per group: min =         1
                                                               avg =       5.8
                                                               max =         8

                                                Average RVI        =   34.8575
                                                Largest FMI        =    0.0474
                                                Complete DF        =      3139
                                                DF:     min        =   1148.73
                                                        avg        =   2772.81
DF adjustment:   Small sample                           max        =   3136.00
                                                F(   2,      .)    =         .
Within VCE type:       Robust                   Prob > F           =         .

                            (Within VCE adjusted for 3140 clusters in mprid)
------------------------------------------------------------------------------
    LnAssets |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Net_Profit |       0.000      0.000     1.500   0.134       -0.000       0.000
1.Home_Based |      -0.634      0.155    -4.081   0.000       -0.938      -0.329
   1.Have_IP |       0.326      0.097     3.362   0.001        0.136       0.516
OO_D_educa~r |       0.206      0.203     1.015   0.310       -0.192       0.605
OO_work_ex~r |      -0.022      0.016    -1.368   0.172       -0.053       0.009
OO_race_wh~r |       0.339      0.475     0.712   0.476       -0.594       1.271
   PO_gender |       0.000  (omitted)
1.f13_trad~n |       0.377      0.065     5.778   0.000        0.249       0.505
1.f19_res_~v |       0.497      0.093     5.345   0.000        0.315       0.680
             |
        year |
       2005  |       0.637      0.072     8.831   0.000        0.496       0.778
       2006  |       0.681      0.081     8.421   0.000        0.522       0.840
       2007  |       0.677      0.088     7.738   0.000        0.506       0.849
       2008  |       0.637      0.091     7.017   0.000        0.459       0.815
       2009  |       0.563      0.093     6.037   0.000        0.380       0.746
       2010  |       0.587      0.099     5.923   0.000        0.393       0.782
       2011  |       0.498      0.107     4.632   0.000        0.287       0.708
             |
       _cons |       9.292      0.422    22.005   0.000        8.464      10.120
-------------+----------------------------------------------------------------
     sigma_u |       2.768
     sigma_e |       2.210
         rho |       0.611   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Strong evidence of time/age effect.

6.4.3.1 Between and Within Groups

There is one more way to look at the above model:

$$y_{it} = \mu_i + \beta x_{it} + \nu_{it}$$

Average this equation over time for each *i* (between estimators):

$$\bar{y}_i = \mu_i + \beta_1 \bar{x}_i + \bar{\nu}_i$$

Subtract the second equation from the first for each t (within estimators / fixed effects estimators):

$$y_{it} - \bar{y}_i = \beta_2 (x_{it} - \bar{x}_i) + (\nu_{it} - \bar{\nu}_i)$$

These three equations provide the basis for estimating β.

Examples 6.5 Between and Within Groups

```
*Between and Within Groups

preserve
mi unset
keep if master==0
*Declare survey design for dataset
svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
gen wgt_7_long0=wgt_7_long
xtdata mprid tdca   LnAssets Net_Profit Home_Based  Have_IP OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 f13_trade_fin f19_res_dev , be   clear
merge    m:1    mprid            using    KFS8_L7_w1,    keepusing(wgt_7_long
sampleinfo_samplestrata_0)
keep if _merge==3
svyset  mprid  [pweight=wgt_7_long]
svy:regress tdca    LnAssets Net_Profit Home_Based  Have_IP OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 f13_trade_fin f19_res_dev  ,   cformat(%6.3f)  sformat(%6.3f)   nolstretch
restore

Number of strata   =         1                  Number of obs      =      2951
Number of PSUs     =      2951                  Population size    = 68599.605
                                                 Design df          =      2950
                                                 F(  10,   2941)    =     13.95
                                                 Prob > F           =    0.0000
                                                 R-squared          =    0.0539
------------------------------------------------------------------------------
             |             Linearized
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      0.015      0.003    5.455   0.000        0.009       0.020
  Net_Profit |     -0.000      0.000   -5.336   0.000       -0.000      -0.000
  Home_Based |     -0.027      0.014   -1.881   0.060       -0.054       0.001
     Have_IP |     -0.026      0.021   -1.284   0.199       -0.067       0.014
OO_D_educa~r |     -0.032      0.013   -2.519   0.012       -0.058      -0.007
OO_work_ex~r |     -0.003      0.001   -4.148   0.000       -0.004      -0.001
OO_race_wh~r |     -0.019      0.018   -1.034   0.301       -0.054       0.017
   PO_gender |     -0.002      0.014   -0.129   0.897       -0.029       0.026
f13_trade_~n |      0.067      0.020    3.297   0.001        0.027       0.107
 f19_res_dev |     -0.018      0.022   -0.796   0.426       -0.061       0.026
       _cons |      0.214      0.034    6.255   0.000        0.147       0.281
------------------------------------------------------------------------------
```

```
preserve
mi unset
keep if master==0
*Declare survey design for dataset
svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
gen wgt_7_long0=wgt_7_long
xtdata mprid tdca   LnAssets Net_Profit Home_Based  Have_IP OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 f13_trade_fin f19_res_dev , fe   clear
merge    m:1    mprid             using    KFS8_L7_w1,    keepusing(wgt_7_long
sampleinfo_samplestrata_0)
keep if _merge==3
svyset  mprid  [pweight=wgt_7_long]
svy:regress tdca    LnAssets Net_Profit Home_Based  Have_IP OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 f13_trade_fin f19_res_dev  ,   cformat(%6.3f)  sformat(%6.3f)   nolstretch
restore
```

```
Number of strata   =         1                 Number of obs      =     12586
Number of PSUs     =      2951                 Population size    = 281145.73
                                                Design df          =      2950
                                                F(   8,   2943)    =         .
                                                Prob > F           =         .
                                                R-squared          =    0.0032

------------------------------------------------------------------------------
             |             Linearized
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |       0.005      0.002     2.879   0.004        0.001       0.008
  Net_Profit |      -0.000      0.000    -1.398   0.162       -0.000       0.000
  Home_Based |      -0.014      0.023    -0.606   0.544       -0.060       0.032
     Have_IP |       0.011      0.013     0.863   0.388       -0.014       0.036
OO_D_educa~r |      -0.008      0.027    -0.311   0.756       -0.060       0.044
OO_work_ex~r |      -0.001      0.002    -0.433   0.665       -0.005       0.003
OO_race_wh~r |       0.069      0.079     0.879   0.380       -0.085       0.223
   PO_gender |       0.000  (omitted)
f13_trade_~n |      -0.001      0.011    -0.065   0.948       -0.023       0.021
 f19_res_dev |       0.034      0.011     2.992   0.003        0.012       0.056
       _cons |       0.166      0.075     2.205   0.028        0.018       0.313
------------------------------------------------------------------------------
```

6.4.4 Random Effects (Random-Intercept) Models

In random effects models, the model to be fit is

$$y_{it} = \alpha + \beta x_{it} + \epsilon_{it} \quad \text{, for } i = 1,2,..n \text{ and } t = 1,2,..T.$$

$$\epsilon_{it} = \mu_i + \nu_{it}$$

$$y_{it} = \alpha + \mu_i + \beta x_{it} + \nu_{it}$$

Here μ_i is a random variable representing a firm-specific effect. We assume $E(\mu_i) = 0$, $Var(\mu_i) = \sigma_\mu^2$, $Cov(\mu_i, \nu_{it}) = 0$ and $Cov(\mu_i, x_{it}) = 0$. The parameter σ_μ^2 summarizes the heterogeneity among firms. The idiosyncratic error (ν_{it}) varies over firms and time, where $E(\nu_{it}) = 0$, $Var(\nu_{it}) = \sigma_\nu^2$, $(\nu_{it}, \nu_{jt}) = 0 \; for \; i \neq j$, $Cov(\mu_i, x_{it}) = 0$ and to identify the model parameters, we assume that the two terms are independent, $Cov(\mu_i, \nu_{it}) = 0$.

Examples 6.6 Random Effects (Random-Intercept)

```
*Random Effects (Mixed) Models

*Gllamm
* Level one  t time points (or waves)
* Level two  i firms
gen  weight = wgt_7_long // wij
gen Baseweight= wgt_ini_0 // wi
gen Level1_w = weight / Baseweight // wi|j=wij/wi
*Rescale the level 1 weights, using Rabe-Hesketh and Skrondal (2006):
gen sqw = Level1_w^2
egen sumsqw = sum(sqw), by(master mprid)
egen sumw = sum(Level1_w), by(master mprid)
gen Level1_w_s = Level1_w * sumw/sumsqw
gen pwt2 = Baseweight
gen pwt1 = Level1_w_s
*Stata
* Level one  i firms
* Level two  t time points (or waves)

mi  xeq  0:xtmixed  tdca       LnAssets  Net_Profit  i.Home_Based     i.Have_IP
OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 i.f13_trade_fin i.f19_res_dev  [pweight=Level1_w] || mprid:, pweight(Baseweight)
pwscale(effective)   cformat(%6.3f)  sformat(%6.3f)   nolstretch
```

```
Mixed-effects regression                        Number of obs      =     12586
Group variable: mprid                           Number of groups   =      2951

                                                Obs per group: min =         1
                                                               avg =       4.3
                                                               max =         8


                                                Wald chi2(9)       =         .
Log pseudolikelihood = -22141.216               Prob > chi2        =         .

                            (Std. Err. adjusted for 2951 clusters in mprid)
------------------------------------------------------------------------------
             |               Robust
        tdca |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |       0.009      0.001     6.762   0.000        0.006       0.011
  Net_Profit |      -0.000      0.000    -3.157   0.002       -0.000      -0.000
1.Home_Based |      -0.038      0.011    -3.600   0.000       -0.059      -0.017
   1.Have_IP |      -0.004      0.010    -0.400   0.689       -0.023       0.015
OO_D_educa~r |      -0.035      0.010    -3.301   0.001       -0.055      -0.014
OO_work_ex~r |      -0.002      0.001    -4.099   0.000       -0.003      -0.001
OO_race_wh~r |      -0.000      0.016    -0.016   0.987       -0.031       0.030
   PO_gender |       0.003      0.013     0.232   0.816       -0.022       0.028
1.f13_trad~n |       0.024      0.009     2.528   0.011        0.005       0.043
1.f19_res_~v |       0.025      0.010     2.549   0.011        0.006       0.044
       _cons |       0.242      0.023    10.507   0.000        0.197       0.287
------------------------------------------------------------------------------

------------------------------------------------------------------------------
                             |               Robust
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
mprid: Identity              |
                   sd(_cons) |      0.207      0.005        0.197       0.217
-----------------------------+------------------------------------------------
                sd(Residual) |      0.268      0.003        0.261       0.275
------------------------------------------------------------------------------
```

```
mi  estimate:xtmixed  tdca      LnAssets  Net_Profit  i.Home_Based    i.Have_IP
OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 i.f13_trade_fin i.f19_res_dev  [pweight=Level1_w] || mprid:, pweight(Baseweight)
pwscale(effective)
mi estimate,    cformat(%6.3f)  sformat(%6.3f)   nolstretch



Multiple-imputation estimates                     Imputations        =         5
Mixed-effects regression                          Number of obs      =     18286

Group variable: mprid                             Number of groups   =      3140
                                                  Obs per group: min =         1
                                                                 avg =       5.8
                                                                 max =         8

                                                  Average RVI        =    0.0259
                                                  Largest FMI        =    0.0678
DF adjustment:   Large sample                     DF:     min        =    923.42
                                                          avg        =  17368.26
                                                          max        =  72170.21
Model F test:       Equal FMI                     F(   9,46164.5)    =     20.66
Within VCE type:       Robust                     Prob > F           =    0.0000

                              (Within VCE adjusted for 3140 clusters in mprid)
------------------------------------------------------------------------------
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      0.009      0.001    8.104   0.000        0.007       0.011
  Net_Profit |     -0.000      0.000   -2.414   0.016       -0.000      -0.000
1.Home_Based |     -0.041      0.009   -4.383   0.000       -0.060      -0.023
   1.Have_IP |     -0.001      0.009   -0.130   0.897       -0.018       0.016
OO_D_educa~r |     -0.029      0.009   -3.140   0.002       -0.048      -0.011
OO_work_ex~r |     -0.002      0.000   -4.920   0.000       -0.003      -0.001
OO_race_wh~r |     -0.002      0.014   -0.113   0.910       -0.029       0.026
   PO_gender |      0.003      0.011    0.240   0.810       -0.020       0.025
1.f13_trad~n |      0.028      0.008    3.353   0.001        0.012       0.045
1.f19_res_~v |      0.028      0.009    3.098   0.002        0.010       0.045
       _cons |      0.220      0.020   11.043   0.000        0.181       0.259
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
mprid: Identity              |
                   sd(_cons) |      0.197      0.005        0.189       0.206
-----------------------------+------------------------------------------------
                sd(Residual) |      0.267      0.003        0.261       0.273
------------------------------------------------------------------------------
```

```
*Gllamm
mi  xeq  0:xi:  gllamm  tdca      LnAssets  Net_Profit  i.Home_Based    i.Have_IP
OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 i.f13_trade_fin i.f19_res_dev  ,i(mprid) pweight(pwt)adapt

Running adaptive quadrature
Iteration 0:    log likelihood = -33011.766
Iteration 1:    log likelihood = -32136.769
Iteration 2:    log likelihood = -31909.609
Iteration 3:    log likelihood = -31703.037
Iteration 4:    log likelihood = -31331.721
Iteration 5:    log likelihood = -31134.392
Iteration 6:    log likelihood = -30727.581
Iteration 7:    log likelihood = -30538.708
Iteration 8:    log likelihood = -29910.051
Iteration 9:    log likelihood = -29475.752
Iteration 10:    log likelihood = -29341.462
Iteration 11:    log likelihood = -24637.071
Iteration 12:    log likelihood = -22582.412
Iteration 13:    log likelihood = -22143.381
Iteration 14:    log likelihood = -22141.216
Iteration 15:    log likelihood = -22141.216

Adaptive quadrature has converged, running Newton-Raphson
Iteration 0:   log likelihood = -22141.216
Iteration 1:   log likelihood = -22141.216  (backed up)
Iteration 2:   log likelihood = -22141.216

number of level 1 units = 12586
number of level 2 units = 2951
Condition Number = 42365651
gllamm model
log likelihood = -22141.216

Robust standard errors
                                    Std.                      [95%
 tdca                       Coef.   Err.      z       P>z     Conf.     Interval]

 LnAssets                    0.009  0.001    6.690  0.000      0.006      0.011
 Net_Profit                  0.000  0.000   -3.130  0.002      0.000      0.000
 _IHome_Base_1              -0.038  0.011   -3.600  0.000     -0.059     -0.017
 _IHave_IP_1                -0.004  0.010   -0.400  0.690     -0.023      0.015
 OO_D_education_owner       -0.035  0.011   -3.290  0.001     -0.055     -0.014
 OO_work_exp_owner          -0.002  0.001   -4.100  0.000     -0.003     -0.001
 OO_race_white_owner         0.000  0.016   -0.020  0.987     -0.031      0.030
 PO_gender                   0.003  0.013    0.230  0.816     -0.022      0.028
 _If13_trade_1               0.024  0.010    2.520  0.012      0.005      0.043
 _If19_res_d_1               0.025  0.010    2.550  0.011      0.006      0.044
 _cons                       0.242  0.023   10.440  0.000      0.197      0.287

Variance at level 1
------------------------------------------------------------------------------

  .07163889 (.0018452)

Variances and covariances of random effects
------------------------------------------------------------------------------

 ***level 2 (mprid)

    var(1): .04289394 (.00211668)
------------------------------------------------------------------------------
```

```
mi estimate,cmdok: gllamm tdca     LnAssets  Net_Profit   Home_Based     Have_IP
OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner PO_gender   ///
  f13_trade_fin  f19_res_dev  ,i(mprid) pweight(pwt)adapt
mi estimate,    cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                    Imputations     =          5
gllamm model                                     Number of obs   =      18286
                                                 Average RVI     =     0.0261
                                                 Largest FMI     =     0.0678
DF adjustment:   Large sample                    DF:     min     =     923.48
                                                         avg     =   17490.67
Within VCE type:          OIM                            max     =   72283.61

------------------------------------------------------------------------------
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
tdca         |
    LnAssets |       0.009      0.001     8.058   0.000        0.007       0.011
  Net_Profit |      -0.000      0.000    -2.403   0.016       -0.000      -0.000
  Home_Based |      -0.041      0.009    -4.383   0.000       -0.060      -0.023
     Have_IP |      -0.001      0.009    -0.130   0.897       -0.018       0.016
OO_D_educa~r |      -0.029      0.009    -3.140   0.002       -0.048      -0.011
OO_work_ex~r |      -0.002      0.000    -4.922   0.000       -0.003      -0.001
OO_race_wh~r |      -0.002      0.014    -0.113   0.910       -0.029       0.026
   PO_gender |       0.003      0.011     0.240   0.810       -0.020       0.025
f13_trade_~n |       0.028      0.008     3.345   0.001        0.012       0.045
 f19_res_dev |       0.028      0.009     3.097   0.002        0.010       0.045
       _cons |       0.220      0.020    10.951   0.000        0.181       0.260
-------------+----------------------------------------------------------------
lns1         |
       _cons |      -1.320      0.012  -113.693   0.000       -1.343      -1.297
-------------+----------------------------------------------------------------
mpri1        |
       _cons |       0.197      0.005    43.281   0.000        0.188       0.206
------------------------------------------------------------------------------
```

Examples 6.7 Random Effects Models as Weighted Average of the Between and Within Estimators

It can be shown that the random-effects estimator is a weighted average of the between and within estimators, with the weight being a function of the intra-class correlation.

$$y_{it} = \alpha + \mu_i + \beta x_{it} + \nu_{it}$$

$$y_{it} - \theta\bar{y}_i = \alpha(1-\theta) + \beta(x_{it} - \theta\bar{x}_i) + \mu_i(1-\theta) + (\nu_{it} - \theta\bar{\nu}_i)$$

$$\text{where } \theta = 1 - \frac{\sigma_\nu}{\sqrt{\sigma_\mu^2 T + \sigma_\nu^2}}$$

The model can be estimated by OLS directly. It is also important to note the following cases: when $\theta = 1$ the random effect, (GLS) becomes fixed-effect (LSDV) and when $\theta = 0$, the random effect (GLS) becomes OLS. If $Cov(\mu_i, x_{it}) \neq 0$, the RE-estimator will be biased.

```
preserve
mi unset
keep if master==0
*Declare survey design for dataset
svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
gen wgt_7_long0=wgt_7_long
xtdata mprid tdca   LnAssets Net_Profit Home_Based  Have_IP OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 f13_trade_fin f19_res_dev , re ratio(0.77) clear
merge    m:1    mprid            using    KFS8_L7_w1,    keepusing(wgt_7_long
sampleinfo_samplestrata_0)
keep if _merge==3
svyset  mprid  [pweight=wgt_7_long]
svy:regress tdca   LnAssets Net_Profit Home_Based  Have_IP OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 f13_trade_fin f19_res_dev constant , noconstant cformat(%6.3f)  sformat(%6.3f)
nolstretch
restore
```

```
Number of strata   =         1                  Number of obs      =     12586
Number of PSUs     =      2951                  Population size    = 281145.73
                                                 Design df          =      2950
                                                 F(  10,   2941)    =         .
                                                 Prob > F           =         .
                                                 R-squared          =    0.2363

------------------------------------------------------------------------------
             |             Linearized
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |       0.009      0.001     7.110   0.000        0.007       0.012
  Net_Profit |      -0.000      0.000    -3.430   0.001       -0.000      -0.000
  Home_Based |      -0.031      0.011    -2.828   0.005       -0.053      -0.010
     Have_IP |      -0.006      0.010    -0.592   0.554       -0.026       0.014
OO_D_educa~r |      -0.036      0.011    -3.269   0.001       -0.057      -0.014
OO_work_ex~r |      -0.002      0.001    -3.618   0.000       -0.003      -0.001
OO_race_wh~r |      -0.004      0.016    -0.233   0.815       -0.035       0.028
   PO_gender |       0.005      0.013     0.379   0.705       -0.021       0.031
f13_trade_~n |       0.025      0.010     2.558   0.011        0.006       0.043
 f19_res_dev |       0.020      0.010     1.979   0.048        0.000       0.040
    constant |       0.237      0.023    10.265   0.000        0.192       0.282
------------------------------------------------------------------------------
```

6.4.5 Random-Coefficient Models

In a random-coefficient model, in addition to the random variable representing a firm-specific effect (Random-Intercept), we allow the slope to vary from one firm to the next, randomly:

$$\epsilon_{it} = \mu_i + \zeta_{it} x_{it} + \nu_{it}$$

$$y_{it} = \mu_i + \zeta_{it} x_{it} + \beta x_{it} + \nu_{it}$$

where ζ_{it} is a random variable. Suppose we think that the effect of assets varies from one firm to the next.

Examples 6.8 Random-Coefficient Models

```
mi  xeq  0:xtmixed  tdca      LnAssets  Net_Profit  i.Home_Based    i.Have_IP
OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner PO_gender  ///
 i.f13_trade_fin   i.f19_res_dev      [pweight=Level1_w]   ||   mprid:LnAssets,
pweight(Baseweight)  pwscale(effective)        cformat(%6.3f)    sformat(%6.3f)
nolstretch
```

```
Mixed-effects regression                        Number of obs      =     12586
Group variable: mprid                           Number of groups   =      2951

                                                Obs per group: min =         1
                                                               avg =       4.3
                                                               max =         8


                                                Wald chi2(9)       =         .
Log pseudolikelihood = -22067.817               Prob > chi2        =         .

                            (Std. Err. adjusted for 2951 clusters in mprid)
------------------------------------------------------------------------------
             |               Robust
        tdca |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |       0.009      0.001     7.308   0.000        0.007       0.012
  Net_Profit |      -0.000      0.000    -3.348   0.001       -0.000      -0.000
1.Home_Based |      -0.039      0.011    -3.638   0.000       -0.059      -0.018
   1.Have_IP |      -0.003      0.010    -0.345   0.730       -0.023       0.016
OO_D_educa~r |      -0.034      0.010    -3.297   0.001       -0.055      -0.014
OO_work_ex~r |      -0.002      0.001    -4.011   0.000       -0.003      -0.001
OO_race_wh~r |       0.001      0.016     0.034   0.973       -0.030       0.031
   PO_gender |       0.000      0.013     0.011   0.991       -0.024       0.025
1.f13_trad~n |       0.024      0.009     2.498   0.012        0.005       0.042
1.f19_res_~v |       0.026      0.010     2.606   0.009        0.006       0.045
       _cons |       0.235      0.023    10.170   0.000        0.189       0.280
------------------------------------------------------------------------------

------------------------------------------------------------------------------
                             |               Robust
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
mprid: Independent           |
                sd(LnAssets) |      0.010      0.001         0.008       0.013
                   sd(_cons) |      0.177      0.010         0.159       0.197
-----------------------------+------------------------------------------------
                sd(Residual) |      0.267      0.003         0.260       0.274
------------------------------------------------------------------------------
```

```
mi  estimate:xtmixed  tdca      LnAssets  Net_Profit  i.Home_Based    i.Have_IP
OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 i.f13_trade_fin   i.f19_res_dev       [pweight=Level1_w]   ||   mprid:LnAssets,
pweight(Baseweight)  pwscale(effective)         cformat(%6.3f)    sformat(%6.3f)
nolstretch
mi estimate,    cformat(%6.3f)  sformat(%6.3f)   nolstretch
```

```
Multiple-imputation estimates                   Imputations        =          5
Mixed-effects regression                        Number of obs      =      18286

Group variable: mprid                           Number of groups   =       3140
                                                Obs per group: min =          1
                                                               avg =        5.8
                                                               max =          8

                                                Average RVI        =     0.0236
                                                Largest FMI        =     0.0665
DF adjustment:   Large sample                   DF:     min        =     961.13
                                                        avg        =   18169.32
                                                        max        =   68494.21
Model F test:       Equal FMI                   F(   9,50987.0)    =      24.40
Within VCE type:       Robust                   Prob > F           =     0.0000

                            (Within VCE adjusted for 3140 clusters in mprid)
------------------------------------------------------------------------------
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |       0.010      0.001    9.351   0.000        0.008       0.012
  Net_Profit |      -0.000      0.000   -2.369   0.018       -0.000      -0.000
1.Home_Based |      -0.042      0.009   -4.511   0.000       -0.061      -0.024
   1.Have_IP |      -0.001      0.009   -0.080   0.937       -0.018       0.016
OO_D_educa~r |      -0.029      0.009   -3.135   0.002       -0.047      -0.011
OO_work_ex~r |      -0.002      0.000   -4.764   0.000       -0.003      -0.001
OO_race_wh~r |      -0.002      0.014   -0.147   0.884       -0.029       0.025
   PO_gender |      -0.001      0.011   -0.066   0.948       -0.023       0.021
1.f13_trad~n |       0.027      0.008    3.223   0.001        0.011       0.044
1.f19_res_~v |       0.028      0.009    3.140   0.002        0.011       0.046
       _cons |       0.211      0.020   10.692   0.000        0.172       0.249
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
mprid: Independent           |
                sd(LnAssets) |      0.012      0.001        0.010       0.014
                   sd(_cons) |      0.156      0.009        0.139       0.174
-----------------------------+------------------------------------------------
                sd(Residual) |      0.266      0.003        0.260       0.272
------------------------------------------------------------------------------
```

6.4.6 Hybrid Model

Because fixed effects model methods control for unmeasured characteristics of firms, the estimate of the fixed effects are different from estimates produce by random effects method. Yet, it is possible to decompose each time-varying predictor into two parts: within firm component and between firm component.

Consider the between estimator's model:

$$\bar{y}_i = \mu_i + \beta_1 \bar{x}_i + \bar{v}_i$$

and the within estimators

$$y_{it} - \bar{y}_i = \beta_2 (x_{it} - \bar{x}_i) + (v_{it} - \bar{v}_i)$$

From the above two models we can decompose the time-varying predictor x_{it} into

$$y_{it} = \mu_i + \beta_1 \bar{x}_i + \beta_2 (x_{it} - \bar{x}_i) + v_{it}$$

In this model, the changes in the average value of x_{it} for a firm have a different effect from temporary departures from the average.

The model will produce an identical estimate for the fixed effects model(β_2), allow for random effects estimates for the time-invariant predictors, allow for random slopes for the time-varying predictors, provide a test of fixed effects vs. random effects models, and more flexibility in modeling correlation structure among the errors term (v_{it}).

Examples 6.9 Hybrid Model

```
global vars  "  LnAssets Net_Profit  "
 foreach var in $vars{

egen m_`var'=mean(`var'), by(_mi_m mprid)
gen d_`var'=`var'-m_`var'
}

mi xeq 0:xtmixed tdca    m_LnAssets     m_Net_Profit    PO_gender ///
                         d_LnAssets     d_Net_Profit    ///
                         [pweight=Level1_w]    ||    mprid:,    pweight(Baseweight)
pwscale(effective)


Performing gradient-based optimization:

Iteration 0:   log pseudolikelihood = -22495.567
Iteration 1:   log pseudolikelihood = -22492.233
Iteration 2:   log pseudolikelihood = -22492.231
```

```
Computing standard errors:
Mixed-effects regression                        Number of obs      =     12670
Group variable: mprid                           Number of groups   =      2959

                                                Obs per group: min =         1
                                                               avg =       4.3
                                                               max =         8
                                                Wald chi2(4)       =         .
Log pseudolikelihood = -22492.231               Prob > chi2        =         .
                                (Std. Err. adjusted for 2959 clusters in mprid)
------------------------------------------------------------------------------
             |               Robust
        tdca |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  m_LnAssets |   .0195254   .0021063     9.27   0.000     .0153972    .0236536
m_Net_Profit |  -6.78e-09   1.38e-09    -4.91   0.000    -9.49e-09   -4.07e-09
   PO_gender |  -.0103867   .0127296    -0.82   0.415    -.0353364    .0145629
  d_LnAssets |   .0048035   .0014833     3.24   0.001     .0018963    .0077108
d_Net_Profit |  -2.73e-11   2.63e-11    -1.04   0.300    -7.89e-11    2.43e-11
       _cons |   .0944752   .0217125     4.35   0.000     .0519195    .1370309
------------------------------------------------------------------------------

------------------------------------------------------------------------------
                             |              Robust
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
mprid: Identity              |
                   sd(_cons) |   .2086916   .0049087      .1992891    .2185378
-----------------------------+------------------------------------------------
                sd(Residual) |   .2680359   .0034547      .2613497    .2748933
------------------------------------------------------------------------------
*The model will produce indetical estimate for the fixed effects model

mi xeq 0:xtreg tdca    LnAssets Net_Profit PO_gender [pweight=wgt_7_long], fe
i(mprid) vce(robust)

note: PO_gender omitted because of collinearity

Fixed-effects (within) regression               Number of obs      =     12670
Group variable: mprid                           Number of groups   =      2959

R-sq:  within  = 0.0016                         Obs per group: min =         1
       between = 0.0269                                        avg =       4.3
       overall = 0.0210                                        max =         8

                                                F(1,2958)          =         .
corr(u_i, Xb)  = 0.1383                         Prob > F           =         .

                                (Std. Err. adjusted for 2959 clusters in mprid)
------------------------------------------------------------------------------
             |               Robust
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |   .0051663   .0015771     3.28   0.001      .002074    .0082586
  Net_Profit |  -2.43e-11   2.27e-11    -1.07   0.283    -6.88e-11    2.01e-11
   PO_gender |          0  (omitted)
       _cons |    .221726   .0156232    14.19   0.000     .1910927    .2523594
-------------+----------------------------------------------------------------
     sigma_u |  .27352343
     sigma_e |  .26552843
         rho |  .51482834   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

```
mi estimate (diff1: _b[m_LnAssets]-_b[d_LnAssets]) ///
(diff2:   _b[m_Net_Profit]-_b[d_Net_Profit]   )   ,   post         saving(miest,
replace):xtmixed tdca    m_LnAssets     m_Net_Profit   PO_gender ///
                         d_LnAssets     d_Net_Profit   ///
                         [pweight=Level1_w]   ||   mprid:,   pweight(Baseweight)
pwscale(effective)

Multiple-imputation estimates                     Imputations        =         5
Mixed-effects regression                          Number of obs      =     18286

Group variable: mprid                             Number of groups   =      3140
                                                  Obs per group: min =         1
                                                                 avg =       5.8
                                                                 max =         8

                                                  Average RVI        =    0.0226
                                                  Largest FMI        =    0.0615
DF adjustment:   Large sample                     DF:     min        =   1116.20
                                                          avg        = 200890.18
                                                          max        = 910719.84
Model F test:       Equal FMI                     F(   4,25116.8)    =     45.61
Within VCE type:       Robust                     Prob > F           =    0.0000

                              (Within VCE adjusted for 3140 clusters in mprid)
------------------------------------------------------------------------------
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  m_LnAssets |    .021056   .0017055    12.35   0.000     .0177133    .0243987
m_Net_Profit |  -6.50e-09   1.36e-09    -4.77   0.000    -9.17e-09   -3.83e-09
   PO_gender |  -.0126376   .0114551    -1.10   0.270    -.0350895    .0098144
  d_LnAssets |   .0054178   .0013155     4.12   0.000     .0028367     .007999
d_Net_Profit |  -3.52e-11   4.52e-11    -0.78   0.436    -1.24e-10    5.35e-11
       _cons |   .0591617   .0173443     3.41   0.001     .0251675    .0931558
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
mprid: Identity              |
                  sd(_cons) |   .1989873   .0044244      .1905001    .2078528
-----------------------------+------------------------------------------------
               sd(Residual) |   .2671837   .0031172      .2611378    .2733695
------------------------------------------------------------------------------

Transformations                                   Average RVI        =    0.0221
                                                  Largest FMI        =    0.0400
DF adjustment:   Large sample                     DF:     min        =   2599.89
                                                          avg        = 143205.60
Within VCE type:       Robust                             max        = 283811.32

        diff1: _b[m_LnAssets]-_b[d_LnAssets]
        diff2: _b[m_Net_Profit]-_b[d_Net_Profit]

                              (Within VCE adjusted for 3140 clusters in mprid)
------------------------------------------------------------------------------
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       diff1 |   .0156382   .0021671     7.22   0.000     .0113888    .0198875
       diff2 |  -6.46e-09   1.33e-09    -4.84   0.000    -9.08e-09   -3.85e-09
------------------------------------------------------------------------------
```

```
mi testtransform diff1 diff2


        diff1: _b[m_LnAssets]-_b[d_LnAssets]
        diff2: _b[m_Net_Profit]-_b[d_Net_Profit]

 ( 1)  diff1 = 0
 ( 2)  diff2 = 0

       F(  2,4888.6) =   35.54
            Prob > F =    0.0000

*The model will produce indetical estimate for the fixed effects model

mi estimate:xtreg tdca   LnAssets Net_Profit PO_gender [pweight=wgt_7_long], fe
i(mprid) vce(robust)

Multiple-imputation estimates                   Imputations        =         5
Fixed-effects (within) regression               Number of obs      =     18286

Group variable: mprid                           Number of groups   =      3140
                                                Obs per group: min =         1
                                                               avg =       5.8
                                                               max =         8

                                                Average RVI        =  1.02e+04
                                                Largest FMI        =    0.0839
                                                Complete DF        =      3139
DF adjustment:   Small sample                   DF:     min        =    504.67
                                                        avg        =    833.29
                                                        max        =   1334.32
Model F test:       Equal FMI                   F(   1,  660.9)    =     15.30
Within VCE type:       Robust                   Prob > F           =    0.0001

                              (Within VCE adjusted for 3140 clusters in mprid)
------------------------------------------------------------------------------
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |   .0054175   .0013849     3.91   0.000     .0026981    .0081369
  Net_Profit |  -3.61e-11   4.17e-11    -0.87   0.386    -1.18e-10    4.57e-11
   PO_gender |           0  (omitted)
       _cons |   .1958908   .0137121    14.29   0.000      .168951    .2228306
-------------+----------------------------------------------------------------
     sigma_u |  .25231258
     sigma_e |  .26595924
         rho |  .47368715   (fraction of variance due to u_i)
------------------------------------------------------------------------------
Note: sigma_u and sigma_e are combined in the original metric.
```

The random effects model is a special case of the fixed effects model. If the random effects model is correct, then the coefficient for d_var should be the same as the one for m_var. The above results clearly indicate that the random effects model should be rejected in favor of the fixed effects model.

6.5 Nonlinear Panel Models

As in the case of linear panel models, the assumption that observations are independent is not appropriate when dealing with longitudinal data, clustered data, and multilevel data. As a result of dependence among observations within and across groups, special methods of estimations are needed. Those methods include one or more of the following approaches:

1. Robust Standard Errors
2. Generalized Estimating Equations (FGLS) [population averaged models]
3. Fixed Effects Model
4. Random Effects Models

6.5.1 Logit Models for Binary Response Variables

Logistic regression or logit regression is a type of probabilistic model used for predicting the outcome of a categorical dependent variable (or binary response) based on one or more predictor variables.

Let us consider the case where the response y_i is binary. We view y_i as a realization of a random variable Y_i that can take the values one and zero with probabilities p_i and 1-p_i, respectively. Assuming that Y_i has Bernoulli distribution with parameter p_i then

$$\Pr(Y_i = y_i) = p_i^{y_i}\,(1-p_i)^{1-y_i}$$

where $\mathrm{E}(Y_i) = p_i$ and $Var(Y_i) = p_i(1-p_i)$

Assume that the probabilities p_i depend linearly on a vector of observed covariates x_i.

$$p_i = \beta x_i$$

Because $0 \le p_i \le 1$ and βx_i could take any value, we need to remove the range restrictions. By transforming the probability to the odds ratio (ratio of success to failure) and taking the logarithms of the odds ratio

$$Odds = \frac{p_i}{1-p_i}$$

$$Logit(p_i) = Log\left[\frac{p_i}{1-p_i}\right]$$

Now, we have $-\infty \le Logit(p_i) \le +\infty$. Assuming that the log odds (logit) is linearly related to x_i

$$Logit(p_i) = Log\left[\frac{p_i}{1-p_i}\right] = \beta x_i$$

and

$$\frac{p_i}{1-p_i} = e^{\beta x_i} \Longrightarrow p_i = \frac{e^{\beta x_i}}{1+e^{\beta x_i}}$$

Examples 6.10 Robust Standard Errors

```
mi xeq 0:svy: logit Have_IP LnAssets     i.Home_Based   i.OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner PO_gender  i.Comp_advantage ///
i.hightech i.d1a_provide_service i.d1b_provide_product   ,    or   cformat(%6.3f)
sformat(%6.3f)   nolstretch

Logistic regression                                     Number of obs   =      16544
                                                        Wald chi2(10)   =     350.35
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -164703.74                       Pseudo R2       =     0.0827

                              (Std. Err. adjusted for 3092 clusters in mprid)
------------------------------------------------------------------------------
             |               Robust
     Have_IP | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      1.023      0.013    1.789   0.074        0.998       1.048
1.Home_Based |      1.042      0.097    0.436   0.663        0.867       1.251
1.OO_D_edu~r |      1.641      0.152    5.361   0.000        1.369       1.967
OO_work_ex~r |      1.006      0.005    1.213   0.225        0.997       1.015
OO_race_wh~r |      1.023      0.134    0.172   0.863        0.791       1.322
   PO_gender |      1.192      0.133    1.575   0.115        0.958       1.482
1.Comp_adv~e |      2.550      0.192   12.408   0.000        2.200       2.957
  1.hightech |      1.905      0.235    5.228   0.000        1.496       2.426
1.d1a_prov~e |      0.624      0.072   -4.077   0.000        0.497       0.783
1.d1b_prov~t |      2.076      0.189    8.039   0.000        1.738       2.481
       _cons |      0.060      0.014  -11.844   0.000        0.038       0.095
------------------------------------------------------------------------------

mi  xeq  0:  logit  Have_IP  LnAssets      i.Home_Based    i.OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner PO_gender  i.Comp_advantage ///
i.hightech  i.d1a_provide_service  i.d1b_provide_product  [pweight=wgt_7_long],
vce(cluster mprid) or  cformat(%6.3f)  sformat(%6.3f)   nolstretch

Logistic regression                                     Number of obs   =      16544
                                                        Wald chi2(10)   =     350.35
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -164703.74                       Pseudo R2       =     0.0827
                              (Std. Err. adjusted for 3092 clusters in mprid)
------------------------------------------------------------------------------
             |               Robust
     Have_IP | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      1.023      0.013    1.789   0.074        0.998       1.048
1.Home_Based |      1.042      0.097    0.436   0.663        0.867       1.251
1.OO_D_edu~r |      1.641      0.152    5.361   0.000        1.369       1.967
OO_work_ex~r |      1.006      0.005    1.213   0.225        0.997       1.015
OO_race_wh~r |      1.023      0.134    0.172   0.863        0.791       1.322
   PO_gender |      1.192      0.133    1.575   0.115        0.958       1.482
1.Comp_adv~e |      2.550      0.192   12.408   0.000        2.200       2.957
  1.hightech |      1.905      0.235    5.228   0.000        1.496       2.426
1.d1a_prov~e |      0.624      0.072   -4.077   0.000        0.497       0.783
1.d1b_prov~t |      2.076      0.189    8.039   0.000        1.738       2.481
       _cons |      0.060      0.014  -11.844   0.000        0.038       0.095
------------------------------------------------------------------------------

mi estimate:svy: logit Have_IP LnAssets   i.Home_Based  i.OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner PO_gender  i.Comp_advantage ///
i.hightech i.d1a_provide_service i.d1b_provide_product  ,   or  cformat(%6.3f)
sformat(%6.3f)   nolstretch
mi estimate,   or  cformat(%6.3f)  sformat(%6.3f)   nolstretch
```

```
Multiple-imputation estimates                   Imputations       =          5
Survey: Logistic regression                     Number of obs     =      18286

Number of strata  =         6                   Population size   = 408495.43
Number of PSUs    =      3140
                                                Average RVI       =     0.0038
                                                Largest FMI       =     0.0228
                                                Complete DF       =       3134
DF adjustment:   Small sample                   DF:     min       =    2205.77
                                                        avg       =    3027.79
                                                        max       =    3129.76
Model F test:       Equal FMI                   F(  10, 3127.3)   =      38.00
Within VCE type:   Linearized                   Prob > F          =     0.0000
------------------------------------------------------------------------------
     Have_IP | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      1.026      0.013    2.049   0.041        1.001       1.052
1.Home_Based |      1.043      0.094    0.469   0.639        0.874       1.246
1.OO_D_edu~r |      1.692      0.151    5.888   0.000        1.420       2.015
OO_work_ex~r |      1.005      0.005    1.111   0.267        0.996       1.014
OO_race_wh~r |      0.989      0.126   -0.089   0.929        0.769       1.270
   PO_gender |      1.176      0.128    1.486   0.137        0.949       1.457
1.Comp_adv~e |      2.550      0.186   12.803   0.000        2.209       2.943
  1.hightech |      1.965      0.235    5.654   0.000        1.555       2.484
1.d1a_prov~e |      0.625      0.070   -4.178   0.000        0.501       0.779
1.d1b_prov~t |      2.080      0.184    8.266   0.000        1.748       2.475
       _cons |      0.059      0.014  -12.234   0.000        0.038       0.093
------------------------------------------------------------------------------

mi estimate: logit Have_IP LnAssets   i.Home_Based  i.OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner PO_gender  i.Comp_advantage ///
i.hightech i.d1a_provide_service i.d1b_provide_product [pweight=wgt_7_long],
vce(cluster mprid) or  cformat(%6.3f)  sformat(%6.3f)   nolstretch
mi estimate,   or  cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                     Imputations     =          5
Logistic regression                               Number of obs   =      18286
                                                  Average RVI     =     0.0038
                                                  Largest FMI     =     0.0228
DF adjustment:   Large sample                     DF:     min     =    7880.45
                                                          avg     = 3322686.99
                                                          max     =   1.53e+07
Model F test:       Equal FMI                     F(  10, 2.1e+06)=      37.96
Within VCE type:       Robust                     Prob > F        =     0.0000
                              (Within VCE adjusted for 3140 clusters in mprid)
------------------------------------------------------------------------------
     Have_IP | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      1.026      0.013    2.050   0.040        1.001       1.052
1.Home_Based |      1.043      0.095    0.469   0.639        0.874       1.246
1.OO_D_edu~r |      1.692      0.151    5.889   0.000        1.420       2.015
OO_work_ex~r |      1.005      0.005    1.111   0.267        0.996       1.014
OO_race_wh~r |      0.989      0.126   -0.089   0.929        0.770       1.270
   PO_gender |      1.176      0.128    1.485   0.137        0.950       1.456
1.Comp_adv~e |      2.550      0.186   12.806   0.000        2.210       2.943
  1.hightech |      1.965      0.235    5.656   0.000        1.555       2.483
1.d1a_prov~e |      0.625      0.070   -4.176   0.000        0.501       0.779
1.d1b_prov~t |      2.080      0.184    8.268   0.000        1.749       2.475
       _cons |      0.059      0.014  -12.222   0.000        0.038       0.093
------------------------------------------------------------------------------
```

```
mi estimate:svy: glm Have_IP LnAssets   i.Home_Based  i.OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner PO_gender  i.Comp_advantage ///
i.hightech i.d1a_provide_service i.d1b_provide_product   , family(binomial)
mi estimate,   eform  cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                   Imputations       =         5
Survey: Generalized linear models               Number of obs     =     18286

Number of strata  =         6                   Population size   = 408495.43
Number of PSUs    =      3140
                                                Average RVI       =    0.0038
                                                Largest FMI       =    0.0228
                                                Complete DF       =      3134
DF adjustment:   Small sample                   DF:     min       =   2205.77
                                                        avg       =   3027.79
Within VCE type:   Linearized                           max       =   3129.76
------------------------------------------------------------------------------
     Have_IP |     exp(b)   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      1.026      0.013    2.049   0.041        1.001       1.052
1.Home_Based |      1.043      0.094    0.469   0.639        0.874       1.246
1.OO_D_edu~r |      1.692      0.151    5.888   0.000        1.420       2.015
OO_work_ex~r |      1.005      0.005    1.111   0.267        0.996       1.014
OO_race_wh~r |      0.989      0.126   -0.089   0.929        0.769       1.270
   PO_gender |      1.176      0.128    1.486   0.137        0.949       1.457
1.Comp_adv~e |      2.550      0.186   12.803   0.000        2.209       2.943
  1.hightech |      1.965      0.235    5.654   0.000        1.555       2.484
1.d1a_prov~e |      0.625      0.070   -4.178   0.000        0.501       0.779
1.d1b_prov~t |      2.080      0.184    8.266   0.000        1.748       2.475
       _cons |      0.059      0.014  -12.234   0.000        0.038       0.093
------------------------------------------------------------------------------

mi estimate, cmdok: gllamm Have_IP LnAssets   Home_Based  OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner PO_gender  Comp_advantage ///
hightech d1a_provide_service d1b_provide_product  ,i(mprid) pweight(pwt) nip(30)
adapt   link(logit) fam(binom) init robust eform
mi estimate,   eform        cformat(%6.3f)  sformat(%6.3f)   nolstretch

Robust standard errors
Multiple-imputation estimates                     Imputations     =         5
fixed effects model                               Number of obs   =     18286
                                                  Average RVI     =    0.0088
                                                  Largest FMI     =    0.0401
DF adjustment:   Large sample                     DF:     min     =   2586.61
                                                          avg     =  272356.71
Within VCE type:          OIM                             max     =  720650.81
------------------------------------------------------------------------------
     Have_IP |     exp(b)   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      1.026      0.008    3.109   0.002        1.009       1.042
  Home_Based |      1.065      0.051    1.318   0.187        0.970       1.169
OO_D_educa~r |      1.668      0.078   10.932   0.000        1.522       1.828
OO_work_ex~r |      1.006      0.002    2.545   0.011        1.001       1.010
OO_race_wh~r |      0.949      0.063   -0.783   0.434        0.834       1.081
   PO_gender |      1.200      0.065    3.373   0.001        1.079       1.334
Comp_advan~e |      2.504      0.127   18.068   0.000        2.267       2.766
    hightech |      1.842      0.112   10.069   0.000        1.635       2.074
d1a_provid~e |      0.650      0.041   -6.814   0.000        0.575       0.736
d1b_provid~t |      2.065      0.106   14.118   0.000        1.867       2.284
       _cons |      0.058      0.008  -21.482   0.000        0.045       0.075
------------------------------------------------------------------------------
```

Examples 6.11 Population-Averaged Model

```
mi  xeq  0:xtgee  Have_IP  LnAssets      i.Home_Based    i.OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner PO_gender  i.Comp_advantage ///
i.hightech  i.d1a_provide_service  i.d1b_provide_product  [pweight=wgt_7_long],
family(binomial)   eform   corr( exchangeable ) cformat(%6.3f)  sformat(%6.3f)
nolstretch

GEE population-averaged model                   Number of obs      =     16544
Group variable:                       mprid     Number of groups   =      3092
Link:                                 logit     Obs per group: min =         1
Family:                            binomial                    avg =       5.1
Correlation:                   exchangeable                    max =         8
                                               Wald chi2(10)      =    217.85
Scale parameter:                          1    Prob > chi2        =    0.0000
                                 (Std. Err. adjusted for clustering on mprid)
------------------------------------------------------------------------------
             |               Robust
     Have_IP | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      1.031      0.009    3.559   0.000        1.014       1.048
1.Home_Based |      1.008      0.078    0.103   0.918        0.866       1.174
1.OO_D_edu~r |      1.390      0.115    3.970   0.000        1.181       1.635
OO_work_ex~r |      1.003      0.005    0.572   0.567        0.994       1.012
OO_race_wh~r |      0.951      0.112   -0.424   0.672        0.754       1.199
   PO_gender |      1.225      0.128    1.945   0.052        0.998       1.502
1.Comp_adv~e |      1.587      0.081    9.039   0.000        1.436       1.755
  1.hightech |      1.523      0.180    3.554   0.000        1.208       1.921
1.d1a_prov~e |      0.873      0.069   -1.716   0.086        0.747       1.020
1.d1b_prov~t |      1.594      0.101    7.335   0.000        1.407       1.806
       _cons |      0.079      0.015  -13.406   0.000        0.054       0.114
------------------------------------------------------------------------------

estat   wcorr

Estimated within-mprid correlation matrix R:

      |  c1      c2       c3         c4         c5         c6         c7         c8
------+--------------------------------------------------------------------------
   r1 |       1
   r2 | .533296        1
   r3 | .533296   .533296         1
   r4 | .533296   .533296    .533296         1
   r5 | .533296   .533296    .533296   .533296      1
   r6 | .533296   .533296    .533296   .533296  .533296         1
   r7 | .533296   .533296    .533296   .533296  .533296   .533296         1
   r8 | .533296   .533296    .533296   .533296  .533296   .533296    .533296   1
```

```
mi  xeq  0:xtlogit  Have_IP  LnAssets      i.Home_Based    i.OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner PO_gender  i.Comp_advantage ///
i.hightech i.d1a_provide_service i.d1b_provide_product [pweight=wgt_7_long],  pa
eform corr( exchangeable ) cformat(%6.3f)  sformat(%6.3f)   nolstretch

GEE population-averaged model                   Number of obs      =     16544
Group variable:                      mprid      Number of groups   =      3092
Link:                                logit      Obs per group: min =         1
Family:                           binomial                     avg =       5.1
Correlation:                  exchangeable                     max =         8
                                                Wald chi2(10)      =    217.85
Scale parameter:                         1      Prob > chi2        =    0.0000
                                (Std. Err. adjusted for clustering on mprid)
------------------------------------------------------------------------------
             |               Robust
     Have_IP | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      1.031      0.009    3.559   0.000        1.014       1.048
1.Home_Based |      1.008      0.078    0.103   0.918        0.866       1.174
1.OO_D_edu~r |      1.390      0.115    3.970   0.000        1.181       1.635
OO_work_ex~r |      1.003      0.005    0.572   0.567        0.994       1.012
OO_race_wh~r |      0.951      0.112   -0.424   0.672        0.754       1.199
   PO_gender |      1.225      0.128    1.945   0.052        0.998       1.502
1.Comp_adv~e |      1.587      0.081    9.039   0.000        1.436       1.755
  1.hightech |      1.523      0.180    3.554   0.000        1.208       1.921
1.d1a_prov~e |      0.873      0.069   -1.716   0.086        0.747       1.020
1.d1b_prov~t |      1.594      0.101    7.335   0.000        1.407       1.806
       _cons |      0.079      0.015  -13.406   0.000        0.054       0.114
------------------------------------------------------------------------------

/*
Default with xtgee is to impose the exchangeable structure, which means that the
correlations
between values of the response variable at any two points is the same, it may be
better not to impose any structure
*/
mi  xeq  0:xtgee  Have_IP  LnAssets      i.Home_Based    i.OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner PO_gender  i.Comp_advantage ///
i.hightech  i.d1a_provide_service  i.d1b_provide_product  [pweight=wgt_7_long],
family(binomial)    eform corr( unstructured  ) cformat(%6.3f)  sformat(%6.3f)
nolstretch
```

```
GEE population-averaged model                   Number of obs      =     16544
Group and time vars:            mprid year      Number of groups   =      3092
Link:                                logit      Obs per group: min =         1
Family:                            binomial                     avg =       5.1
Correlation:                    unstructured                     max =         8
                                                Wald chi2(10)      =    214.69
Scale parameter:                           1    Prob > chi2        =    0.0000

                                  (Std. Err. adjusted for clustering on mprid)
------------------------------------------------------------------------------
             |               Robust
     Have_IP | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      1.026      0.008    3.319   0.001        1.011       1.042
1.Home_Based |      1.021      0.079    0.267   0.789        0.877       1.189
1.OO_D_edu~r |      1.441      0.117    4.488   0.000        1.228       1.690
OO_work_ex~r |      1.002      0.004    0.544   0.587        0.994       1.011
OO_race_wh~r |      0.969      0.113   -0.273   0.785        0.771       1.217
   PO_gender |      1.227      0.127    1.979   0.048        1.002       1.501
1.Comp_adv~e |      1.549      0.076    8.951   0.000        1.407       1.705
  1.hightech |      1.586      0.167    4.370   0.000        1.290       1.951
1.d1a_prov~e |      0.876      0.065   -1.799   0.072        0.758       1.012
1.d1b_prov~t |      1.547      0.095    7.111   0.000        1.372       1.745
       _cons |      0.082      0.015  -13.552   0.000        0.057       0.117
------------------------------------------------------------------------------

estat   wcorr

Estimated within-mprid correlation matrix R:

|     c1         c2         c3         c4         c5      c6         c7         c8
-------------------------------------------------------------------------------
r1 |         1
r2 |  .5557433          1
r3 |  .5224484   .6202249          1
r4 |  .4486707   .5806228   .6256935          1
r5 |  .4727351   .5234209   .5678863   .5908751          1
r6 |  .4116451   .4818799   .5589694   .5619538   .5675952          1
r7 |  .3698166   .4087015   .5054729    .557027    .587197   .6158786          1
r8 |  .3637366   .4371442   .5423424   .5949154   .5759982   .6119614   .6365366
```

```
mi  estimate:xtgee  Have_IP  LnAssets      i.Home_Based    i.OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner PO_gender  i.Comp_advantage ///
i.hightech  i.d1a_provide_service  i.d1b_provide_product  [pweight=wgt_7_long],
family(binomial)    eform corr( unstructured   ) cformat(%6.3f)   sformat(%6.3f)
nolstretch
mi estimate,  eform   cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                   Imputations        =         5
GEE population-averaged model                   Number of obs      =     18286

Group and time vars:            mprid year      Number of groups   =      3140
Link:                                logit      Obs per group: min =         1
Family:                           binomial                     avg =       5.6
Correlation:                   unstructured                    max =         8
Scale parameter:                          1
                                                Average RVI        =    0.0190
                                                Largest FMI        =    0.0889
DF adjustment:                 Large sample     DF:     min        =    546.71
                                                        avg        =  1.13e+06
                                                        max        =  8.89e+06
Model F test:                     Equal FMI     F(  10,79044.8)    =     22.06
Within VCE type:                     Robust     Prob > F           =    0.0000

                                (Within VCE adjusted for clustering on mprid)
------------------------------------------------------------------------------
     Have_IP |     exp(b)   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      1.026      0.008    3.168   0.002        1.010       1.043
1.Home_Based |      1.014      0.078    0.181   0.856        0.872       1.179
1.OO_D_edu~r |      1.478      0.116    4.967   0.000        1.267       1.724
OO_work_ex~r |      1.002      0.004    0.551   0.582        0.994       1.011
OO_race_wh~r |      0.917      0.104   -0.770   0.442        0.734       1.145
   PO_gender |      1.184      0.121    1.656   0.098        0.969       1.446
1.Comp_adv~e |      1.548      0.073    9.315   0.000        1.412       1.697
  1.hightech |      1.611      0.166    4.623   0.000        1.316       1.971
1.d1a_prov~e |      0.869      0.062   -1.968   0.049        0.755       0.999
1.d1b_prov~t |      1.501      0.094    6.502   0.000        1.328       1.697
       _cons |      0.089      0.016  -13.550   0.000        0.063       0.126
------------------------------------------------------------------------------
```

Examples 6.12 Fixed Effects Model

The fixed effects logit model can be written as

$$Logit(p_i) = Log\left[\frac{p_i}{1-p_i}\right] = \alpha_i + \beta x_{it}$$

and

$$p_{it} = \frac{e^{\alpha_i+\beta x_{it}}}{1+e^{\alpha_i+\beta x_{it}}}$$

The model cannot be estimated by full maximum-likelihood. The conditional maximum-likelihood method, which "conditions" the α_i out of the likelihood function, will be able to estimate the fixed-effects logit model. The conditional maximum-likelihood will control for stable characteristics (that do not change across time) whether measured or not. Of note, if a firm has ones or zeroes for all eight years, these response patterns are eliminated from the analysis; also, fixed effects models are not useful for looking at the effects of variables that do not change across time.

```
mi xeq 0:svy: clog Have_IP LnAssets     i.Home_Based   i.OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner PO_gender  i.Comp_advantage ///
i.hightech  i.d1a_provide_service  i.d1b_provide_product  ,group  (mprid)    or
cformat(%6.3f)  sformat(%6.3f)   nolstretch

note: multiple positive outcomes within groups encountered.
note: 2303 groups (11460 obs) dropped because of all positive or all negative
outcomes.
Survey: Conditional (fixed-effects) logistic regression
Number of strata   =         6                  Number of obs      =      5084
Number of PSUs     =       789                  Population size    = 17245.036
                                                Design df          =       783
                                                F(   9,    775)    =      7.96
                                                Prob > F           =    0.0000
------------------------------------------------------------------------------
             |               Linearized
     Have_IP | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      1.070      0.021    3.479   0.001        1.030       1.111
1.Home_Based |      0.928      0.273   -0.255   0.799        0.521       1.652
1.OO_D_edu~r |      1.054      0.282    0.195   0.845        0.623       1.782
OO_work_ex~r |      1.018      0.033    0.559   0.577        0.956       1.084
OO_race_wh~r |      0.271      0.199   -1.782   0.075        0.064       1.142
   PO_gender |      1.000  (omitted)
1.Comp_adv~e |      2.009      0.234    5.991   0.000        1.599       2.526
  1.hightech |      0.995      0.379   -0.012   0.990        0.471       2.102
1.d1a_prov~e |      1.168      0.263    0.691   0.490        0.751       1.818
1.d1b_prov~t |      1.694      0.261    3.419   0.001        1.252       2.293
------------------------------------------------------------------------------
```

```
mi estimate, esampvaryok :svy: clog Have_IP LnAssets i.Home_Based
i.OO_D_education_owner OO_work_exp_owner OO_race_white_owner PO_gender
i.Comp_advantage i.hightech i.d1a_provide_service i.d1b_provide_product ,group
(mprid) ///
mi estimate, or cformat(%6.3f) sformat(%6.3f) nolstretch

Multiple-imputation estimates
Survey: Conditional (fixed-effects) logistic regression
                                            Imputations        =         5
                                            Number of obs      =      5817
Number of strata  =         6               Population size    = 18462.974
Number of PSUs    =       837
                                            Average RVI        =    0.0675
                                            Largest FMI        =    0.3182
                                            Complete DF        =       831
DF adjustment:   Small sample               DF:     min        =     44.11
                                                    avg        =    651.22
                                                    max        =    828.64
Model F test:       Equal FMI               F(   9,  742.5)    =      6.71
Within VCE type:   Linearized               Prob > F           =    0.0000
------------------------------------------------------------------------------
     Have_IP | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      1.057      0.021    2.760   0.006        1.016       1.099
1.Home_Based |      0.987      0.283   -0.047   0.963        0.561       1.734
1.OO_D_edu~r |      1.133      0.281    0.504   0.615        0.697       1.842
OO_work_ex~r |      1.011      0.034    0.317   0.751        0.946       1.080
OO_race_wh~r |      0.481      0.392   -0.897   0.375        0.093       2.489
   PO_gender |      1.000  (omitted)
1.Comp_adv~e |      2.011      0.225    6.244   0.000        1.615       2.505
  1.hightech |      1.066      0.380    0.180   0.857        0.530       2.146
1.d1a_prov~e |      1.188      0.266    0.770   0.442        0.766       1.843
1.d1b_prov~t |      1.565      0.238    2.949   0.003        1.162       2.108
------------------------------------------------------------------------------
Warning: estimation sample varies across imputations; results may be biased.
         Sample sizes vary between 5817 and 5845.
Note: number of primary clusters varies among imputations.
Note: population size varies among imputations.
```

Examples 6.13 Random Effects (Random-Intercept)

```
*Gllamm
mi xeq 0:xi: gllamm Have_IP LnAssets    i.Home_Based   i.OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner PO_gender  i.Comp_advantage ///
i.hightech i.d1a_provide_service i.d1b_provide_product   ,i(mprid) pweight(pwt)
nip(30) adapt   link(logit) fam(binom) eform

Running adaptive quadrature
Iteration 0:    log likelihood = -49224.013
Iteration 1:    log likelihood = -39226.504
Iteration 2:    log likelihood = -38457.994
Iteration 3:    log likelihood = -38230.506
Iteration 4:    log likelihood = -38228.513
Iteration 5:    log likelihood = -38228.513


Adaptive quadrature has converged, running Newton-Raphson
Iteration 0:   log likelihood = -38228.513
Iteration 1:   log likelihood = -38228.513  (backed up)
Iteration 2:   log likelihood = -38228.512

number of level 1 units = 16544
number of level 2 units = 3092

Condition Number = 83.584689

gllamm model

log likelihood = -38228.512

Robust standard errors
------------------------------------------------------------------------------
            Have_IP |      exp(b)    Std. Err.       z     P>|z|      [95% Conf.
Interval]
--------------------+---------------------------------------------------------
           LnAssets |   1.063503   .0175671     3.73   0.000     1.029623    1.098497
      _IHome_Base_1 |   .9764546    .156579    -0.15   0.882     .7131124    1.337045
      _IOO_D_educ_1 |   1.935961   .3152606     4.06   0.000     1.406964    2.663852
  OO_work_exp_owner |   1.004006   .0092292     0.43   0.664     .9860785    1.022258
OO_race_white_owner |   .8006374   .1959506    -0.91   0.364     .4955757    1.293486
          PO_gender |   1.497545   .3130597     1.93   0.053     .9941167    2.255913
      _IComp_adva_1 |   2.578003   .2735617     8.92   0.000     2.093916    3.174005
       _Ihightech_1 |   2.337021   .6051605     3.28   0.001     1.406851    3.882194
      _Id1a_provi_1 |   .7856036   .1375449    -1.38   0.168     .5574085    1.107219
      _Id1b_provi_1 |    2.82769   .3787664     7.76   0.000     2.174775    3.676626
              _cons |   .0041076   .0016238   -13.90   0.000     .0018927    .0089144
------------------------------------------------------------------------------


Variances and covariances of random effects
------------------------------------------------------------------------------


***level 2 (mprid)

    var(1): 11.198038 (.88019865)
------------------------------------------------------------------------------
```

```
mi estimate,cmdok: gllamm Have_IP LnAssets    Home_Based   OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner PO_gender  Comp_advantage ///
hightech d1a_provide_service d1b_provide_product  ,i(mprid) pweight(pwt) nip(30)
adapt   link(logit) fam(binom) eform
mi estimate,eform   cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                     Imputations     =          5
gllamm model                                      Number of obs   =      18286
                                                  Average RVI     =     0.0119
                                                  Largest FMI     =     0.0569
DF adjustment:   Large sample                     DF:     min     =    1302.69
                                                          avg     =  346554.32
Within VCE type:          OIM                             max     = 1570507.11

------------------------------------------------------------------------------
     Have_IP |     exp(b)   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Have_IP      |
    LnAssets |      1.063      0.018    3.583   0.000        1.028       1.098
  Home_Based |      0.950      0.150   -0.324   0.746        0.697       1.295
OO_D_educa~r |      1.994      0.316    4.357   0.000        1.462       2.720
OO_work_ex~r |      1.003      0.009    0.309   0.757        0.985       1.021
OO_race_wh~r |      0.758      0.184   -1.140   0.254        0.471       1.221
   PO_gender |      1.448      0.298    1.799   0.072        0.967       2.167
Comp_advan~e |      2.545      0.257    9.241   0.000        2.087       3.102
    hightech |      2.373      0.582    3.525   0.000        1.468       3.837
d1a_provid~e |      0.797      0.136   -1.328   0.184        0.570       1.114
d1b_provid~t |      2.638      0.344    7.434   0.000        2.043       3.407
       _cons |      0.005      0.002  -13.993   0.000        0.002       0.010
-------------+----------------------------------------------------------------
mpri1        |
       _cons |      3.345      0.127   26.324   0.000        3.096       3.594
------------------------------------------------------------------------------
```

Examples 6.14 Hybrid Model

```
mi xeq 0:xi: gllamm Have_IP   m_LnAssets       m_Home_Based
m_OO_D_education_owner     m_OO_work_exp_owner        m_OO_race_white_owner
       m_Comp_advantage PO_gender ///
d_LnAssets    d_Home_Based d_OO_D_education_owner     d_OO_work_exp_owner
       d_OO_race_white_owner      d_Comp_advantage  ,i(mprid) pweight(pwt) nip(30)
adapt   link(logit) fam(binom) eform

Running adaptive quadrature
Adaptive quadrature has converged, running Newton-Raphson

number of level 1 units = 16544
number of level 2 units = 3092

Condition Number = 171.36288

gllamm model

log likelihood = -38188.256

Robust standard errors
 Robust standard errors

                              Std.                    [95%
 Have_IP                 exp(b)  Err.    z       P>z    Conf.     Interval]

 m_LnAssets               1.047  0.040   1.200  0.229     0.971      1.129
 m_Home_Based             1.157  0.244   0.690  0.488     0.766      1.748
 m_OO_D_education_owner   2.741  0.546   5.060  0.000     1.855      4.051
 m_OO_work_exp_owner      0.999  0.010  -0.070  0.942     0.980      1.019
 m_OO_race_white_owner    0.810  0.221  -0.770  0.440     0.475      1.382
 m_Comp_advantage        28.485  8.299  11.500  0.000    16.092     50.422
 PO_gender                1.545  0.340   1.980  0.048     1.004      2.379
 d_LnAssets               1.070  0.019   3.770  0.000     1.033      1.109
 d_Home_Based             0.852  0.244  -0.560  0.575     0.486      1.493
 d_OO_D_education_owner   1.026  0.298   0.090  0.931     0.581      1.811
 d_OO_work_exp_owner      1.009  0.030   0.290  0.773     0.951      1.069
 d_OO_race_white_owner    0.319  0.235  -1.550  0.121     0.075      1.354
 d_Comp_advantage         1.933  0.213   5.990  0.000     1.558      2.398
                                          -
 _cons                    0.001  0.001  12.300  0.000     0.000      0.003

Variances and covariances of random effects
------------------------------------------------------------------------------


***level 2 (mprid)

    var(1): 12.115952 (.97713957)
------------------------------------------------------------------------------
```

```
test (m_LnAssets=d_LnAssets) ///
(m_Home_Based=d_Home_Based ) ///
(m_OO_D_education_owner=d_OO_D_education_owner  ) ///
(m_OO_work_exp_owner=  d_OO_work_exp_owner) ///
(m_OO_race_white_owner=d_OO_race_white_owner) ///
(m_Comp_advantage=d_Comp_advantage) ///
(m_Comp_advantage=d_Comp_advantage )


( 1)  [Have_IP]m_LnAssets - [Have_IP]d_LnAssets = 0
 ( 2)  [Have_IP]m_Home_Based - [Have_IP]d_Home_Based = 0
 ( 3)  [Have_IP]m_OO_D_education_owner - [Have_IP]d_OO_D_education_owner = 0
 ( 4)  [Have_IP]m_OO_work_exp_owner - [Have_IP]d_OO_work_exp_owner = 0
 ( 5)  [Have_IP]m_OO_race_white_owner - [Have_IP]d_OO_race_white_owner = 0
 ( 6)  [Have_IP]m_Comp_advantage - [Have_IP]d_Comp_advantage = 0
 ( 7)  [Have_IP]m_Comp_advantage - [Have_IP]d_Comp_advantage = 0
       Constraint 7 dropped

           chi2(  6) =   87.88
         Prob > chi2 =    0.0000

mi estimate,cmdok: gllamm Have_IP   m_LnAssets       m_Home_Based
m_OO_D_education_owner     m_OO_work_exp_owner       m_OO_race_white_owner
       m_Comp_advantage PO_gender ///
d_LnAssets   d_Home_Based d_OO_D_education_owner     d_OO_work_exp_owner
       d_OO_race_white_owner     d_Comp_advantage  ,i(mprid) pweight(pwt) nip(30)
adapt   link(logit) fam(binom) eform

mi estimate,eform   cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                     Imputations     =          5
gllamm model                                      Number of obs   =      18286
                                                  Average RVI     =     0.0263
                                                  Largest FMI     =     0.2247
DF adjustment:   Large sample                     DF:     min     =      92.37
                                                          avg     =  413367.44
Within VCE type:          OIM                             max     = 2375363.55

------------------------------------------------------------------------------
     Have_IP |     exp(b)   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Have_IP      |
  m_LnAssets |      1.077      0.042    1.929   0.054        0.999       1.162
m_Home_Based |      1.203      0.248    0.895   0.371        0.803       1.803
m_OO_D_edu~r |      2.889      0.568    5.392   0.000        1.964       4.248
m_OO_work_~r |      0.999      0.010   -0.145   0.885        0.980       1.018
m_OO_race_~r |      0.712      0.194   -1.246   0.213        0.417       1.215
m_Comp_adv~e |     25.025      7.172   11.234   0.000       14.269      43.888
   PO_gender |      1.458      0.317    1.731   0.083        0.951       2.234
  d_LnAssets |      1.062      0.019    3.280   0.001        1.024       1.101
d_Home_Based |      0.832      0.236   -0.650   0.516        0.477       1.450
d_OO_D_edu~r |      1.117      0.304    0.407   0.684        0.656       1.904
d_OO_work_~r |      1.003      0.031    0.105   0.917        0.944       1.066
d_OO_race_~r |      0.518      0.357   -0.954   0.343        0.132       2.036
d_Comp_adv~e |      1.963      0.206    6.426   0.000        1.598       2.412
       _cons |      0.001      0.001  -12.618   0.000        0.000       0.003
-------------+----------------------------------------------------------------
mpri1        |
       _cons |      3.488      0.136   25.643   0.000        3.222       3.755
------------------------------------------------------------------------------
```

6.5.2 Multinomial Logit Models for Catagorical Response Variables

Suppose we have a nominal dependent variable y_j where i indexes choices (k of them) and j indexes firms (n of them). Multinomial Logit is designed to estimate what observed covariates (x_j) predict choice of *m* from the other *k-1* alternatives. The probability you choose for option *i* is

$$p_{ij=}\Pr(y_j = i) = \begin{cases} \dfrac{1}{1+\sum_{m=2}^{k} e^{x_j\beta_m}}, if\ i = 1 \\ \dfrac{e^{\beta_j x_i}}{1+\sum_{m=2}^{k} e^{x_j\beta_m}}, if\ i > 1 \end{cases}$$

where β_m is the coefficient vector for outcome *m*.

The multinomial logit is designed for outcomes that are not interrelated (Independence of Irrelevant Alternatives (IIA)). The IIA assumption means that the odds of one outcome versus another should be independent of other alternatives.

Examples 6.15 Robust Standard Errors

```
mi   xeq   0:svy:   mlogit   Legal_Form   i.Have_IP   LnAssets        i.Home_Based
i.OO_D_education_owner      OO_work_exp_owner      OO_race_white_owner   PO_gender
i.Comp_advantage ///
i.hightech i.d1a_provide_service i.d1b_provide_product   ,    rrr   cformat(%6.3f)
sformat(%6.3f)   nolstretch  baseoutcome(1)




Survey: Multinomial logistic regression

Number of strata   =         6                  Number of obs      =     16544
Number of PSUs     =      3092                  Population size    =  368623.1
                                                 Design df          =      3086
                                                 F(  33,   3054)    =     11.55
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
  Legal_Form |        RRR   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |  (base outcome)
-------------+----------------------------------------------------------------
2            |
   1.Have_IP |      1.590      0.188     3.915   0.000        1.260       2.006
    LnAssets |      1.124      0.018     7.441   0.000        1.090       1.160
1.Home_Based |      0.555      0.063    -5.155   0.000        0.444       0.695
1.OO_D_edu~r |      1.834      0.203     5.473   0.000        1.476       2.279
OO_work_ex~r |      0.999      0.006    -0.219   0.827        0.987       1.010
OO_race_wh~r |      1.326      0.219     1.707   0.088        0.959       1.833
   PO_gender |      1.816      0.236     4.596   0.000        1.408       2.342
1.Comp_adv~e |      0.876      0.071    -1.632   0.103        0.747       1.027
  1.hightech |      1.673      0.301     2.862   0.004        1.176       2.381
1.d1a_prov~e |      0.726      0.108    -2.151   0.032        0.543       0.972
1.d1b_prov~t |      0.588      0.064    -4.892   0.000        0.475       0.728
       _cons |      0.274      0.084    -4.231   0.000        0.150       0.499
-------------+----------------------------------------------------------------
```

```
3             |
   1.Have_IP |      1.606      0.212     3.580   0.000        1.239       2.081
    LnAssets |      1.161      0.020     8.605   0.000        1.122       1.201
1.Home_Based |      0.365      0.046    -7.914   0.000        0.285       0.469
1.OO_D_edu~r |      1.214      0.147     1.600   0.110        0.957       1.541
OO_work_ex~r |      0.997      0.006    -0.479   0.632        0.984       1.010
OO_race_wh~r |      1.261      0.235     1.242   0.214        0.874       1.818
   PO_gender |      1.617      0.236     3.293   0.001        1.215       2.153
1.Comp_adv~e |      0.879      0.081    -1.404   0.160        0.735       1.052
  1.hightech |      1.916      0.363     3.435   0.001        1.322       2.777
1.d1a_prov~e |      1.096      0.183     0.545   0.586        0.789       1.521
1.d1b_prov~t |      0.729      0.086    -2.663   0.008        0.578       0.920
       _cons |      0.155      0.048    -6.001   0.000        0.084       0.285
-------------+----------------------------------------------------------------
4             |
   1.Have_IP |      1.885      0.280     4.260   0.000        1.408       2.523
    LnAssets |      1.075      0.019     4.081   0.000        1.038       1.113
1.Home_Based |      0.319      0.049    -7.501   0.000        0.237       0.430
1.OO_D_edu~r |      1.005      0.148     0.033   0.974        0.753       1.342
OO_work_ex~r |      0.995      0.007    -0.684   0.494        0.980       1.010
OO_race_wh~r |      0.863      0.172    -0.735   0.462        0.584       1.277
   PO_gender |      1.446      0.249     2.136   0.033        1.031       2.028
1.Comp_adv~e |      1.039      0.112     0.351   0.725        0.841       1.283
  1.hightech |      1.756      0.403     2.451   0.014        1.119       2.754
1.d1a_prov~e |      0.802      0.150    -1.176   0.240        0.555       1.159
1.d1b_prov~t |      0.689      0.095    -2.710   0.007        0.526       0.902
       _cons |      0.340      0.125    -2.930   0.003        0.165       0.700
------------------------------------------------------------------------------

mi   xeq   0:      mlogit   Legal_Form   i.Have_IP   LnAssets        i.Home_Based
i.OO_D_education_owner     OO_work_exp_owner     OO_race_white_owner   PO_gender
i.Comp_advantage ///
i.hightech  i.d1a_provide_service  i.d1b_provide_product    [pweight=wgt_7_long],
vce(cluster   mprid)     rrr     cformat(%6.3f)     sformat(%6.3f)       nolstretch
baseoutcome(1)

Multinomial logistic regression                  Number of obs   =      16544
                                                 Wald chi2(33)   =     382.09
                                                 Prob > chi2     =     0.0000
Log pseudolikelihood = -456636.78                Pseudo R2       =     0.0574

                              (Std. Err. adjusted for 3092 clusters in mprid)
------------------------------------------------------------------------------
             |               Robust
  Legal_Form |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |  (base outcome)
-------------+----------------------------------------------------------------
2            |
   1.Have_IP |      1.590      0.189     3.909   0.000        1.260       2.007
    LnAssets |      1.124      0.018     7.443   0.000        1.090       1.160
1.Home_Based |      0.555      0.063    -5.148   0.000        0.444       0.695
1.OO_D_edu~r |      1.834      0.203     5.468   0.000        1.476       2.279
OO_work_ex~r |      0.999      0.006    -0.219   0.827        0.987       1.010
OO_race_wh~r |      1.326      0.219     1.707   0.088        0.959       1.832
   PO_gender |      1.816      0.236     4.584   0.000        1.407       2.343
1.Comp_adv~e |      0.876      0.071    -1.633   0.103        0.747       1.027
  1.hightech |      1.673      0.301     2.863   0.004        1.176       2.380
1.d1a_prov~e |      0.726      0.108    -2.152   0.031        0.543       0.972
1.d1b_prov~t |      0.588      0.064    -4.885   0.000        0.475       0.728
       _cons |      0.274      0.084    -4.230   0.000        0.150       0.499
-------------+----------------------------------------------------------------
```

```
3             |
   1.Have_IP |      1.606      0.213     3.576   0.000        1.239       2.081
    LnAssets |      1.161      0.020     8.597   0.000        1.122       1.201
1.Home_Based |      0.365      0.046    -7.916   0.000        0.285       0.469
1.OO_D_edu~r |      1.214      0.147     1.600   0.110        0.957       1.540
OO_work_ex~r |      0.997      0.006    -0.479   0.632        0.984       1.010
OO_race_wh~r |      1.261      0.235     1.242   0.214        0.875       1.818
   PO_gender |      1.617      0.237     3.282   0.001        1.214       2.154
1.Comp_adv~e |      0.879      0.081    -1.404   0.160        0.735       1.052
  1.hightech |      1.916      0.363     3.435   0.001        1.322       2.777
1.d1a_prov~e |      1.096      0.184     0.545   0.586        0.789       1.521
1.d1b_prov~t |      0.729      0.086    -2.663   0.008        0.578       0.920
       _cons |      0.155      0.048    -5.984   0.000        0.084       0.286
-------------+------------------------------------------------------------------
4             |
   1.Have_IP |      1.885      0.281     4.258   0.000        1.408       2.523
    LnAssets |      1.075      0.019     4.082   0.000        1.038       1.113
1.Home_Based |      0.319      0.049    -7.501   0.000        0.237       0.430
1.OO_D_edu~r |      1.005      0.148     0.033   0.974        0.753       1.342
OO_work_ex~r |      0.995      0.007    -0.684   0.494        0.980       1.010
OO_race_wh~r |      0.863      0.172    -0.735   0.462        0.584       1.277
   PO_gender |      1.446      0.250     2.127   0.033        1.029       2.030
1.Comp_adv~e |      1.039      0.112     0.351   0.725        0.841       1.283
  1.hightech |      1.756      0.403     2.451   0.014        1.119       2.753
1.d1a_prov~e |      0.802      0.150    -1.176   0.239        0.555       1.158
1.d1b_prov~t |      0.689      0.095    -2.704   0.007        0.526       0.903
       _cons |      0.340      0.125    -2.931   0.003        0.166       0.700
------------------------------------------------------------------------------
```

```
mi estimate:svy: mlogit Legal_Form i.Have_IP LnAssets i.Home_Based
i.OO_D_education_owner OO_work_exp_owner OO_race_white_owner PO_gender
i.Comp_advantage ///
i.hightech i.d1a_provide_service i.d1b_provide_product , rrr cformat(%6.3f)
sformat(%6.3f) nolstretch baseoutcome(1)
mi estimate, rrr cformat(%6.3f) sformat(%6.3f) nolstretch
```

```
Multiple-imputation estimates                   Imputations       =          5
Survey: Multinomial logistic regression         Number of obs     =      18286

Number of strata  =         6                   Population size   = 408495.43
Number of PSUs    =      3140
                                                Average RVI       =     0.0007
                                                Largest FMI       =     0.0062
                                                Complete DF       =       3134
DF adjustment:   Small sample                   DF:     min       =    3024.35
                                                        avg       =    3124.75
                                                        max       =    3131.98
Model F test:       Equal FMI                   F(  33, 3131.9)   =      11.94
Within VCE type:   Linearized                   Prob > F          =     0.0000

------------------------------------------------------------------------------
  Legal_Form |        RRR   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |  (base outcome)
-------------+----------------------------------------------------------------
2            |
   1.Have_IP |      1.557      0.181    3.820   0.000        1.241       1.955
    LnAssets |      1.123      0.017    7.508   0.000        1.090       1.158
1.Home_Based |      0.539      0.060   -5.526   0.000        0.432       0.671
1.OO_D_edu~r |      1.762      0.192    5.207   0.000        1.423       2.180
OO_work_ex~r |      0.999      0.006   -0.199   0.842        0.988       1.010
OO_race_wh~r |      1.304      0.209    1.656   0.098        0.952       1.786
   PO_gender |      1.790      0.227    4.580   0.000        1.395       2.296
1.Comp_adv~e |      0.912      0.073   -1.157   0.247        0.780       1.066
  1.hightech |      1.674      0.303    2.847   0.004        1.174       2.387
1.d1a_prov~e |      0.762      0.112   -1.846   0.065        0.571       1.017
1.d1b_prov~t |      0.596      0.063   -4.887   0.000        0.485       0.734
       _cons |      0.273      0.082   -4.322   0.000        0.152       0.492
-------------+----------------------------------------------------------------
3            |
   1.Have_IP |      1.557      0.199    3.453   0.001        1.211       2.001
    LnAssets |      1.154      0.020    8.427   0.000        1.117       1.194
1.Home_Based |      0.349      0.044   -8.427   0.000        0.274       0.446
1.OO_D_edu~r |      1.164      0.138    1.277   0.202        0.922       1.469
OO_work_ex~r |      0.997      0.006   -0.431   0.667        0.985       1.010
OO_race_wh~r |      1.306      0.235    1.482   0.139        0.917       1.859
   PO_gender |      1.642      0.236    3.448   0.001        1.238       2.176
1.Comp_adv~e |      0.920      0.082   -0.939   0.348        0.772       1.096
  1.hightech |      1.912      0.363    3.409   0.001        1.317       2.775
1.d1a_prov~e |      1.107      0.182    0.617   0.537        0.802       1.527
1.d1b_prov~t |      0.739      0.085   -2.622   0.009        0.590       0.927
       _cons |      0.156      0.048   -6.069   0.000        0.086       0.285
-------------+----------------------------------------------------------------
4            |
   1.Have_IP |      1.926      0.276    4.569   0.000        1.454       2.552
    LnAssets |      1.083      0.020    4.225   0.000        1.043       1.123
1.Home_Based |      0.306      0.046   -7.908   0.000        0.229       0.411
1.OO_D_edu~r |      1.007      0.145    0.047   0.962        0.759       1.336
OO_work_ex~r |      0.997      0.007   -0.424   0.672        0.983       1.011
OO_race_wh~r |      0.888      0.174   -0.606   0.544        0.604       1.305
   PO_gender |      1.480      0.252    2.305   0.021        1.060       2.066
1.Comp_adv~e |      1.072      0.114    0.652   0.514        0.870       1.321
  1.hightech |      1.799      0.406    2.601   0.009        1.155       2.802
1.d1a_prov~e |      0.769      0.139   -1.455   0.146        0.540       1.096
1.d1b_prov~t |      0.654      0.090   -3.097   0.002        0.499       0.855
       _cons |      0.321      0.116   -3.131   0.002        0.157       0.654
------------------------------------------------------------------------------
```

Examples 6.16 Fixed Effects Model

For Multinomial Logit Models, conditional ML is neither possible nor available yet. Another approach is to use multiple logistic regression analyses: one for each pair of outcomes (a set of binomial models estimated separately). One potential problem with this approach is that each analysis is run on a different sample and doing it sequentially could lead to misestimation of the standard errors; thus, we highly recommend avoid using this approach.

```
tab Legal_Form,gen(Legal_Form)

mi   xeq   0:svy:   clog   Legal_Form2   i.Have_IP   LnAssets        i.Home_Based
i.OO_D_education_owner      OO_work_exp_owner      OO_race_white_owner   PO_gender
i.Comp_advantage ///
i.hightech   i.d1a_provide_service   i.d1b_provide_product   if   Legal_Form==1   |
Legal_Form==2 ,group(mprid) or     cformat(%6.3f)  sformat(%6.3f)   nolstretch

Survey: Conditional (fixed-effects) logistic regression

Number of strata   =         6                  Number of obs      =       619
Number of PSUs     =        92                  Population size    = 2169.6744
                                                Design df          =        86
                                                F(  10,     77)    =     54.90
                                                Prob > F           =    0.0000

------------------------------------------------------------------------------
             |               Linearized
 Legal_Form2 | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   1.Have_IP |      2.598      1.640    1.512   0.134        0.741       9.113
    LnAssets |      1.081      0.078    1.090   0.279        0.938       1.247
1.Home_Based |      0.201      0.163   -1.985   0.050        0.040       1.002
1.OO_D_edu~r |      1.870      2.861    0.409   0.683        0.089      39.135
OO_work_ex~r |      0.960      0.062   -0.631   0.530        0.845       1.091
OO_race_wh~r |      0.000      0.000  -20.023   0.000        0.000       0.000
   PO_gender |      1.000  (omitted)
1.Comp_adv~e |      0.809      0.232   -0.739   0.462        0.457       1.432
  1.hightech |      0.526      0.586   -0.576   0.566        0.057       4.825
1.d1a_prov~e |      0.319      0.196   -1.857   0.067        0.094       1.084
1.d1b_prov~t |      0.172      0.080   -3.784   0.000        0.068       0.434
------------------------------------------------------------------------------

mi   xeq   0:svy:   clog   Legal_Form3   i.Have_IP   LnAssets        i.Home_Based
i.OO_D_education_owner      OO_work_exp_owner      OO_race_white_owner   PO_gender
i.Comp_advantage ///
i.hightech   i.d1a_provide_service   i.d1b_provide_product   if   Legal_Form==1   |
Legal_Form==3 ,group(mprid) or     cformat(%6.3f)  sformat(%6.3f)   nolstretch
```

```
Survey: Conditional (fixed-effects) logistic regression

Number of strata   =         5                  Number of obs      =       353
Number of PSUs     =        58                  Population size    = 1275.7755
                                                Design df          =        53
                                                F(  10,     44)    =     82.17
                                                Prob > F           =    0.0000

------------------------------------------------------------------------------
             |               Linearized
 Legal_Form3 | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   1.Have_IP |      1.970      1.320    1.011   0.317        0.513       7.557
    LnAssets |      2.083      0.649    2.356   0.022        1.115       3.892
1.Home_Based |      0.331      0.272   -1.346   0.184        0.064       1.719
1.OO_D_edu~r |      0.000      0.000  -23.276   0.000        0.000       0.000
OO_work_ex~r |      0.864      0.073   -1.725   0.090        0.729       1.024
OO_race_wh~r |     77.731    365.695    0.925   0.359        0.006    9.74e+05
   PO_gender |      1.000  (omitted)
1.Comp_adv~e |      0.724      0.285   -0.821   0.416        0.329       1.595
  1.hightech |      3.621      8.674    0.537   0.593        0.030     441.817
1.d1a_prov~e |      8.507      5.573    3.268   0.002        2.287      31.653
1.d1b_prov~t |      0.800      0.781   -0.229   0.820        0.113       5.679
------------------------------------------------------------------------------

mi   xeq   0:svy:   clog   Legal_Form4   i.Have_IP   LnAssets        i.Home_Based
i.OO_D_education_owner      OO_work_exp_owner      OO_race_white_owner   PO_gender
i.Comp_advantage ///
i.hightech   i.d1a_provide_service   i.d1b_provide_product   if   Legal_Form==1   |
Legal_Form==4 ,group(mprid) or     cformat(%6.3f)  sformat(%6.3f)   nolstretch

Survey: Conditional (fixed-effects) logistic regression

Number of strata   =         4                  Number of obs      =       154
Number of PSUs     =        27                  Population size    = 636.09717
                                                Design df          =        23
                                                F(  10,     14)    =     36.76
                                                Prob > F           =    0.0000

------------------------------------------------------------------------------
             |               Linearized
 Legal_Form4 | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   1.Have_IP |      3.549      3.810    1.180   0.250        0.385      32.699
    LnAssets |      1.128      0.095    1.429   0.167        0.948       1.341
1.Home_Based |      0.000      0.000  -12.908   0.000        0.000       0.000
1.OO_D_edu~r |      1.249      2.485    0.112   0.912        0.020      76.659
OO_work_ex~r |      1.048      0.085    0.580   0.568        0.886       1.241
OO_race_wh~r |      0.638      2.251   -0.127   0.900        0.000     938.246
   PO_gender |      1.000  (omitted)
1.Comp_adv~e |      0.291      0.206   -1.741   0.095        0.067       1.262
  1.hightech |      0.787      1.464   -0.129   0.899        0.017      36.943
1.d1a_prov~e |      0.162      0.148   -1.996   0.058        0.025       1.068
1.d1b_prov~t |      0.074      0.066   -2.902   0.008        0.012       0.473
------------------------------------------------------------------------------
```

Examples 6.17 Hybrid Model

```
  mi  xeq  0:svy:  mlogit  Legal_Form  m_Have_IP  m_LnAssets       m_Home_Based
m_OO_D_education_owner   m_OO_work_exp_owner   m_OO_race_white_owner  i.PO_gender
m_Comp_advantage m_hightech m_d1a_provide_service m_d1b_provide_product ///
                                d_Have_IP    d_LnAssets           d_Home_Based
d_OO_D_education_owner         d_OO_work_exp_owner         d_OO_race_white_owner
d_Comp_advantage d_hightech d_d1a_provide_service d_d1b_provide_product ///
  ,   rrr  cformat(%6.3f)  sformat(%6.3f)   nolstretch  baseoutcome(1)

Survey: Multinomial logistic regression

Number of strata   =         6                  Number of obs      =     16544
Number of PSUs     =      3092                  Population size    =  368623.1
                                                 Design df          =      3086
                                                 F(  63,   3024)    =      6.05
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
  Legal_Form |        RRR   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |  (base outcome)
-------------+----------------------------------------------------------------
2            |
   m_Have_IP |      2.076      0.412    3.684   0.000        1.407       3.062
  m_LnAssets |      1.225      0.034    7.417   0.000        1.161       1.293
m_Home_Based |      0.577      0.076   -4.174   0.000        0.445       0.747
m_OO_D_edu~r |      1.869      0.230    5.077   0.000        1.468       2.380
m_OO_work_~r |      0.997      0.006   -0.454   0.650        0.985       1.009
m_OO_race_~r |      1.254      0.216    1.313   0.189        0.895       1.757
 1.PO_gender |      1.734      0.230    4.154   0.000        1.337       2.248
m_Comp_adv~e |      0.777      0.134   -1.471   0.141        0.554       1.088
  m_hightech |      1.903      0.403    3.038   0.002        1.256       2.882
m_d1a_prov~e |      0.583      0.133   -2.362   0.018        0.373       0.913
m_d1b_prov~t |      0.441      0.070   -5.133   0.000        0.322       0.603
   d_Have_IP |      1.144      0.058    2.644   0.008        1.035       1.264
  d_LnAssets |      1.004      0.004    0.845   0.398        0.995       1.012
d_Home_Based |      0.932      0.096   -0.680   0.496        0.762       1.141
d_OO_D_edu~r |      0.930      0.085   -0.793   0.428        0.776       1.113
d_OO_work_~r |      0.999      0.014   -0.064   0.949        0.972       1.027
d_OO_race_~r |      0.935      0.308   -0.204   0.838        0.490       1.785
d_Comp_adv~e |      0.924      0.029   -2.547   0.011        0.869       0.982
  d_hightech |      0.918      0.064   -1.232   0.218        0.800       1.052
d_d1a_prov~e |      0.931      0.064   -1.032   0.302        0.813       1.066
d_d1b_prov~t |      0.932      0.042   -1.564   0.118        0.853       1.018
       _cons |      0.182      0.078   -3.994   0.000        0.079       0.420
-------------+----------------------------------------------------------------
3            |
   m_Have_IP |      2.152      0.476    3.462   0.001        1.394       3.321
  m_LnAssets |      1.279      0.038    8.391   0.000        1.208       1.355
m_Home_Based |      0.376      0.055   -6.648   0.000        0.282       0.502
m_OO_D_edu~r |      1.193      0.161    1.304   0.192        0.915       1.554
m_OO_work_~r |      0.995      0.007   -0.710   0.477        0.982       1.008
m_OO_race_~r |      1.177      0.227    0.847   0.397        0.807       1.717
 1.PO_gender |      1.512      0.226    2.767   0.006        1.128       2.027
m_Comp_adv~e |      0.780      0.151   -1.282   0.200        0.533       1.141
  m_hightech |      2.181      0.484    3.514   0.000        1.412       3.370
m_d1a_prov~e |      1.066      0.275    0.250   0.803        0.643       1.768
m_d1b_prov~t |      0.597      0.103   -2.988   0.003        0.426       0.838
   d_Have_IP |      1.134      0.057    2.522   0.012        1.028       1.251
  d_LnAssets |      1.013      0.005    2.813   0.005        1.004       1.021
```

```
d_Home_Based |      0.816      0.092   -1.811   0.070        0.655       1.017
d_OO_D_edu~r |      0.918      0.108   -0.726   0.468        0.729       1.156
d_OO_work_~r |      0.983      0.012   -1.369   0.171        0.959       1.007
d_OO_race_~r |      0.985      0.456   -0.033   0.974        0.397       2.441
d_Comp_adv~e |      0.926      0.029   -2.489   0.013        0.872       0.984
  d_hightech |      1.040      0.110    0.367   0.714        0.845       1.279
d_d1a_prov~e |      1.007      0.054    0.124   0.901        0.906       1.119
d_d1b_prov~t |      0.962      0.048   -0.771   0.441        0.873       1.061
       _cons |      0.078      0.034   -5.905   0.000        0.034       0.182
-------------+----------------------------------------------------------------
4            |
   m_Have_IP |      2.701      0.658    4.079   0.000        1.675       4.356
  m_LnAssets |      1.138      0.036    4.056   0.000        1.069       1.212
m_Home_Based |      0.301      0.052   -6.915   0.000        0.215       0.424
m_OO_D_edu~r |      0.938      0.154   -0.386   0.699        0.680       1.296
m_OO_work_~r |      0.993      0.008   -0.937   0.349        0.978       1.008
m_OO_race_~r |      0.823      0.169   -0.946   0.344        0.550       1.232
 1.PO_gender |      1.402      0.246    1.926   0.054        0.994       1.978
m_Comp_adv~e |      1.075      0.244    0.319   0.750        0.689       1.679
  m_hightech |      1.970      0.522    2.559   0.011        1.172       3.313
m_d1a_prov~e |      0.673      0.192   -1.392   0.164        0.385       1.176
m_d1b_prov~t |      0.529      0.107   -3.153   0.002        0.356       0.786
   d_Have_IP |      1.102      0.087    1.233   0.218        0.944       1.286
  d_LnAssets |      0.999      0.007   -0.176   0.860        0.986       1.012
d_Home_Based |      1.031      0.119    0.262   0.793        0.822       1.293
d_OO_D_edu~r |      1.068      0.157    0.448   0.654        0.801       1.424
d_OO_work_~r |      1.020      0.018    1.137   0.256        0.986       1.056
d_OO_race_~r |      0.828      0.547   -0.285   0.775        0.227       3.022
d_Comp_adv~e |      0.951      0.043   -1.116   0.265        0.870       1.039
  d_hightech |      0.935      0.130   -0.484   0.629        0.711       1.229
d_d1a_prov~e |      1.007      0.072    0.091   0.927        0.875       1.158
d_d1b_prov~t |      0.963      0.057   -0.646   0.518        0.858       1.081
       _cons |      0.276      0.140   -2.544   0.011        0.102       0.744
------------------------------------------------------------------------------
```

```
mi estimate:svy: mlogit Legal_Form m_Have_IP m_LnAssets m_Home_Based
m_OO_D_education_owner m_OO_work_exp_owner m_OO_race_white_owner i.PO_gender
m_Comp_advantage m_hightech m_d1a_provide_service m_d1b_provide_product ///
                              d_Have_IP  d_LnAssets          d_Home_Based
d_OO_D_education_owner      d_OO_work_exp_owner      d_OO_race_white_owner
d_Comp_advantage d_hightech d_d1a_provide_service d_d1b_provide_product ///
  ,  rrr cformat(%6.3f) sformat(%6.3f)  nolstretch baseoutcome(1)
mi estimate, rrr cformat(%6.3f) sformat(%6.3f)  nolstretch
```

```
Multiple-imputation estimates                   Imputations       =          5
Survey: Multinomial logistic regression         Number of obs     =      18286
Number of strata  =         6                   Population size   = 408495.43
Number of PSUs    =      3140
                                                Average RVI       =     0.0046
                                                Largest FMI       =     0.0365
                                                Complete DF       =       3134
DF adjustment:   Small sample                   DF:     min       =    1531.34
                                                        avg       =    3031.52
                                                        max       =    3131.99
Model F test:       Equal FMI                   F(  63, 3131.1)   =       6.21
Within VCE type:   Linearized                   Prob > F          =     0.0000
------------------------------------------------------------------------------
  Legal_Form |        RRR   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |  (base outcome)
-------------+----------------------------------------------------------------
2            |
   m_Have_IP |      2.049      0.397    3.699   0.000        1.401       2.996
  m_LnAssets |      1.227      0.034    7.383   0.000        1.162       1.295
m_Home_Based |      0.562      0.073   -4.460   0.000        0.437       0.724
m_OO_D_edu~r |      1.792      0.217    4.812   0.000        1.413       2.273
m_OO_work_~r |      0.997      0.006   -0.467   0.640        0.986       1.009
m_OO_race_~r |      1.244      0.210    1.295   0.196        0.894       1.731
 1.PO_gender |      1.701      0.221    4.096   0.000        1.319       2.193
m_Comp_adv~e |      0.807      0.135   -1.276   0.202        0.581       1.122
  m_hightech |      1.890      0.402    2.994   0.003        1.246       2.867
m_d1a_prov~e |      0.632      0.141   -2.054   0.040        0.408       0.979
m_d1b_prov~t |      0.454      0.071   -5.069   0.000        0.334       0.616
   d_Have_IP |      1.097      0.047    2.145   0.032        1.008       1.194
  d_LnAssets |      1.004      0.004    1.077   0.282        0.997       1.011
d_Home_Based |      0.900      0.087   -1.088   0.277        0.745       1.088
d_OO_D_edu~r |      0.921      0.077   -0.984   0.325        0.782       1.085
d_OO_work_~r |      1.004      0.014    0.317   0.751        0.978       1.031
d_OO_race_~r |      1.081      0.191    0.438   0.661        0.764       1.529
d_Comp_adv~e |      0.964      0.026   -1.363   0.173        0.915       1.016
  d_hightech |      0.928      0.056   -1.249   0.212        0.825       1.044
d_d1a_prov~e |      0.934      0.055   -1.164   0.245        0.833       1.048
d_d1b_prov~t |      0.921      0.031   -2.469   0.014        0.862       0.983
       _cons |      0.170      0.072   -4.196   0.000        0.074       0.389
-------------+----------------------------------------------------------------
3            |
   m_Have_IP |      2.075      0.446    3.393   0.001        1.361       3.163
  m_LnAssets |      1.274      0.038    8.144   0.000        1.202       1.351
m_Home_Based |      0.363      0.052   -7.026   0.000        0.274       0.482
m_OO_D_edu~r |      1.145      0.152    1.018   0.309        0.882       1.485
m_OO_work_~r |      0.995      0.007   -0.693   0.489        0.983       1.008
m_OO_race_~r |      1.239      0.233    1.141   0.254        0.857       1.792
 1.PO_gender |      1.531      0.225    2.890   0.004        1.147       2.043
m_Comp_adv~e |      0.818      0.154   -1.066   0.287        0.565       1.184
  m_hightech |      2.178      0.484    3.500   0.000        1.408       3.369
m_d1a_prov~e |      1.073      0.268    0.282   0.778        0.658       1.750
m_d1b_prov~t |      0.612      0.103   -2.920   0.004        0.440       0.851
   d_Have_IP |      1.085      0.042    2.080   0.038        1.005       1.171
  d_LnAssets |      1.011      0.004    2.501   0.012        1.002       1.020
d_Home_Based |      0.748      0.092   -2.358   0.018        0.588       0.952
d_OO_D_edu~r |      0.868      0.098   -1.247   0.213        0.696       1.084
d_OO_work_~r |      0.991      0.012   -0.727   0.468        0.967       1.015
d_OO_race_~r |      0.872      0.240   -0.496   0.620        0.508       1.497
d_Comp_adv~e |      0.968      0.027   -1.152   0.249        0.916       1.023
  d_hightech |      0.991      0.071   -0.123   0.902        0.862       1.140
d_d1a_prov~e |      1.041      0.055    0.764   0.445        0.939       1.154
```

```
d_d1b_prov~t |      0.943      0.036   -1.549   0.122        0.876       1.016
       _cons |      0.076      0.032   -6.034   0.000        0.033       0.175
-------------+----------------------------------------------------------------
4            |
   m_Have_IP |      2.735      0.645    4.267   0.000        1.723       4.343
  m_LnAssets |      1.153      0.040    4.145   0.000        1.078       1.234
m_Home_Based |      0.293      0.050   -7.214   0.000        0.210       0.409
m_OO_D_edu~r |      0.931      0.150   -0.443   0.658        0.679       1.277
m_OO_work_~r |      0.995      0.008   -0.676   0.499        0.980       1.010
m_OO_race_~r |      0.845      0.172   -0.828   0.408        0.566       1.260
 1.PO_gender |      1.431      0.247    2.073   0.038        1.020       2.009
m_Comp_adv~e |      1.126      0.253    0.529   0.597        0.725       1.748
  m_hightech |      2.014      0.524    2.691   0.007        1.209       3.355
m_d1a_prov~e |      0.628      0.173   -1.692   0.091        0.366       1.077
m_d1b_prov~t |      0.491      0.100   -3.492   0.000        0.329       0.732
   d_Have_IP |      1.134      0.077    1.848   0.065        0.992       1.296
  d_LnAssets |      1.002      0.007    0.233   0.815        0.988       1.016
d_Home_Based |      0.952      0.106   -0.445   0.656        0.765       1.184
d_OO_D_edu~r |      1.116      0.182    0.675   0.500        0.811       1.537
d_OO_work_~r |      1.020      0.016    1.207   0.228        0.988       1.052
d_OO_race_~r |      1.010      0.345    0.030   0.976        0.517       1.973
d_Comp_adv~e |      0.972      0.038   -0.732   0.464        0.901       1.049
  d_hightech |      0.995      0.128   -0.036   0.971        0.774       1.281
d_d1a_prov~e |      1.015      0.053    0.283   0.778        0.916       1.124
d_d1b_prov~t |      0.930      0.043   -1.561   0.119        0.848       1.019
       _cons |      0.251      0.128   -2.720   0.007        0.092       0.680
------------------------------------------------------------------------------
```

6.5.3 Ordered Logit Models for Catagorical Response Variables

Suppose we have a ordinal dependent variable y_i where κ indexes choices (K of them) and i indexes firms (n of them). Ordered logit is designed to estimate what observed covariates (x_i) predict choice of κ from the other K-1 alternatives.

$$y_i^* = \beta x_i + \epsilon_i$$

where the K observed response categories α_κ, κ = 1, 2,... K are generated by applying thresholds λ_κ, κ = 1,2,……. K-1 to y_i^* as follows:

$$y_j = \begin{cases} \alpha_1 , if\ y_i^* \le \lambda_1 \\ \alpha_2 , if\ \lambda_2 < y_i^* \le \lambda_2 \\ \alpha_3 , if\ \lambda_3 < y_i^* \le \lambda_4 \\ \quad \ldots\ldots \\ \alpha_K , if\ \lambda_{K-1} < y_i^* \end{cases}$$

where the thresholds λ_κ do not vary between subjects.
The probability of the kth response category is

$$\Pr(\ y_i = \ \alpha_K\ ) = \Pr(\lambda_{\kappa-1} < y_i^* \le \lambda_\kappa)$$

Examples 6.18 Robust Standard Errors

```
mi   xeq   0:svy:   ologit   d8b_perc_international_sales   i.Have_IP   LnAssets
i.Home_Based   i.OO_D_education_owner   OO_work_exp_owner   OO_race_white_owner
PO_gender  i.Comp_advantage ///
i.hightech  i.d1a_provide_service  i.d1b_provide_product  if  year>2006  ,  or
cformat(%6.3f)  sformat(%6.3f)   nolstretch

Survey: Ordered logistic regression
Number of strata   =         6                 Number of obs      =      1361
Number of PSUs     =       521                 Population size    =  23713.85
                                               Design df          =       515
                                               F(  11,    505)    =      1.57
                                               Prob > F           =    0.1041

------------------------------------------------------------------------------
             |             Linearized
d8b_perc_i~s | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   1.Have_IP |      0.922      0.185   -0.406   0.685        0.622       1.366
    LnAssets |      1.009      0.036    0.255   0.799        0.940       1.083
1.Home_Based |      1.611      0.375    2.048   0.041        1.020       2.546
1.OO_D_edu~r |      1.437      0.335    1.553   0.121        0.908       2.273
OO_work_ex~r |      1.011      0.011    1.009   0.313        0.989       1.034
OO_race_wh~r |      0.855      0.293   -0.456   0.649        0.436       1.677
   PO_gender |      1.968      0.527    2.529   0.012        1.163       3.329
1.Comp_adv~e |      1.198      0.229    0.948   0.344        0.824       1.743
  1.hightech |      0.909      0.252   -0.344   0.731        0.527       1.568
1.d1a_prov~e |      0.801      0.184   -0.964   0.336        0.509       1.259
1.d1b_prov~t |      1.090      0.276    0.339   0.735        0.663       1.792
-------------+----------------------------------------------------------------
       /cut1 |      1.151      0.670    1.717   0.087       -0.166       2.467
       /cut2 |      2.700      0.677    3.990   0.000        1.370       4.029
       /cut3 |      3.303      0.686    4.814   0.000        1.955       4.651
       /cut4 |      3.927      0.748    5.250   0.000        2.457       5.396
------------------------------------------------------------------------------
```

```
mi estimate, esampvaryok:svy: ologit d8b_perc_international_sales i.Have_IP
LnAssets        i.Home_Based      i.OO_D_education_owner      OO_work_exp_owner
OO_race_white_owner PO_gender  i.Comp_advantage ///
i.hightech  i.d1a_provide_service  i.d1b_provide_product  if  year>2006  ,  or
cformat(%6.3f)  sformat(%6.3f)   nolstretch
mi estimate,   or  cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                   Imputations       =         5
Survey: Ordered logistic regression             Number of obs     =      1497

Number of strata  =         6                   Population size   = 26019.671
Number of PSUs    =       537
                                                Average RVI       =    0.0045
                                                Largest FMI       =    0.0120
                                                Complete DF       =       531
DF adjustment:   Small sample                   DF:     min       =    513.36
                                                        avg       =    522.06
                                                        max       =    528.24
Model F test:       Equal FMI                   F(  11,  528.8)   =      1.79
Within VCE type:   Linearized                   Prob > F          =    0.0534

------------------------------------------------------------------------------
d8b_perc_i~s | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   1.Have_IP |      0.897      0.173   -0.567   0.571        0.614       1.308
    LnAssets |      1.008      0.034    0.247   0.805        0.944       1.077
1.Home_Based |      1.636      0.367    2.196   0.029        1.053       2.541
1.OO_D_edu~r |      1.438      0.321    1.626   0.104        0.927       2.229
OO_work_ex~r |      1.012      0.011    1.067   0.286        0.990       1.033
OO_race_wh~r |      0.867      0.282   -0.439   0.661        0.458       1.642
   PO_gender |      1.940      0.504    2.549   0.011        1.164       3.233
1.Comp_adv~e |      1.160      0.205    0.841   0.401        0.820       1.643
  1.hightech |      0.924      0.253   -0.288   0.773        0.539       1.583
1.d1a_prov~e |      0.747      0.167   -1.306   0.192        0.482       1.158
1.d1b_prov~t |      1.021      0.242    0.086   0.932        0.640       1.627
-------------+----------------------------------------------------------------
       /cut1 |      1.032      0.639    1.615   0.107       -0.224       2.288
       /cut2 |      2.568      0.643    3.992   0.000        1.304       3.832
       /cut3 |      3.174      0.650    4.881   0.000        1.897       4.452
       /cut4 |      3.811      0.714    5.339   0.000        2.409       5.213
------------------------------------------------------------------------------
Warning: estimation sample varies across imputations; results may be biased.
         Sample sizes vary between 1497 and 1502.
Note: number of primary clusters varies among imputations.
Note: population size varies among imputations.
```

```
mi  estimate,cmdok  esampvaryok:  gllamm  d8b_perc_international_sales  Have_IP
LnAssets           Home_Based       OO_D_education_owner       OO_work_exp_owner
OO_race_white_owner PO_gender  Comp_advantage ///
hightech  d1a_provide_service  d1b_provide_product  if  year>2006  ,  i(mprid)
pweight(pwt)     link(ologit) fam(binom) eform init robust
 mi estimate, eform
```

```
Multiple-imputation estimates                   Imputations     =          5
fixed effects model                             Number of obs   =       1497
                                                Average RVI     =     0.0080
                                                Largest FMI     =     0.0171
DF adjustment:   Large sample                   DF:     min     =   13903.78
                                                        avg     =  353058.11
Within VCE type:          OIM                           max     = 3202948.02
```

d8b_perc_international_sales	exp(b)	Std. Err.	t	P>t	[95% Conf.	Interval]
d8b_perc_international_sales						
Have_IP	0.91	0.12	-0.76	0.45	0.70	1.17
LnAssets	1.01	0.03	0.33	0.74	0.96	1.06
Home_Based	1.60	0.22	3.36	0.00	1.22	2.11
OO_D_education_owner	1.44	0.20	2.57	0.01	1.09	1.90
OO_work_exp_owner	1.01	0.01	1.70	0.09	1.00	1.02
OO_race_white_owner	0.99	0.19	-0.07	0.94	0.68	1.44
PO_gender	1.80	0.31	3.39	0.00	1.28	2.54
Comp_advantage	1.11	0.15	0.76	0.45	0.85	1.45
hightech	1.02	0.16	0.14	0.89	0.75	1.40
d1a_provide_service	0.74	0.12	-1.85	0.06	0.54	1.02
d1b_provide_product	0.93	0.15	-0.46	0.65	0.67	1.28
_cut11						
_cons	0.99	0.45	2.21	0.03	0.11	1.88
_cut12						
_cons	2.53	0.45	5.59	0.00	1.64	3.42
_cut13						
_cons	3.15	0.45	6.94	0.00	2.26	4.04
_cut14						
_cons	3.79	0.48	7.96	0.00	2.85	4.72

```
Warning: estimation sample varies across imputations; results may be biased.
Sample sizes vary between 1497 and 1502.
```

Examples 6.19 Random Effects (Random-Intercept)

```
mi xeq 0:gllamm d8b_perc_international_sales Have_IP LnAssets Home_Based
OO_D_education_owner OO_work_exp_owner OO_race_white_owner PO_gender
Comp_advantage ///
hightech d1a_provide_service d1b_provide_product if year>2006 , i(mprid)
pweight(pwt) link(ologit) fam(binom) eform robust

number of level 1 units = 1361
number of level 2 units = 521

Condition Number = 235.38122

gllamm model

log likelihood = -7512.1349

Robust standard errors
```

d8b_perc_international_sales	exp(b)	Std. Err.	z	P>z	[95% Conf.	Interval]
d8b_perc_international_sales						
Have_IP	0.65	0.30	-0.92	0.36	0.26	1.62
LnAssets	1.02	0.06	0.31	0.75	0.90	1.15
Home_Based	2.36	1.02	1.98	0.05	1.01	5.51
OO_D_education_owner	1.84	1.05	1.07	0.29	0.60	5.63
OO_work_exp_owner	1.00	0.02	-0.08	0.94	0.96	1.04
OO_race_white_owner	0.56	0.31	-1.04	0.30	0.19	1.67
PO_gender	3.33	2.81	1.43	0.15	0.64	17.41
Comp_advantage	1.16	0.35	0.50	0.62	0.65	2.08
hightech	1.92	0.73	1.72	0.09	0.91	4.04
d1a_provide_service	1.28	0.55	0.58	0.56	0.56	2.96
d1b_provide_product	1.10	0.40	0.25	0.80	0.54	2.24
_cut11						
_cons	2.24	1.08	2.08	0.04	0.13	4.35
_cut12						
_cons	5.47	1.15	4.76	0.00	3.22	7.72
_cut13						
_cons	6.74	1.19	5.64	0.00	4.40	9.08
_cut14						
_cons	7.928	1.2594	6.3	0	5.459504	10.3962

```
Variances and covariances of random effects
------------------------------------------------------------------------------


***level 2 (mprid)

    var(1): 8.5909724 (1.6645685)

mi estimate,cmdok esampvaryok: gllamm d8b_perc_international_sales Have_IP
LnAssets Home_Based OO_D_education_owner OO_work_exp_owner
OO_race_white_owner PO_gender Comp_advantage ///
hightech d1a_provide_service d1b_provide_product if year>2006 , i(mprid)
pweight(pwt) link(ologit) fam(binom) eform robust
 mi estimate, eform
```

```
Multiple-imputation estimates                   Imputations     =          5
gllamm model                                    Number of obs   =       1497
                                                Average RVI     =     0.0074
                                                Largest FMI     =     0.0195
DF adjustment:   Large sample                   DF:     min     =   10753.50
                                                        avg     =  453421.66
Within VCE type:          OIM                           max     = 3180007.07
```

d8b_perc_international_sales	exp(b)	Std. Err.	t	P>t	[95% Conf.	Interval]
d8b_perc_international_sales						
Have_IP	0.99	0.53	-0.02	0.99	0.35	2.82
LnAssets	1.00	0.04	0.04	0.97	0.92	1.09
Home_Based	2.76	0.97	2.87	0.00	1.38	5.50
OO_D_education_owner	1.81	1.05	1.02	0.31	0.58	5.62
OO_work_exp_owner	1.00	0.03	-0.06	0.95	0.94	1.06
OO_race_white_owner	0.45	0.23	-1.58	0.12	0.16	1.22
PO_gender	1.19	1.22	0.17	0.87	0.16	8.88
Comp_advantage	0.92	0.25	-0.31	0.76	0.55	1.55
hightech	1.88	0.62	1.93	0.05	0.99	3.58
d1a_provide_service	1.17	0.40	0.47	0.64	0.60	2.30
d1b_provide_product	1.10	0.41	0.26	0.79	0.53	2.30
_cut11						
_cons	1.00	1.20	0.84	0.40	-1.35	3.35
_cut12						
_cons	4.14	1.25	3.32	0.00	1.70	6.58
_cut13						
_cons	5.44	1.22	4.44	0.00	3.04	7.84
_cut14						
_cons	6.67	1.24	5.38	0.00	4.24	9.10
mpri1						
_cons	2.97	0.31	9.63	0.00	2.37	3.58

```
Warning: estimation sample varies across imputations; results may be biased.
Sample sizes vary between 1497 and 1502.
```

6.5.4 Poisson Models for Count Data

Suppose we have the random variable Y_i which takes integer values (1,2,3….). Y_i is said to have a Poisson distribution with parameter μ and the probability of occurrences of the event y_i over a fixed exposure period of time is

$$\Pr(Y_i = y_i) = \frac{e^{-\mu}\,\mu^{y_i}}{y_i!}$$

where $\mu > 0$ and the mean equal variance; $E(Y_i) = \sigma^2(Y_i) = \mu$. When $\sigma^2(Y_i) > E(Y_i) = \mu$ we have an overdispersion problem.

Examples 6.20 Robust Standard Errors

Robust standard errors can correct for both overdispersion and dependence.

```
egen N_Credit_Cards=rowtotal( f7b_pers_credcard_numused f7b_bus_credcard_numused
f9b_pers_credcard_numused f9b_bus_credcard_numused)

mi  xeq  0:svy:poisson    N_Credit_Cards        Have_IP  LnAssets      Home_Based
OO_D_education_owner    OO_work_exp_owner    OO_race_white_owner  OO_gender_owner
Comp_advantage ///
hightech     d1a_provide_service     d1b_provide_product     c4_numowners_confirm
,cformat(%6.3f)  sformat(%6.3f)   nolstretch

Survey: Poisson regression

Number of strata   =         6                  Number of obs      =     16541
Number of PSUs     =      3092                  Population size    = 368607.96
                                                 Design df          =      3086
                                                 F(  12,   3075)    =     21.72
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
N_Credit_C~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Have_IP |       0.126      0.049     2.595   0.009        0.031       0.222
    LnAssets |       0.078      0.006    12.430   0.000        0.066       0.090
  Home_Based |      -0.006      0.041    -0.156   0.876       -0.088       0.075
OO_D_educa~r |      -0.074      0.041    -1.834   0.067       -0.154       0.005
OO_work_ex~r |      -0.008      0.002    -3.863   0.000       -0.012      -0.004
OO_race_wh~r |      -0.091      0.064    -1.425   0.154       -0.216       0.034
OO_gender_~r |      -0.027      0.050    -0.549   0.583       -0.125       0.070
Comp_advan~e |       0.145      0.035     4.182   0.000        0.077       0.214
    hightech |      -0.102      0.078    -1.312   0.190       -0.254       0.050
d1a_provid~e |       0.065      0.064     1.014   0.311       -0.061       0.191
d1b_provid~t |       0.054      0.040     1.352   0.176       -0.025       0.133
c4_numowne~m |       0.049      0.025     1.958   0.050       -0.000       0.098
       _cons |      -0.602      0.115    -5.243   0.000       -0.827      -0.377
------------------------------------------------------------------------------

mi  xeq  0:  poisson    N_Credit_Cards        Have_IP  LnAssets      Home_Based
OO_D_education_owner    OO_work_exp_owner    OO_race_white_owner  OO_gender_owner
Comp_advantage ///
hightech     d1a_provide_service     d1b_provide_product     c4_numowners_confirm
[pw=wgt_7_long],   cluster(mprid) cformat(%6.3f)  sformat(%6.3f)   nolstretch
```

```
Poisson regression                                Number of obs   =      16541
                                                  Wald chi2(12)   =     260.57
Log pseudolikelihood = -618265.45                 Prob > chi2     =     0.0000

                              (Std. Err. adjusted for 3092 clusters in mprid)
------------------------------------------------------------------------------
             |               Robust
N_Credit_C~s |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Have_IP |       0.126      0.049     2.595   0.009        0.031       0.222
    LnAssets |       0.078      0.006    12.417   0.000        0.066       0.090
  Home_Based |      -0.006      0.041    -0.156   0.876       -0.088       0.075
OO_D_educa~r |      -0.074      0.041    -1.834   0.067       -0.154       0.005
OO_work_ex~r |      -0.008      0.002    -3.864   0.000       -0.012      -0.004
OO_race_wh~r |      -0.091      0.064    -1.424   0.154       -0.216       0.034
OO_gender_~r |      -0.027      0.050    -0.549   0.583       -0.125       0.070
Comp_advan~e |       0.145      0.035     4.177   0.000        0.077       0.214
    hightech |      -0.102      0.078    -1.313   0.189       -0.254       0.050
d1a_provid~e |       0.065      0.064     1.014   0.311       -0.061       0.191
d1b_provid~t |       0.054      0.040     1.352   0.176       -0.024       0.133
c4_numowne~m |       0.049      0.025     1.958   0.050       -0.000       0.098
       _cons |      -0.602      0.115    -5.242   0.000       -0.827      -0.377
------------------------------------------------------------------------------


mi  estimate:svy:poisson    N_Credit_Cards       Have_IP  LnAssets     Home_Based
OO_D_education_owner    OO_work_exp_owner    OO_race_white_owner  OO_gender_owner
Comp_advantage ///
hightech     d1a_provide_service     d1b_provide_product     c4_numowners_confirm
,cformat(%6.3f)  sformat(%6.3f)   nolstretch
mi estimate,cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                     Imputations        =         5
Survey: Poisson regression                        Number of obs      =     18286

Number of strata  =         6                     Population size    = 408495.43
Number of PSUs    =      3140
                                                  Average RVI        =    0.0218
                                                  Largest FMI        =    0.0589
                                                  Complete DF        =      3134
DF adjustment:   Small sample                     DF:     min        =    862.53
                                                          avg        =   2334.08
                                                          max        =   3105.26
Model F test:       Equal FMI                     F(  12, 3013.4)    =     22.00
Within VCE type:   Linearized                     Prob > F           =    0.0000
------------------------------------------------------------------------------
N_Credit_C~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Have_IP |       0.086      0.043     1.986   0.047        0.001       0.172
    LnAssets |       0.076      0.006    12.915   0.000        0.064       0.087
  Home_Based |      -0.001      0.038    -0.026   0.980       -0.076       0.074
OO_D_educa~r |      -0.034      0.038    -0.900   0.368       -0.108       0.040
OO_work_ex~r |      -0.006      0.002    -3.492   0.000       -0.010      -0.003
OO_race_wh~r |      -0.065      0.058    -1.122   0.262       -0.180       0.049
OO_gender_~r |      -0.014      0.046    -0.304   0.761       -0.104       0.076
Comp_advan~e |       0.126      0.032     3.882   0.000        0.062       0.189
    hightech |      -0.075      0.068    -1.101   0.271       -0.209       0.059
d1a_provid~e |       0.089      0.059     1.511   0.131       -0.026       0.204
d1b_provid~t |       0.053      0.037     1.434   0.152       -0.020       0.127
c4_numowne~m |       0.050      0.032     1.559   0.119       -0.013       0.112
       _cons |      -0.533      0.113    -4.742   0.000       -0.754      -0.313
```

Examples 6.21 Population-Averaged Model

```
mi   xeq   0:xtgee   N_Credit_Cards          Have_IP   LnAssets        Home_Based
OO_D_education_owner    OO_work_exp_owner    OO_race_white_owner  OO_gender_owner
Comp_advantage ///
hightech     d1a_provide_service     d1b_provide_product     c4_numowners_confirm
[pweight=wgt_7_long], vce ( robust) family( poisson )   corr( unstructured   )
cformat(%6.3f)  sformat(%6.3f)   nolstretch

GEE population-averaged model                   Number of obs      =     16541
Group and time vars:            mprid year      Number of groups   =      3092
Link:                                  log      Obs per group: min =         1
Family:                            Poisson                     avg =       5.1
Correlation:                   unstructured                    max =         8
                                                Wald chi2(12)      =    222.24
Scale parameter:                           1    Prob > chi2        =    0.0000

                                (Std. Err. adjusted for clustering on mprid)
------------------------------------------------------------------------------
             |               Robust
N_Credit_C~s |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Have_IP |      0.091      0.040    2.269   0.023        0.012       0.170
    LnAssets |      0.055      0.005   10.416   0.000        0.044       0.065
  Home_Based |     -0.083      0.038   -2.191   0.028       -0.157      -0.009
OO_D_educa~r |     -0.077      0.038   -2.018   0.044       -0.153      -0.002
OO_work_ex~r |     -0.008      0.002   -3.882   0.000       -0.011      -0.004
OO_race_wh~r |     -0.037      0.060   -0.611   0.541       -0.155       0.081
OO_gender_~r |     -0.029      0.050   -0.582   0.560       -0.127       0.069
Comp_advan~e |      0.108      0.029    3.741   0.000        0.051       0.165
    hightech |     -0.101      0.063   -1.615   0.106       -0.225       0.022
d1a_provid~e |      0.079      0.045    1.760   0.078       -0.009       0.168
d1b_provid~t |      0.054      0.034    1.604   0.109       -0.012       0.120
c4_numowne~m |      0.069      0.026    2.593   0.010        0.017       0.121
       _cons |     -0.395      0.102   -3.873   0.000       -0.595      -0.195
------------------------------------------------------------------------------

estat   wcorr

Estimated within-mprid correlation matrix R:

      |        c1         c2         c3         c4         c5         c6         c7
------+---------------------------------------------------------------------------
   r1 |         1
   r2 |  .4151724          1
   r3 |  .3106036   .4826708          1
   r4 |  .2725235   .3905605   .5067727          1
   r5 |  .1801521   .3118357   .4431083   .4978497          1
   r6 |  .1698682   .3690488   .4176272   .5071768   .4938269          1
   r7 |  .1390472   .2934081   .2810183   .3930939   .3828957   .4707738          1
   r8 |  .1268394   .1841205    .270074   .3368679   .2790039   .3964584   .3912923


mi  estimate:xtgee  Have_IP  LnAssets      i.Home_Based    i.OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner PO_gender  i.Comp_advantage ///
i.hightech i.d1a_provide_service i.d1b_provide_product [pweight=wgt_7_long], vce
( robust) family( poisson )  corr( unstructured  ) cformat(%6.3f)  sformat(%6.3f)
nolstretch

mi estimate,      cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                   Imputations        =         5
```

```
GEE population-averaged model                   Number of obs      =     18286

Group and time vars:            mprid year      Number of groups   =      3140
Link:                                  log      Obs per group: min =         1
Family:                            Poisson                    avg =       5.6
Correlation:                   unstructured                    max =         8
Scale parameter:                          1
                                                Average RVI        =    0.0184
                                                Largest FMI        =    0.0876
DF adjustment:                 Large sample     DF:     min        =    561.59
                                                        avg        =  1.04e+06
                                                        max        =  8.34e+06
Model F test:                     Equal FMI     F(  10,83787.7)    =     22.62
Within VCE type:                     Robust     Prob > F           =    0.0000

                                (Within VCE adjusted for clustering on mprid)
------------------------------------------------------------------------------
     Have_IP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |       0.020      0.007     2.982   0.003        0.007       0.033
1.Home_Based |       0.014      0.061     0.226   0.821       -0.105       0.133
1.OO_D_edu~r |       0.299      0.063     4.753   0.000        0.176       0.422
OO_work_ex~r |       0.002      0.003     0.557   0.577       -0.005       0.009
OO_race_wh~r |      -0.067      0.089    -0.750   0.453       -0.241       0.108
   PO_gender |       0.130      0.082     1.586   0.113       -0.031       0.290
1.Comp_adv~e |       0.355      0.039     9.049   0.000        0.278       0.432
  1.hightech |       0.331      0.074     4.489   0.000        0.187       0.476
1.d1a_prov~e |      -0.105      0.055    -1.922   0.055       -0.212       0.002
1.d1b_prov~t |       0.323      0.051     6.379   0.000        0.224       0.422
       _cons |      -2.451      0.145   -16.938   0.000       -2.734      -2.167
------------------------------------------------------------------------------
```

```
mi estimate:xtpoisson Have_IP LnAssets    i.Home_Based   i.OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner PO_gender  i.Comp_advantage ///
i.hightech i.d1a_provide_service i.d1b_provide_product [pweight=wgt_7_long], vce
( robust) pa corr( unstructured  )cformat(%6.3f)  sformat(%6.3f)   nolstretch
mi estimate,      cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                     Imputations        =         5
Population-averaged Poisson regression            Number of obs      =     18286

Group and time vars:            mprid year        Number of groups   =      3140
Link:                                 log         Obs per group: min =         1
Family:                           Poisson                        avg =       5.6
Correlation:                  unstructured                       max =         8
Scale parameter:                         1
                                                  Average RVI        =    0.0184
                                                  Largest FMI        =    0.0876
DF adjustment:                Large sample        DF:     min        =    561.59
                                                          avg        =  1.04e+06
                                                          max        =  8.34e+06
Model F test:                    Equal FMI        F(  10,83787.7)    =     22.62
Within VCE type:                    Robust        Prob > F           =    0.0000

                                (Within VCE adjusted for clustering on mprid)
------------------------------------------------------------------------------
     Have_IP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |       0.020      0.007     2.982   0.003        0.007       0.033
1.Home_Based |       0.014      0.061     0.226   0.821       -0.105       0.133
1.OO_D_edu~r |       0.299      0.063     4.753   0.000        0.176       0.422
OO_work_ex~r |       0.002      0.003     0.557   0.577       -0.005       0.009
OO_race_wh~r |      -0.067      0.089    -0.750   0.453       -0.241       0.108
   PO_gender |       0.130      0.082     1.586   0.113       -0.031       0.290
1.Comp_adv~e |       0.355      0.039     9.049   0.000        0.278       0.432
  1.hightech |       0.331      0.074     4.489   0.000        0.187       0.476
1.d1a_prov~e |      -0.105      0.055    -1.922   0.055       -0.212       0.002
1.d1b_prov~t |       0.323      0.051     6.379   0.000        0.224       0.422
       _cons |      -2.451      0.145   -16.938   0.000       -2.734      -2.167
------------------------------------------------------------------------------
```

Examples 6.22 Random Effects (Random-Intercept)

```
mi   xeq   0:gllamm   N_Credit_Cards          Have_IP   LnAssets        Home_Based
OO_D_education_owner    OO_work_exp_owner    OO_race_white_owner  OO_gender_owner
Comp_advantage ///
hightech d1a_provide_service d1b_provide_product c4_numowners_confirm, i(mprid)
pweight(pwt) family(poisson) link(log)  robust  adapt

number of level 1 units = 16541
number of level 2 units = 3092

Condition Number = 85.934339

gllamm model

log likelihood = -172560.75

Robust standard errors
                              Std.                          [95%
 N_Credit_Cards        Coef.    Err.      z         P>z      Conf.    Interval]

 Have_IP                0.086    0.037     2.290    0.022     0.012      0.159
 LnAssets               0.059    0.006    10.100    0.000     0.048      0.071
 Home_Based            -0.102    0.039    -2.640    0.008    -0.177     -0.026
 OO_D_education_owner  -0.089    0.037    -2.380    0.017    -0.162     -0.016
 OO_work_exp_owner     -0.006    0.002    -3.180    0.001    -0.010     -0.002
 OO_race_white_owner    0.032    0.063     0.500    0.619    -0.093      0.156
 OO_gender_owner       -0.043    0.053    -0.810    0.416    -0.147      0.061
 Comp_advantage         0.141    0.027     5.180    0.000     0.088      0.194
 hightech              -0.085    0.056    -1.500    0.134    -0.195      0.026
 d1a_provide_service    0.066    0.045     1.450    0.146    -0.023      0.155
 d1b_provide_product    0.056    0.032     1.760    0.079    -0.006      0.119
 c4_numowners_confirm   0.072    0.029     2.530    0.011     0.016      0.128
 _cons                 -0.829    0.110    -7.520    0.000    -1.045     -0.613


Variances and covariances of random effects
------------------------------------------------------------------------------


***level 2 (mprid)

    var(1): .69175126 (.03314396)
------------------------------------------------------------------------------
```

```
mi  estimate,cmdok  esampvaryok:gllamm  N_Credit_Cards         Have_IP  LnAssets
Home_Based     OO_D_education_owner     OO_work_exp_owner     OO_race_white_owner
OO_gender_owner  Comp_advantage ///
hightech d1a_provide_service d1b_provide_product c4_numowners_confirm, i(mprid)
pweight(pwt) family(poisson) link(log)  robust  adapt
```

```
Multiple-imputation estimates                   Imputations     =          5
gllamm model                                    Number of obs   =      18286
                                                Average RVI     =     0.0482
                                                Largest FMI     =     0.1678
DF adjustment:   Large sample                   DF:     min     =     161.22
                                                        avg     =  131871.32
Within VCE type:          OIM                           max     = 1674644.73
```

N_Credit_Cards	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
N_Credit_Cards						
Have_IP	0.054	0.032	1.680	0.094	-0.009	0.118
LnAssets	0.055	0.005	10.550	0.000	0.045	0.066
Home_Based	-0.096	0.037	-2.590	0.010	-0.169	-0.023
OO_D_education_owner	-0.045	0.039	-1.140	0.255	-0.121	0.032
OO_work_exp_owner	-0.005	0.002	-2.430	0.015	-0.009	-0.001
OO_race_white_owner	0.026	0.065	0.410	0.682	-0.100	0.153
OO_gender_owner	-0.019	0.055	-0.350	0.730	-0.128	0.089
Comp_advantage	0.103	0.025	4.150	0.000	0.054	0.152
hightech	-0.089	0.054	-1.660	0.096	-0.194	0.016
d1a_provide_service	0.082	0.040	2.070	0.039	0.004	0.161
d1b_provide_product	0.039	0.031	1.260	0.206	-0.022	0.100
c4_numowners_confirm	0.065	0.027	2.420	0.016	0.012	0.117
_cons	-0.700	0.106	-6.620	0.000	-0.907	-0.493
mpri1						
_cons	0.837	0.021	40.640	0.000	0.796	0.877

Examples 6.23 Hybrid Model

```
global  vars     "     Have_IP  LnAssets       Home_Based     OO_D_education_owner
OO_work_exp_owner   OO_race_white_owner OO_gender_owner   Comp_advantage hightech
d1a_provide_service d1b_provide_product c4_numowners_confirm"

 foreach var in $vars{

egen m_`var'=mean(`var'), by(_mi_m mprid)
gen d_`var'=`var'-m_`var'
}



mi xeq 0:svy:poisson  N_Credit_Cards m_Have_IP       m_LnAssets    m_Home_Based
       m_OO_D_education_owner     m_OO_work_exp_owner
       m_OO_race_white_owner      m_OO_gender_owner   m_Comp_advantage    ///
m_hightech    m_d1a_provide_service      m_d1b_provide_product
       m_c4_numowners_confirm  d_Have_IP        d_LnAssets          d_Home_Based
        d_OO_D_education_owner     d_OO_work_exp_owner
d_OO_race_white_owner ///
        d_OO_gender_owner          d_Comp_advantage    d_hightech
d_d1a_provide_service       d_d1b_provide_product      d_c4_numowners_confirm,
cformat(%6.3f)  sformat(%6.3f)   nolstretch

Survey: Poisson regression
Number of strata   =         6                  Number of obs      =     16541
Number of PSUs     =      3092                  Population size    = 368607.96
                                                 Design df          =      3086
                                                 F(  24,   3063)    =     13.09
                                                 Prob > F           =    0.0000
------------------------------------------------------------------------------
             |             Linearized
N_Credit_C~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   m_Have_IP |      0.154      0.075    2.053   0.040        0.007       0.300
  m_LnAssets |      0.098      0.009   11.362   0.000        0.081       0.115
m_Home_Based |      0.040      0.046    0.882   0.378       -0.049       0.130
m_OO_D_edu~r |     -0.074      0.044   -1.702   0.089       -0.160       0.011
m_OO_work_~r |     -0.008      0.002   -4.003   0.000       -0.012      -0.004
m_OO_race_~r |     -0.115      0.064   -1.795   0.073       -0.240       0.011
m_OO_gende~r |     -0.038      0.052   -0.734   0.463       -0.141       0.064
m_Comp_adv~e |      0.164      0.063    2.615   0.009        0.041       0.288
  m_hightech |     -0.094      0.087   -1.086   0.277       -0.265       0.076
m_d1a_prov~e |      0.075      0.091    0.819   0.413       -0.104       0.253
m_d1b_prov~t |      0.037      0.056    0.656   0.512       -0.073       0.147
m_c4_numow~m |      0.025      0.031    0.787   0.431       -0.037       0.086
   d_Have_IP |      0.097      0.047    2.046   0.041        0.004       0.190
  d_LnAssets |      0.043      0.007    6.622   0.000        0.031       0.056
d_Home_Based |     -0.188      0.065   -2.883   0.004       -0.316      -0.060
d_OO_D_edu~r |     -0.169      0.078   -2.156   0.031       -0.322      -0.015
d_OO_work_~r |     -0.003      0.008   -0.425   0.671       -0.019       0.012
d_OO_race_~r |      0.262      0.260    1.009   0.313       -0.247       0.771
d_OO_gende~r |     -0.192      0.155   -1.240   0.215       -0.497       0.112
d_Comp_adv~e |      0.125      0.032    3.903   0.000        0.062       0.188
  d_hightech |     -0.115      0.122   -0.939   0.348       -0.355       0.125
d_d1a_prov~e |      0.037      0.054    0.695   0.487       -0.068       0.143
d_d1b_prov~t |      0.073      0.044    1.670   0.095       -0.013       0.159
d_c4_numow~m |      0.078      0.042    1.839   0.066       -0.005       0.161
       _cons |     -0.777      0.138   -5.610   0.000       -1.048      -0.505
------------------------------------------------------------------------------
```

```
mi estimate:svy:poisson  N_Credit_Cards m_Have_IP    m_LnAssets   m_Home_Based
       m_OO_D_education_owner     m_OO_work_exp_owner
       m_OO_race_white_owner      m_OO_gender_owner   m_Comp_advantage    ///
m_hightech   m_d1a_provide_service      m_d1b_provide_product
       m_c4_numowners_confirm  d_Have_IP        d_LnAssets          d_Home_Based
        d_OO_D_education_owner     d_OO_work_exp_owner
d_OO_race_white_owner ///
       d_OO_gender_owner          d_Comp_advantage    d_hightech
d_d1a_provide_service      d_d1b_provide_product      d_c4_numowners_confirm,
cformat(%6.3f)  sformat(%6.3f)   nolstretch
mi estimate,cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                   Imputations        =         5
Survey: Poisson regression                      Number of obs      =     18286

Number of strata  =         6                   Population size    = 408495.43
Number of PSUs    =      3140
                                                Average RVI        =    0.0402
                                                Largest FMI        =    0.1885
                                                Complete DF        =      3134
DF adjustment:   Small sample                   DF:     min        =    123.04
                                                        avg        =   2003.13
                                                        max        =   3118.16
Model F test:       Equal FMI                   F(  24, 2965.7)    =     12.66
Within VCE type:   Linearized                   Prob > F           =    0.0000

------------------------------------------------------------------------------
N_Credit_C~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   m_Have_IP |       0.116      0.068     1.707   0.088       -0.017       0.249
  m_LnAssets |       0.097      0.008    11.686   0.000        0.081       0.113
m_Home_Based |       0.048      0.043     1.135   0.257       -0.035       0.132
m_OO_D_edu~r |      -0.037      0.041    -0.901   0.368       -0.116       0.043
m_OO_work_~r |      -0.007      0.002    -3.740   0.000       -0.011      -0.003
m_OO_race_~r |      -0.090      0.059    -1.526   0.127       -0.207       0.026
m_OO_gende~r |      -0.029      0.048    -0.605   0.545       -0.123       0.065
m_Comp_adv~e |       0.163      0.058     2.811   0.005        0.049       0.277
  m_hightech |      -0.062      0.076    -0.815   0.415       -0.211       0.087
m_d1a_prov~e |       0.097      0.082     1.179   0.238       -0.064       0.259
m_d1b_prov~t |       0.041      0.051     0.789   0.430       -0.060       0.141
m_c4_numow~m |       0.030      0.038     0.786   0.432       -0.045       0.105
   d_Have_IP |       0.038      0.039     0.979   0.328       -0.038       0.114
  d_LnAssets |       0.041      0.006     6.985   0.000        0.029       0.053
d_Home_Based |      -0.159      0.062    -2.558   0.011       -0.281      -0.037
d_OO_D_edu~r |      -0.121      0.083    -1.456   0.146       -0.283       0.042
d_OO_work_~r |       0.002      0.007     0.277   0.782       -0.012       0.016
d_OO_race_~r |       0.138      0.207     0.670   0.503       -0.267       0.544
d_OO_gende~r |      -0.130      0.153    -0.853   0.394       -0.431       0.170
d_Comp_adv~e |       0.086      0.029     2.973   0.003        0.029       0.143
  d_hightech |      -0.155      0.117    -1.329   0.184       -0.384       0.074
d_d1a_prov~e |       0.076      0.048     1.587   0.113       -0.018       0.170
d_d1b_prov~t |       0.048      0.040     1.189   0.235       -0.031       0.126
d_c4_numow~m |       0.064      0.032     2.038   0.042        0.002       0.126
       _cons |      -0.734      0.135    -5.424   0.000       -0.999      -0.468
------------------------------------------------------------------------------
```

6.5.5 Negative Binomial Models for Count Data

While the Poisson models for count data do a good job of correcting for overdispersion and dependence among observations, negative binomial (NB) models can do better with overdispersion and dependence. Negative binomial regression models the counts of an event when the event has extra-Poisson variation (overdispersion).

Examples 6.24 Robust Standard Errors

```
mi  xeq  0:svy:nbreg    N_Credit_Cards        Have_IP  LnAssets      Home_Based
OO_D_education_owner    OO_work_exp_owner    OO_race_white_owner  OO_gender_owner
Comp_advantage ///
hightech     d1a_provide_service     d1b_provide_product     c4_numowners_confirm
,cformat(%6.3f)  sformat(%6.3f)   nolstretch

Survey: Negative binomial regression

Number of strata   =         6                 Number of obs      =     16541
Number of PSUs     =      3092                 Population size    = 368607.96
                                                Design df          =      3086
                                                F(  12,   3075)    =     22.76
                                                Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
N_Credit_C~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Have_IP |      0.126      0.048    2.654   0.008        0.033       0.219
    LnAssets |      0.079      0.006   12.747   0.000        0.067       0.091
  Home_Based |     -0.010      0.040   -0.237   0.813       -0.088       0.069
OO_D_educa~r |     -0.069      0.040   -1.756   0.079       -0.147       0.008
OO_work_ex~r |     -0.008      0.002   -3.936   0.000       -0.012      -0.004
OO_race_wh~r |     -0.091      0.064   -1.421   0.155       -0.216       0.035
OO_gender_~r |     -0.010      0.050   -0.203   0.839       -0.107       0.087
Comp_advan~e |      0.152      0.034    4.488   0.000        0.086       0.219
    hightech |     -0.099      0.076   -1.302   0.193       -0.248       0.050
d1a_provid~e |      0.073      0.063    1.149   0.251       -0.051       0.197
d1b_provid~t |      0.056      0.040    1.427   0.154       -0.021       0.134
c4_numowne~m |      0.069      0.030    2.319   0.020        0.011       0.127
       _cons |     -0.660      0.117   -5.628   0.000       -0.890      -0.430
-------------+----------------------------------------------------------------
    /lnalpha |     -0.106      0.046                       -0.196      -0.015
-------------+----------------------------------------------------------------
       alpha |      0.900      0.042                        0.822       0.985
------------------------------------------------------------------------------
```

```
mi  xeq  0:  nbreg    N_Credit_Cards         Have_IP  LnAssets       Home_Based
OO_D_education_owner    OO_work_exp_owner    OO_race_white_owner  OO_gender_owner
Comp_advantage ///
hightech     d1a_provide_service     d1b_provide_product     c4_numowners_confirm
[pw=wgt_7_long],   cluster(mprid) cformat(%6.3f)  sformat(%6.3f)   nolstretch

Negative binomial regression                      Number of obs   =      16541
Dispersion           = mean                       Wald chi2(12)   =     273.25
Log pseudolikelihood = -562393.97                 Prob > chi2     =     0.0000

                             (Std. Err. adjusted for 3092 clusters in mprid)
------------------------------------------------------------------------------
             |               Robust
N_Credit_C~s |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Have_IP |      0.126      0.048    2.654   0.008        0.033       0.219
    LnAssets |      0.079      0.006   12.734   0.000        0.067       0.091
  Home_Based |     -0.010      0.040   -0.237   0.813       -0.088       0.069
OO_D_educa~r |     -0.069      0.040   -1.756   0.079       -0.147       0.008
OO_work_ex~r |     -0.008      0.002   -3.937   0.000       -0.012      -0.004
OO_race_wh~r |     -0.091      0.064   -1.421   0.155       -0.216       0.035
OO_gender_~r |     -0.010      0.050   -0.203   0.839       -0.107       0.087
Comp_advan~e |      0.152      0.034    4.484   0.000        0.086       0.219
    hightech |     -0.099      0.076   -1.302   0.193       -0.248       0.050
d1a_provid~e |      0.073      0.063    1.149   0.251       -0.051       0.197
d1b_provid~t |      0.056      0.040    1.427   0.153       -0.021       0.134
c4_numowne~m |      0.069      0.030    2.319   0.020        0.011       0.127
       _cons |     -0.660      0.117   -5.627   0.000       -0.890      -0.430
-------------+----------------------------------------------------------------
    /lnalpha |     -0.106      0.046                       -0.196      -0.015
-------------+----------------------------------------------------------------
       alpha |      0.900      0.042                        0.822       0.985
------------------------------------------------------------------------------
```

```
mi  estimate:svy:nbreg    N_Credit_Cards        Have_IP  LnAssets      Home_Based
OO_D_education_owner    OO_work_exp_owner    OO_race_white_owner  OO_gender_owner
Comp_advantage ///
hightech     d1a_provide_service     d1b_provide_product     c4_numowners_confirm
,cformat(%6.3f)  sformat(%6.3f)   nolstretch
mi estimate,cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                   Imputations        =         5
Survey: Negative binomial regression            Number of obs      =     18286

Number of strata  =         6                   Population size    = 408495.43
Number of PSUs    =      3140
                                                Average RVI        =    0.0194
                                                Largest FMI        =    0.0547
                                                Complete DF        =      3134
DF adjustment:   Small sample                   DF:     min        =    954.28
                                                        avg        =   2427.27
                                                        max        =   3085.97
Model F test:       Equal FMI                   F(  12, 3027.1)    =     23.38
Within VCE type:   Linearized                   Prob > F           =    0.0000

------------------------------------------------------------------------------
N_Credit_C~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Have_IP |      0.085      0.042    2.001   0.046        0.002       0.168
    LnAssets |      0.077      0.006   13.328   0.000        0.066       0.088
  Home_Based |     -0.004      0.037   -0.095   0.924       -0.077       0.069
OO_D_educa~r |     -0.028      0.037   -0.757   0.449       -0.100       0.044
OO_work_ex~r |     -0.007      0.002   -3.511   0.000       -0.010      -0.003
OO_race_wh~r |     -0.069      0.059   -1.171   0.242       -0.184       0.046
OO_gender_~r |      0.001      0.046    0.032   0.975       -0.089       0.092
Comp_advan~e |      0.129      0.032    4.075   0.000        0.067       0.191
    hightech |     -0.082      0.067   -1.233   0.218       -0.214       0.049
d1a_provid~e |      0.093      0.058    1.601   0.110       -0.021       0.207
d1b_provid~t |      0.050      0.037    1.376   0.169       -0.021       0.122
c4_numowne~m |      0.073      0.034    2.159   0.031        0.007       0.140
       _cons |     -0.591      0.113   -5.243   0.000       -0.813      -0.370
-------------+----------------------------------------------------------------
    /lnalpha |     -0.336      0.048                       -0.430      -0.242
-------------+----------------------------------------------------------------
       alpha |      0.715      0.034                        0.651       0.785
------------------------------------------------------------------------------
```

Alpha is the overdispersion parameter. If Alpha =0, then there is no overdispersion. Thus, we have evidence of overdispersion in our examples.

Examples 6.25 Population-Averaged Model

```
mi   xeq   0:xtgee   N_Credit_Cards          Have_IP   LnAssets        Home_Based
OO_D_education_owner    OO_work_exp_owner    OO_race_white_owner  OO_gender_owner
Comp_advantage ///
hightech     d1a_provide_service     d1b_provide_product     c4_numowners_confirm
[pweight=wgt_7_long], vce ( robust) family(  nbinomial )  corr( unstructured  )
cformat(%6.3f)  sformat(%6.3f)   nolstretch

GEE population-averaged model                   Number of obs      =     16541
Group and time vars:            mprid year      Number of groups   =      3092
Link:                                 log       Obs per group: min =         1
Family:             negative binomial(k=1)                     avg =       5.1
Correlation:                  unstructured                     max =         8
                                                Wald chi2(12)      =    253.99
Scale parameter:                         1      Prob > chi2        =    0.0000

                                 (Std. Err. adjusted for clustering on mprid)
------------------------------------------------------------------------------
             |             Semirobust
N_Credit_C~s |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Have_IP |      0.093      0.039    2.410   0.016        0.017       0.168
    LnAssets |      0.055      0.005   10.476   0.000        0.044       0.065
  Home_Based |     -0.087      0.037   -2.346   0.019       -0.161      -0.014
OO_D_educa~r |     -0.071      0.038   -1.886   0.059       -0.145       0.003
OO_work_ex~r |     -0.008      0.002   -4.112   0.000       -0.012      -0.004
OO_race_wh~r |     -0.034      0.060   -0.556   0.578       -0.152       0.085
OO_gender_~r |     -0.013      0.049   -0.262   0.793       -0.109       0.083
Comp_advan~e |      0.112      0.028    3.936   0.000        0.056       0.167
    hightech |     -0.097      0.060   -1.604   0.109       -0.215       0.022
d1a_provid~e |      0.078      0.045    1.730   0.084       -0.010       0.167
d1b_provid~t |      0.056      0.033    1.718   0.086       -0.008       0.121
c4_numowne~m |      0.096      0.024    4.018   0.000        0.049       0.143
       _cons |     -0.451      0.100   -4.525   0.000       -0.646      -0.255
------------------------------------------------------------------------------
```

```
mi   estimate:xtgee   N_Credit_Cards   Have_IP   LnAssets          i.Home_Based
i.OO_D_education_owner     OO_work_exp_owner     OO_race_white_owner   PO_gender
i.Comp_advantage ///
i.hightech i.d1a_provide_service i.d1b_provide_product [pweight=wgt_7_long], vce
( robust) family( nbinomial )  corr( unstructured  ) cformat(%6.3f)  sformat(%6.3f)
nolstretch
mi estimate,      cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                   Imputations       =         5
GEE population-averaged model                   Number of obs     =     18286

Group and time vars:            mprid year      Number of groups  =      3140
Link:                                 log       Obs per group: min =         1
Family:             negative binomial(k=1)                     avg =       5.6
Correlation:                  unstructured                     max =         8
Scale parameter:                         1
                                                Average RVI       =    0.0491
                                                Largest FMI       =    0.1465
DF adjustment:                Large sample      DF:     min       =    208.97
                                                        avg       =  16989.61
                                                        max       =  64086.76
Model F test:                    Equal FMI      F(  11,14341.7)   =     20.35
Within VCE type:                Semirobust      Prob > F          =    0.0000

                              (Within VCE adjusted for clustering on mprid)
------------------------------------------------------------------------------
N_Credit_C~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Have_IP |      0.057      0.033    1.701   0.089       -0.009       0.122
    LnAssets |      0.052      0.005   10.949   0.000        0.042       0.061
1.Home_Based |     -0.104      0.034   -3.009   0.003       -0.171      -0.036
1.OO_D_edu~r |     -0.034      0.036   -0.946   0.344       -0.106       0.037
OO_work_ex~r |     -0.007      0.002   -4.140   0.000       -0.011      -0.004
OO_race_wh~r |     -0.010      0.057   -0.183   0.855       -0.123       0.102
   PO_gender |      0.039      0.044    0.889   0.374       -0.047       0.124
1.Comp_adv~e |      0.089      0.026    3.357   0.001        0.037       0.141
  1.hightech |     -0.071      0.055   -1.299   0.194       -0.179       0.036
1.d1a_prov~e |      0.100      0.041    2.422   0.016        0.019       0.180
1.d1b_prov~t |      0.053      0.032    1.678   0.094       -0.009       0.115
       _cons |     -0.243      0.093   -2.622   0.009       -0.424      -0.061
------------------------------------------------------------------------------
```

```
mi   estimate:xtnbreg   N_Credit_Cards   Have_IP   LnAssets         i.Home_Based
i.OO_D_education_owner     OO_work_exp_owner     OO_race_white_owner   PO_gender
i.Comp_advantage ///
i.hightech i.d1a_provide_service i.d1b_provide_product [pweight=wgt_7_long], vce
( robust) pa corr( unstructured  )cformat(%6.3f)  sformat(%6.3f)   nolstretch
mi estimate,      cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates
Population-averaged negative binomial regression

                                                Imputations        =         5
                                                Number of obs      =     18286

Group and time vars:            mprid year      Number of groups   =      3140
Link:                                 log       Obs per group: min =         1
Family:             negative binomial(k=1)                     avg =       5.6
Correlation:                  unstructured                     max =         8
Scale parameter:                         1
                                                Average RVI        =    0.0491
                                                Largest FMI        =    0.1465
DF adjustment:                Large sample      DF:     min        =    208.97
                                                        avg        =  16989.61
                                                        max        =  64086.76
Model F test:                    Equal FMI      F(  11,14341.7)    =     20.35
Within VCE type:                 Semirobust     Prob > F           =    0.0000

                                 (Within VCE adjusted for clustering on mprid)
------------------------------------------------------------------------------
N_Credit_C~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Have_IP |      0.057      0.033    1.701   0.089       -0.009       0.122
    LnAssets |      0.052      0.005   10.949   0.000        0.042       0.061
1.Home_Based |     -0.104      0.034   -3.009   0.003       -0.171      -0.036
1.OO_D_edu~r |     -0.034      0.036   -0.946   0.344       -0.106       0.037
OO_work_ex~r |     -0.007      0.002   -4.140   0.000       -0.011      -0.004
OO_race_wh~r |     -0.010      0.057   -0.183   0.855       -0.123       0.102
   PO_gender |      0.039      0.044    0.889   0.374       -0.047       0.124
1.Comp_adv~e |      0.089      0.026    3.357   0.001        0.037       0.141
  1.hightech |     -0.071      0.055   -1.299   0.194       -0.179       0.036
1.d1a_prov~e |      0.100      0.041    2.422   0.016        0.019       0.180
1.d1b_prov~t |      0.053      0.032    1.678   0.094       -0.009       0.115
       _cons |     -0.243      0.093   -2.622   0.009       -0.424      -0.061
------------------------------------------------------------------------------
```

Examples 6.26 Hybrid Model

```
mi xeq 0:xtgee N_Credit_Cards m_Have_IP        m_LnAssets    m_Home_Based
       m_OO_D_education_owner     m_OO_work_exp_owner
       m_OO_race_white_owner      m_OO_gender_owner   m_Comp_advantage    ///
m_hightech    m_d1a_provide_service      m_d1b_provide_product
       m_c4_numowners_confirm  d_Have_IP        d_LnAssets          d_Home_Based
        d_OO_D_education_owner     d_OO_work_exp_owner
d_OO_race_white_owner ///
        d_OO_gender_owner          d_Comp_advantage    d_hightech
d_d1a_provide_service       d_d1b_provide_product      d_c4_numowners_confirm
[pweight=wgt_7_long], vce ( robust) family(  nbinomial )  corr( unstructured  )
cformat(%6.3f)  sformat(%6.3f)   nolstretch

GEE population-averaged model                   Number of obs      =     16541
Group and time vars:            mprid year      Number of groups   =      3092
Link:                                 log       Obs per group: min =         1
Family:             negative binomial(k=1)                       avg =       5.1
Correlation:                  unstructured                       max =         8
                                                Wald chi2(24)      =    321.22
Scale parameter:                         1      Prob > chi2        =    0.0000

                                (Std. Err. adjusted for clustering on mprid)
------------------------------------------------------------------------------
             |             Semirobust
N_Credit_C~s |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   m_Have_IP |      0.113      0.068    1.658   0.097       -0.021       0.246
  m_LnAssets |      0.094      0.008   11.240   0.000        0.078       0.110
m_Home_Based |      0.014      0.044    0.316   0.752       -0.073       0.101
m_OO_D_edu~r |     -0.054      0.041   -1.297   0.195       -0.135       0.027
m_OO_work_~r |     -0.009      0.002   -4.549   0.000       -0.013      -0.005
m_OO_race_~r |     -0.078      0.061   -1.276   0.202       -0.197       0.042
m_OO_gende~r |     -0.025      0.050   -0.504   0.614       -0.122       0.072
m_Comp_adv~e |      0.182      0.060    3.041   0.002        0.065       0.299
  m_hightech |     -0.096      0.072   -1.335   0.182       -0.237       0.045
m_d1a_prov~e |      0.106      0.080    1.318   0.188       -0.051       0.262
m_d1b_prov~t |      0.035      0.054    0.659   0.510       -0.070       0.141
m_c4_numow~m |      0.036      0.031    1.184   0.236       -0.024       0.096
   d_Have_IP |      0.087      0.045    1.945   0.052       -0.001       0.175
  d_LnAssets |      0.040      0.006    6.214   0.000        0.028       0.053
d_Home_Based |     -0.231      0.066   -3.531   0.000       -0.360      -0.103
d_OO_D_edu~r |     -0.147      0.079   -1.861   0.063       -0.302       0.008
d_OO_work_~r |     -0.004      0.008   -0.480   0.631       -0.021       0.012
d_OO_race_~r |      0.133      0.274    0.486   0.627       -0.404       0.671
d_OO_gende~r |     -0.138      0.168   -0.825   0.409       -0.467       0.191
d_Comp_adv~e |      0.099      0.032    3.129   0.002        0.037       0.161
  d_hightech |     -0.103      0.114   -0.904   0.366       -0.326       0.120
d_d1a_prov~e |      0.064      0.054    1.194   0.232       -0.041       0.169
d_d1b_prov~t |      0.055      0.042    1.325   0.185       -0.026       0.137
d_c4_numow~m |      0.125      0.044    2.873   0.004        0.040       0.211
       _cons |     -0.816      0.131   -6.220   0.000       -1.073      -0.559
------------------------------------------------------------------------------
```

```
mi estimate:xtgee N_Credit_Cards m_Have_IP     m_LnAssets    m_Home_Based
       m_OO_D_education_owner     m_OO_work_exp_owner
       m_OO_race_white_owner      m_OO_gender_owner   m_Comp_advantage    ///
m_hightech    m_d1a_provide_service      m_d1b_provide_product
       m_c4_numowners_confirm  d_Have_IP        d_LnAssets          d_Home_Based
        d_OO_D_education_owner     d_OO_work_exp_owner
d_OO_race_white_owner ///
       d_OO_gender_owner          d_Comp_advantage    d_hightech
d_d1a_provide_service       d_d1b_provide_product      d_c4_numowners_confirm
[pweight=wgt_7_long], vce ( robust) family( nbinomial )  corr( unstructured  )
cformat(%6.3f)  sformat(%6.3f)   nolstretch
mi estimate,      cformat(%6.3f)  sformat(%6.3f)   nolstretch


Multiple-imputation estimates                   Imputations        =         5
GEE population-averaged model                   Number of obs      =     18286

Group and time vars:            mprid year      Number of groups   =      3140
Link:                                 log       Obs per group: min =         1
Family:             negative binomial(k=1)                     avg =       5.6
Correlation:                  unstructured                     max =         8
Scale parameter:                         1
                                                Average RVI        =    0.0504
                                                Largest FMI        =    0.1690
DF adjustment:                Large sample      DF:     min        =    159.07
                                                        avg        = 153802.02
                                                        max        =  1.98e+06
Model F test:                    Equal FMI      F(  24,35582.4)    =     13.38
Within VCE type:                 Semirobust     Prob > F           =    0.0000

                                 (Within VCE adjusted for clustering on mprid)
------------------------------------------------------------------------------
N_Credit_C~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   m_Have_IP |      0.091      0.063    1.444   0.149       -0.032       0.214
  m_LnAssets |      0.093      0.008   11.796   0.000        0.078       0.109
m_Home_Based |      0.028      0.042    0.676   0.499       -0.054       0.110
m_OO_D_edu~r |     -0.018      0.039   -0.451   0.652       -0.095       0.059
m_OO_work_~r |     -0.008      0.002   -4.236   0.000       -0.011      -0.004
m_OO_race_~r |     -0.064      0.059   -1.095   0.274       -0.180       0.051
m_OO_gende~r |     -0.006      0.047   -0.132   0.895       -0.098       0.086
m_Comp_adv~e |      0.182      0.056    3.252   0.001        0.072       0.291
  m_hightech |     -0.087      0.065   -1.356   0.175       -0.214       0.039
m_d1a_prov~e |      0.119      0.076    1.575   0.115       -0.029       0.267
m_d1b_prov~t |      0.034      0.050    0.666   0.505       -0.065       0.133
m_c4_numow~m |      0.067      0.031    2.187   0.029        0.007       0.127
   d_Have_IP |      0.028      0.036    0.765   0.444       -0.044       0.099
  d_LnAssets |      0.037      0.006    6.509   0.000        0.026       0.048
d_Home_Based |     -0.210      0.060   -3.528   0.000       -0.327      -0.093
d_OO_D_edu~r |     -0.093      0.078   -1.181   0.238       -0.246       0.061
d_OO_work_~r |     -0.001      0.007   -0.090   0.928       -0.015       0.013
d_OO_race_~r |      0.024      0.213    0.113   0.910       -0.394       0.442
d_OO_gende~r |     -0.086      0.163   -0.527   0.598       -0.407       0.235
d_Comp_adv~e |      0.069      0.029    2.403   0.017        0.013       0.125
  d_hightech |     -0.114      0.104   -1.096   0.273       -0.318       0.090
d_d1a_prov~e |      0.095      0.047    2.004   0.045        0.002       0.187
d_d1b_prov~t |      0.047      0.039    1.202   0.230       -0.030       0.123
d_c4_numow~m |      0.088      0.033    2.637   0.009        0.022       0.153
       _cons |     -0.813      0.127   -6.383   0.000       -1.063      -0.564
------------------------------------------------------------------------------
```

6.6 Analysis of Subpopulations

Analysis of subpopulations (also called domains analysis, subgroup analysis, subpopulation analysis, or subdomain analysis) refers to the computation of descriptive and analytical statistics for subpopulations, e.g., African-owned businesses, team-owned businesses, or home-based businesses.

A common mistake is the elimination of cases that do not belong to the subpopulation under study while carrying out the computation of descriptive and analytical statistics using the remainder cases. Given that the formation of a subpopulation is unrelated to the sample design, the subpopulation sample size is a random variable.[2] To calculate correctly the variance of an estimate for subpopulation, we should take the randomness in the subpopulation sample size into account by including all the data in the analysis. The implications of ignoring the randomness in the subpopulation sample size will lead to underestimated standard errors.

The subpop command allows us to conducted subpopulation analysis. In Stata, the subpopulation command (subpop) option only works with commands supported by the svy prefix command. The subpop (varname) option takes a zero/nonzero variable, and the subpopulation is defined by varname ≠ 0 and not missing, but we highly recommend generating a subpopulation dummy variable with 1 for the subpopulation you are studying, 0 for the observations not in the subpopulation and missing for either.

Stata commands that support the pweight option do not allow for the subpopulation command (subpop) option.

6.6.1 Pooled Regression

Examples 6.27 Robust Standard Errors

```
mi xeq 0:svy, subpop( Home_Based): reg tdca    LnAssets Net_Profit     i.Have_IP
OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 i.f13_trade_fin i.f19_res_dev, cformat(%6.3f)  sformat(%6.3f)   nolstretch
```

[2] In very rare cases where a stratum is the subpopulation (domain has a fixed sample size), eliminate cases are not a problem.

```
Number of strata   =         6                  Number of obs      =     15094
Number of PSUs     =      3051                  Population size    = 341730.39
                                                 Subpop. no. of obs =      6765
                                                 Subpop. size       = 141289.38
                                                 Design df          =      3045
                                                 F(   9,   3037)    =     11.88
                                                 Prob > F           =    0.0000
                                                 R-squared          =    0.0433
------------------------------------------------------------------------------
             |             Linearized
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |       0.016      0.002     7.810   0.000        0.012       0.020
  Net_Profit |      -0.000      0.000    -0.341   0.733       -0.000       0.000
   1.Have_IP |      -0.020      0.016    -1.273   0.203       -0.052       0.011
OO_D_educa~r |      -0.045      0.015    -2.976   0.003       -0.074      -0.015
OO_work_ex~r |      -0.003      0.001    -4.106   0.000       -0.004      -0.002
OO_race_wh~r |      -0.019      0.022    -0.870   0.384       -0.062       0.024
   PO_gender |       0.014      0.017     0.818   0.413       -0.020       0.048
1.f13_trad~n |       0.052      0.020     2.605   0.009        0.013       0.091
1.f19_res_~v |       0.016      0.017     0.981   0.326       -0.016       0.049
       _cons |       0.158      0.026     6.030   0.000        0.106       0.209
------------------------------------------------------------------------------


mi estimate :svy, subpop( Home_Based):reg tdca    LnAssets Net_Profit i.Have_IP
OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner PO_gender   ///
 i.f13_trade_fin i.f19_res_dev
mi estimate,    cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                    Imputations        =         5
Survey: Linear regression                        Number of obs      =     18286

Number of strata  =         6                  Population size    = 408495.43
Number of PSUs    =      3140                  Subpop. no. of obs =      9844
                                                 Subpop. size       =  205125.2
                                                 Average RVI        =    0.0142
                                                 Largest FMI        =    0.0413
                                                 Complete DF        =      3134
DF adjustment:   Small sample                    DF:     min        =   1346.85
                                                         avg        =   2606.30
                                                         max        =   3112.13
Model F test:       Equal FMI                    F(   9, 3059.4)    =     18.74
Within VCE type:   Linearized                    Prob > F           =    0.0000

------------------------------------------------------------------------------
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |       0.016      0.002    10.068   0.000        0.013       0.019
  Net_Profit |       0.000      0.000     0.547   0.585       -0.000       0.000
   1.Have_IP |      -0.008      0.014    -0.587   0.557       -0.036       0.020
OO_D_educa~r |      -0.046      0.013    -3.534   0.000       -0.071      -0.020
OO_work_ex~r |      -0.003      0.001    -5.210   0.000       -0.004      -0.002
OO_race_wh~r |      -0.020      0.019    -1.042   0.298       -0.056       0.017
   PO_gender |       0.022      0.015     1.464   0.143       -0.007       0.052
1.f13_trad~n |       0.049      0.017     2.880   0.004        0.016       0.083
1.f19_res_~v |       0.017      0.015     1.143   0.253       -0.012       0.046
       _cons |       0.131      0.022     5.920   0.000        0.088       0.175
------------------------------------------------------------------------------
```

6.6.2 Logit Models for Binary Response Variables

Examples 6.28 Robust Standard Errors

```
mi    xeq    0:svy,    subpop(    Home_Based):    logit    Have_IP    LnAssets
i.OO_D_education_owner     OO_work_exp_owner     OO_race_white_owner   PO_gender
i.Comp_advantage ///
i.hightech i.d1a_provide_service i.d1b_provide_product   ,    or   cformat(%6.3f)
sformat(%6.3f)   nolstretch

Survey: Logistic regression

Number of strata   =         6                 Number of obs      =     17213
Number of PSUs     =      3119                 Population size    = 385234.06
                                               Subpop. no. of obs =      8884
                                               Subpop. size       = 184793.04
                                               Design df          =      3113
                                               F(   9,   3105)    =     20.96
                                               Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
     Have_IP | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      1.006      0.017    0.371   0.710        0.974       1.039
1.OO_D_edu~r |      1.444      0.191    2.781   0.005        1.114       1.871
OO_work_ex~r |      1.008      0.007    1.118   0.264        0.994       1.021
OO_race_wh~r |      0.962      0.178   -0.211   0.833        0.669       1.382
   PO_gender |      0.834      0.129   -1.171   0.242        0.616       1.130
1.Comp_adv~e |      2.666      0.277    9.435   0.000        2.174       3.268
  1.hightech |      1.513      0.264    2.375   0.018        1.075       2.129
1.d1a_prov~e |      0.628      0.110   -2.666   0.008        0.446       0.884
1.d1b_prov~t |      2.290      0.288    6.596   0.000        1.790       2.929
       _cons |      0.094      0.029   -7.653   0.000        0.051       0.173
------------------------------------------------------------------------------
```

```
mi estimate:svy, subpop( Home_Based): logit Have_IP LnAssets
i.OO_D_education_owner OO_work_exp_owner OO_race_white_owner PO_gender
i.Comp_advantage ///
i.hightech i.d1a_provide_service i.d1b_provide_product , or cformat(%6.3f)
sformat(%6.3f) nolstretch
mi estimate, or cformat(%6.3f) sformat(%6.3f) nolstretch

Multiple-imputation estimates                   Imputations       =          5
Survey: Logistic regression                     Number of obs     =      18286

Number of strata  =          6                  Population size   = 408495.43
Number of PSUs    =       3140                  Subpop. no. of obs =       9844
                                                Subpop. size      =  205125.2
                                                Average RVI       =     0.0032
                                                Largest FMI       =     0.0109
                                                Complete DF       =       3134
DF adjustment:   Small sample                   DF:     min       =    2837.99
                                                        avg       =    3070.76
                                                        max       =    3130.36
Model F test:       Equal FMI                   F(   9, 3128.0)   =      22.35
Within VCE type:   Linearized                   Prob > F          =     0.0000

------------------------------------------------------------------------------
     Have_IP | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      1.009      0.016    0.578   0.563        0.978       1.042
1.OO_D_edu~r |      1.470      0.189    2.998   0.003        1.143       1.891
OO_work_ex~r |      1.005      0.007    0.809   0.418        0.992       1.019
OO_race_wh~r |      0.915      0.167   -0.490   0.624        0.640       1.308
   PO_gender |      0.874      0.132   -0.894   0.371        0.650       1.175
1.Comp_adv~e |      2.625      0.264    9.593   0.000        2.155       3.198
  1.hightech |      1.575      0.265    2.698   0.007        1.132       2.190
1.d1a_prov~e |      0.591      0.101   -3.081   0.002        0.423       0.826
1.d1b_prov~t |      2.276      0.280    6.689   0.000        1.788       2.896
       _cons |      0.100      0.030   -7.692   0.000        0.056       0.180
------------------------------------------------------------------------------
```

6.6.3 Multinomial Logit Models for Catagorical Response Variables

Examples 6.29 Robust Standard Errors

```
mi xeq 0:svy, subpop( Home_Based): mlogit Legal_Form i.Have_IP LnAssets
i.OO_D_education_owner     OO_work_exp_owner     OO_race_white_owner   PO_gender
i.Comp_advantage ///
i.hightech i.d1a_provide_service i.d1b_provide_product  ,   rrr  cformat(%6.3f)
sformat(%6.3f)   nolstretch  baseoutcome(1)

Survey: Multinomial logistic regression

Number of strata   =         6                 Number of obs      =     17213
Number of PSUs     =      3119                 Population size    = 385234.06
                                                Subpop. no. of obs =      8884
                                                Subpop. size       = 184793.04
                                                Design df          =      3113
                                                F(  30,   3084)    =      4.52
                                                Prob > F           =    0.0000
```

Legal_Form	RRR	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
1	(base outcome)					
2						
1.Have_IP	1.248	0.195	1.417	0.157	0.919	1.695
LnAssets	1.096	0.022	4.558	0.000	1.053	1.139
1.OO_D_edu~r	1.733	0.246	3.873	0.000	1.312	2.289
OO_work_ex~r	0.998	0.007	-0.327	0.743	0.983	1.012
OO_race_wh~r	1.215	0.256	0.925	0.355	0.804	1.837
PO_gender	1.928	0.328	3.855	0.000	1.381	2.692
1.Comp_adv~e	0.899	0.093	-1.029	0.304	0.734	1.101
1.hightech	1.241	0.264	1.017	0.309	0.818	1.883
1.d1a_prov~e	0.662	0.126	-2.159	0.031	0.456	0.963
1.d1b_prov~t	0.625	0.088	-3.339	0.001	0.475	0.824
_cons	0.227	0.084	-3.986	0.000	0.109	0.470
3						
1.Have_IP	1.319	0.255	1.435	0.151	0.903	1.926
LnAssets	1.161	0.029	5.936	0.000	1.105	1.219
1.OO_D_edu~r	1.010	0.171	0.061	0.952	0.725	1.408
OO_work_ex~r	0.997	0.009	-0.358	0.720	0.980	1.014
OO_race_wh~r	0.929	0.240	-0.287	0.774	0.560	1.540
PO_gender	1.564	0.325	2.150	0.032	1.040	2.350
1.Comp_adv~e	0.924	0.118	-0.623	0.534	0.720	1.186
1.hightech	1.845	0.432	2.620	0.009	1.167	2.919
1.d1a_prov~e	1.218	0.335	0.718	0.473	0.711	2.089
1.d1b_prov~t	0.704	0.121	-2.037	0.042	0.502	0.987
_cons	0.074	0.033	-5.855	0.000	0.031	0.178
4						
1.Have_IP	1.829	0.420	2.628	0.009	1.166	2.869
LnAssets	1.052	0.024	2.187	0.029	1.005	1.101
1.OO_D_edu~r	0.789	0.164	-1.141	0.254	0.525	1.186
OO_work_ex~r	0.991	0.011	-0.839	0.402	0.970	1.012
OO_race_wh~r	0.582	0.155	-2.039	0.042	0.346	0.979

```
   PO_gender |      1.432      0.359    1.431   0.153        0.875       2.343
1.Comp_adv~e |      0.802      0.119   -1.485   0.138        0.600       1.073
  1.hightech |      0.945      0.343   -0.157   0.875        0.463       1.926
1.d1a_prov~e |      0.653      0.205   -1.356   0.175        0.353       1.209
1.d1b_prov~t |      0.636      0.142   -2.029   0.043        0.410       0.985
       _cons |      0.302      0.144   -2.515   0.012        0.119       0.768
------------------------------------------------------------------------------



mi estimate:svy, subpop( Home_Based): mlogit Legal_Form i.Have_IP LnAssets
i.OO_D_education_owner     OO_work_exp_owner     OO_race_white_owner   PO_gender
i.Comp_advantage ///
i.hightech i.d1a_provide_service i.d1b_provide_product  ,   rrr  cformat(%6.3f)
sformat(%6.3f)   nolstretch  baseoutcome(1)
mi estimate,   rrr  cformat(%6.3f)  sformat(%6.3f)   nolstretch


Multiple-imputation estimates                   Imputations        =         5
Survey: Multinomial logistic regression         Number of obs      =     18286
Number of strata  =         6                   Population size    = 408495.43
Number of PSUs    =      3140                   Subpop. no. of obs =      9844
                                                Subpop. size       =  205125.2
                                                Average RVI        =    0.0006
                                                Largest FMI        =    0.0041
                                                Complete DF        =      3134
DF adjustment:   Small sample                   DF:     min        =   3078.29
                                                        avg        =   3126.91
                                                        max        =   3131.97
Model F test:       Equal FMI                   F(  30, 3132.0)    =      4.51
Within VCE type:   Linearized                   Prob > F           =    0.0000

------------------------------------------------------------------------------
  Legal_Form |        RRR   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |  (base outcome)
-------------+----------------------------------------------------------------
2            |
   1.Have_IP |      1.223      0.187    1.312   0.190        0.905       1.651
    LnAssets |      1.095      0.022    4.558   0.000        1.053       1.139
1.OO_D_edu~r |      1.649      0.230    3.585   0.000        1.254       2.167
OO_work_ex~r |      0.999      0.007   -0.159   0.874        0.985       1.013
OO_race_wh~r |      1.190      0.244    0.849   0.396        0.796       1.779
   PO_gender |      1.939      0.324    3.967   0.000        1.398       2.691
1.Comp_adv~e |      0.927      0.094   -0.748   0.455        0.760       1.131
  1.hightech |      1.283      0.272    1.174   0.240        0.846       1.944
1.d1a_prov~e |      0.669      0.126   -2.127   0.034        0.462       0.969
1.d1b_prov~t |      0.625      0.086   -3.415   0.001        0.477       0.818
       _cons |      0.224      0.083   -4.056   0.000        0.109       0.462
-------------+----------------------------------------------------------------
3            |
   1.Have_IP |      1.275      0.239    1.295   0.195        0.883       1.841
    LnAssets |      1.154      0.028    5.979   0.000        1.101       1.209
1.OO_D_edu~r |      0.987      0.163   -0.076   0.939        0.714       1.365
OO_work_ex~r |      0.997      0.008   -0.407   0.684        0.980       1.013
OO_race_wh~r |      0.967      0.241   -0.135   0.893        0.593       1.578
   PO_gender |      1.569      0.323    2.193   0.028        1.049       2.348
1.Comp_adv~e |      0.952      0.118   -0.399   0.690        0.747       1.213
  1.hightech |      1.839      0.428    2.619   0.009        1.165       2.903
1.d1a_prov~e |      1.207      0.319    0.713   0.476        0.719       2.027
1.d1b_prov~t |      0.698      0.116   -2.162   0.031        0.503       0.967
       _cons |      0.075      0.032   -6.040   0.000        0.032       0.174
```

```
-------------+----------------------------------------------------------------
4            |
   1.Have_IP |      1.814      0.396     2.724   0.006        1.182       2.784
    LnAssets |      1.051      0.024     2.189   0.029        1.005       1.099
1.OO_D_edu~r |      0.830      0.168    -0.918   0.359        0.558       1.235
OO_work_ex~r |      0.993      0.011    -0.704   0.481        0.972       1.013
OO_race_wh~r |      0.593      0.151    -2.048   0.041        0.360       0.978
   PO_gender |      1.564      0.386     1.812   0.070        0.964       2.536
1.Comp_adv~e |      0.866      0.124    -1.006   0.315        0.654       1.147
  1.hightech |      1.065      0.376     0.177   0.859        0.533       2.128
1.d1a_prov~e |      0.608      0.182    -1.662   0.097        0.338       1.094
1.d1b_prov~t |      0.614      0.132    -2.268   0.023        0.403       0.936
       _cons |      0.279      0.126    -2.835   0.005        0.116       0.675
------------------------------------------------------------------------------
```

6.6.4 Poisson Models for Count Data

Examples 6.30 Robust Standard Errors

```
mi xeq 0:svy, subpop( Home_Based):poisson   N_Credit_Cards     Have_IP LnAssets
OO_D_education_owner    OO_work_exp_owner    OO_race_white_owner  OO_gender_owner
Comp_advantage ///
hightech     d1a_provide_service     d1b_provide_product     c4_numowners_confirm
,cformat(%6.3f)  sformat(%6.3f)   nolstretch


Survey: Poisson regression

Number of strata   =         6                  Number of obs      =     17212
Number of PSUs     =      3119                  Population size    = 385225.99
                                                 Subpop. no. of obs =      8883
                                                 Subpop. size       = 184784.98
                                                 Design df          =      3113
                                                 F(  11,   3103)    =     17.42
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
N_Credit_C~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Have_IP |      0.144      0.062     2.323   0.020        0.023       0.266
    LnAssets |      0.083      0.008     9.931   0.000        0.066       0.099
OO_D_educa~r |     -0.087      0.052    -1.663   0.096       -0.190       0.016
OO_work_ex~r |     -0.009      0.003    -3.307   0.001       -0.014      -0.004
OO_race_wh~r |     -0.097      0.092    -1.055   0.292       -0.278       0.084
OO_gender_~r |      0.096      0.066     1.445   0.149       -0.034       0.226
Comp_advan~e |      0.206      0.044     4.714   0.000        0.120       0.291
    hightech |     -0.018      0.121    -0.150   0.881       -0.256       0.219
d1a_provid~e |      0.026      0.091     0.286   0.775       -0.152       0.204
d1b_provid~t |      0.056      0.055     1.027   0.305       -0.051       0.163
c4_numowne~m |      0.167      0.044     3.753   0.000        0.080       0.254
       _cons |     -0.875      0.158    -5.554   0.000       -1.184      -0.566
------------------------------------------------------------------------------
```

```
mi estimate:svy, subpop( Home_Based):poisson  N_Credit_Cards    Have_IP LnAssets
OO_D_education_owner    OO_work_exp_owner    OO_race_white_owner  OO_gender_owner
Comp_advantage ///
hightech     d1a_provide_service     d1b_provide_product     c4_numowners_confirm
,cformat(%6.3f)  sformat(%6.3f)   nolstretch
mi estimate,cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                   Imputations        =         5
Survey: Poisson regression                      Number of obs      =     18286

Number of strata  =         6                   Population size    = 408495.43
Number of PSUs    =      3140                   Subpop. no. of obs =      9844
                                                Subpop. size       =  205125.2
                                                Average RVI        =    0.0215
                                                Largest FMI        =    0.0567
                                                Complete DF        =      3134
DF adjustment:   Small sample                   DF:     min        =    908.54
                                                        avg        =   2298.52
                                                        max        =   3051.85
Model F test:       Equal FMI                   F(  11, 2999.9)    =     20.35
Within VCE type:   Linearized                   Prob > F           =    0.0000

------------------------------------------------------------------------------
N_Credit_C~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Have_IP |       0.079      0.058     1.363   0.173       -0.035       0.192
    LnAssets |       0.087      0.008    10.934   0.000        0.071       0.102
OO_D_educa~r |      -0.069      0.049    -1.416   0.157       -0.166       0.027
OO_work_ex~r |      -0.007      0.003    -2.922   0.004       -0.012      -0.002
OO_race_wh~r |      -0.099      0.083    -1.195   0.232       -0.262       0.064
OO_gender_~r |       0.082      0.063     1.309   0.191       -0.041       0.204
Comp_advan~e |       0.178      0.041     4.340   0.000        0.098       0.259
    hightech |      -0.056      0.106    -0.528   0.598       -0.264       0.152
d1a_provid~e |       0.042      0.086     0.485   0.627       -0.126       0.209
d1b_provid~t |       0.035      0.051     0.701   0.483       -0.064       0.135
c4_numowne~m |       0.147      0.042     3.545   0.000        0.066       0.229
       _cons |      -0.759      0.146    -5.190   0.000       -1.046      -0.472
------------------------------------------------------------------------------
```

6.6.5 Negative Binomial Models for Count Data

Examples 6.31 Robust Standard Errors

```
mi xeq 0:svy, subpop(if Home_Based==1 & Legal_Form==1):nbreg   N_Credit_Cards
Have_IP LnAssets     OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner
OO_gender_owner  Comp_advantage ///
hightech   d1a_provide_service   d1b_provide_product             ,cformat(%6.3f)
sformat(%6.3f)   nolstretch

Survey: Negative binomial regression

Number of strata   =         6                  Number of obs      =     18021
Number of PSUs     =      3135                  Population size    =  402655.1
                                                 Subpop. no. of obs =      3719
                                                 Subpop. size       = 82353.228
                                                 Design df          =      3129
                                                 F(  10,   3120)    =      8.59
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
N_Credit_C~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Have_IP |      0.077      0.083    0.927   0.354       -0.086       0.241
    LnAssets |      0.085      0.012    7.038   0.000        0.061       0.108
  Home_Based |      0.000  (omitted)
OO_D_educa~r |      0.190      0.074    2.571   0.010        0.045       0.335
OO_work_ex~r |     -0.010      0.003   -2.900   0.004       -0.017      -0.003
OO_race_wh~r |      0.158      0.120    1.308   0.191       -0.079       0.394
OO_gender_~r |      0.192      0.084    2.284   0.022        0.027       0.358
Comp_advan~e |      0.176      0.067    2.626   0.009        0.045       0.307
    hightech |     -0.012      0.147   -0.081   0.936       -0.301       0.277
d1a_provid~e |     -0.171      0.116   -1.472   0.141       -0.400       0.057
d1b_provid~t |     -0.042      0.074   -0.578   0.564       -0.187       0.102
c4_numowne~m |      0.000  (omitted)
       _cons |     -0.934      0.205   -4.564   0.000       -1.335      -0.533
-------------+----------------------------------------------------------------
    /lnalpha |     -0.646      0.132                        -0.904      -0.387
-------------+----------------------------------------------------------------
       alpha |      0.524      0.069                         0.405       0.679
------------------------------------------------------------------------------
```

```
mi estimate:svy, subpop(if Home_Based==1 & Legal_Form==1):nbreg  N_Credit_Cards
Have_IP LnAssets     OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner
OO_gender_owner  Comp_advantage ///
hightech   d1a_provide_service   d1b_provide_product            ,cformat(%6.3f)
sformat(%6.3f)   nolstretch
mi estimate,cformat(%6.3f)  sformat(%6.3f)   nolstretch


Multiple-imputation estimates                   Imputations        =         5
Survey: Negative binomial regression            Number of obs      =     18286

Number of strata  =         6                   Population size    = 408495.43
Number of PSUs    =      3140                   Subpop. no. of obs =      4144
                                                Subpop. size       = 92112.744
                                                Average RVI        =    0.0266
                                                Largest FMI        =    0.0726
                                                Complete DF        =      3134
DF adjustment:   Small sample                   DF:     min        =    633.46
                                                        avg        =   1975.77
                                                        max        =   3116.66
Model F test:       Equal FMI                   F(  10, 2915.4)    =      9.63
Within VCE type:   Linearized                   Prob > F           =    0.0000
------------------------------------------------------------------------------
N_Credit_C~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Have_IP |      0.005      0.077    0.069   0.945       -0.146       0.157
    LnAssets |      0.092      0.012    7.713   0.000        0.068       0.115
OO_D_educa~r |      0.202      0.070    2.880   0.004        0.064       0.339
OO_work_ex~r |     -0.008      0.003   -2.359   0.018       -0.014      -0.001
OO_race_wh~r |      0.089      0.111    0.807   0.420       -0.128       0.306
OO_gender_~r |      0.189      0.078    2.413   0.016        0.036       0.343
Comp_advan~e |      0.139      0.065    2.134   0.033        0.011       0.266
    hightech |     -0.042      0.129   -0.328   0.743       -0.295       0.211
d1a_provid~e |     -0.114      0.106   -1.077   0.282       -0.323       0.094
d1b_provid~t |     -0.031      0.068   -0.456   0.648       -0.165       0.103
       _cons |     -0.869      0.193   -4.506   0.000       -1.248      -0.491
-------------+----------------------------------------------------------------
    /lnalpha |     -0.932      0.147                       -1.220      -0.645
-------------+----------------------------------------------------------------
       alpha |      0.394      0.058                        0.295       0.525
------------------------------------------------------------------------------
```

6.7 Working with Balanced Panel Data

In a balanced panel, the number of time periods is the same for all firms. Thus, working with balance data required that we limit the data to survival firms that responded to every survey. For the KFS data, this means limiting the analysis to the 1,630 survival firms. All methods of analysis that we used in sections 6.4, 6.5, and 6.6 can be used with balanced panel data.

While the weights in the KFS control and make the required adjustment for attrition and sample selection biases, weights do not control for survivorship bias. Thus, working with balance data could require controlling for survivorship bias.

6.8 Structural Equation Modeling (SEM)

In Stata, the "sem" command fits structural equation models, the "sem" command supports the svy command prefix as well as the pweight option. Utilizing the "sem" command required the data to be in the wide format.

Examples 6.32 Cluster-Robust Standard Errors using SEM

Using balanced panel data opens the door for the use of structural equation modeling methods.

```
keep if Duration==8

*Robust Standard Errors
mi estimate :svy:reg tdca  LnAssets Home_Based PO_gender OO_work_exp_owner Have_IP
OO_D_education_owner OO_race_white_owner
mi estimate,    cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                   Imputations        =         5
Survey: Linear regression                       Number of obs      =     13040

Number of strata  =         6                   Population size    = 279136.31
Number of PSUs    =      1630
                                                Average RVI        =    0.0121
                                                Largest FMI        =    0.0353
                                                Complete DF        =      1624
DF adjustment:   Small sample                   DF:     min        =   1064.28
                                                        avg        =   1439.61
                                                        max        =   1619.06
Model F test:       Equal FMI                   F(   7, 1602.7)    =     20.48
Within VCE type:   Linearized                   Prob > F           =    0.0000

------------------------------------------------------------------------------
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |       0.015      0.002   10.030   0.000        0.012       0.018
  Home_Based |      -0.034      0.013   -2.672   0.008       -0.058      -0.009
   PO_gender |       0.010      0.015    0.709   0.478       -0.018       0.039
OO_work_ex~r |      -0.001      0.001   -2.374   0.018       -0.003      -0.000
     Have_IP |      -0.010      0.012   -0.854   0.393       -0.034       0.013
OO_D_educa~r |      -0.037      0.012   -3.000   0.003       -0.060      -0.013
OO_race_wh~r |       0.007      0.018    0.413   0.680       -0.027       0.042
       _cons |       0.127      0.025    5.147   0.000        0.079       0.176
------------------------------------------------------------------------------
```

```
* _OEx, meaning all observed exogenous variables in your model
* _LEx, meaning all latent exogenous variables in your model
*  diagonal all variances unrestricted and all covariances fixed at 0
* _OEn, meaning all error variables associated with observed endogenous variables
in your model
* _LEn, meaning all error variables associated with latent endogenous variables
in your model
* _En, meaning all error variables in your model

preserve

mi xtset,clear
keep   tdca        LnAssets   Home_Based   PO_gender   OO_work_exp_owner   Have_IP
OO_D_education_owner OO_race_white_owner ///
wgt_7_long mprid year sampleinfo_samplestrata _mi_m _mi_id _mi_miss

mi reshape wide   tdca   LnAssets Home_Based PO_gender OO_work_exp_owner Have_IP
OO_D_education_owner OO_race_white_owner , i(mprid) j(year)


 mi svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)

  mi estimate ,cmdok:svy: sem        ///
(      tdca2004      <-      _cons@c      LnAssets2004 @b1
       Home_Based2004       @b2          PO_gender2004                          @b3
       OO_work_exp_owner2004       @b4   Have_IP2004                            @b5
       OO_D_education_owner2004 @b6      OO_race_white_owner2004 @b7        ) ///
(      tdca2005      <-      _cons@c      LnAssets2005 @b1
       Home_Based2005       @b2          PO_gender2004                          @b3
       OO_work_exp_owner2005       @b4   Have_IP2005                            @b5
       OO_D_education_owner2005 @b6      OO_race_white_owner2005 @b7        ) ///
(      tdca2006      <-      _cons@c      LnAssets2006 @b1
       Home_Based2006       @b2          PO_gender2004                          @b3
       OO_work_exp_owner2006       @b4   Have_IP2006                            @b5
       OO_D_education_owner2006 @b6      OO_race_white_owner2006 @b7        ) ///
(      tdca2007      <-      _cons@c      LnAssets2007 @b1
       Home_Based2007       @b2          PO_gender2004                          @b3
       OO_work_exp_owner2007       @b4   Have_IP2007                            @b5
       OO_D_education_owner2007 @b6      OO_race_white_owner2007 @b7        ) ///
(      tdca2008      <-      _cons@c      LnAssets2008 @b1
       Home_Based2008       @b2          PO_gender2004                          @b3
       OO_work_exp_owner2008       @b4   Have_IP2008                            @b5
       OO_D_education_owner2008 @b6      OO_race_white_owner2008 @b7        ) ///
(      tdca2009      <-      _cons@c      LnAssets2009 @b1
       Home_Based2009       @b2          PO_gender2004                          @b3
       OO_work_exp_owner2009       @b4   Have_IP2009                            @b5
       OO_D_education_owner2009 @b6      OO_race_white_owner2009 @b7        ) ///
(      tdca2010      <-      _cons@c      LnAssets2010 @b1
       Home_Based2010       @b2          PO_gender2004                          @b3
       OO_work_exp_owner2010       @b4   Have_IP2010                            @b5
       OO_D_education_owner2010 @b6      OO_race_white_owner2010 @b7        ) ///
(      tdca2011      <-      _cons@c      LnAssets2011 @b1
       Home_Based2011       @b2          PO_gender2004                          @b3
       OO_work_exp_owner2011       @b4   Have_IP2011                            @b5
       OO_D_education_owner2011 @b6      OO_race_white_owner2011 @b7        ) ///
, cov(e._OEn@S_e )      nocapslatent
mi estimate,    cformat(%6.3f)  sformat(%6.3f)   nolstretch

restore
```

```
Multiple-imputation estimates                   Imputations        =         5
Survey: Structural equation model               Number of obs      =      1630

Number of strata  =         6                   Population size    = 34892.039
Number of PSUs    =      1630
                                                Average RVI        =    0.0109
                                                Largest FMI        =    0.0353
                                                Complete DF        =      1624
DF adjustment:   Small sample                   DF:     min        =   1064.28
                                                        avg        =   1457.47
Within VCE type:   Linearized                           max        =   1619.06

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  td~2004 <- |
  LnAss~2004 |      0.015      0.002   10.030   0.000        0.012       0.018
  Home_~2004 |     -0.034      0.013   -2.672   0.008       -0.058      -0.009
  PO_ge~2004 |      0.010      0.015    0.709   0.478       -0.018       0.039
  OO_wo~2004 |     -0.001      0.001   -2.374   0.018       -0.003      -0.000
  Have_~2004 |     -0.010      0.012   -0.854   0.393       -0.034       0.013
  OO_D_~2004 |     -0.037      0.012   -3.000   0.003       -0.060      -0.013
  OO_ra~2004 |      0.007      0.018    0.413   0.680       -0.027       0.042
       _cons |      0.127      0.025    5.147   0.000        0.079       0.176
  -----------+----------------------------------------------------------------
  td~2005 <- |
  PO_ge~2004 |      0.010      0.015    0.709   0.478       -0.018       0.039
  LnAss~2005 |      0.015      0.002   10.030   0.000        0.012       0.018
  Home_~2005 |     -0.034      0.013   -2.672   0.008       -0.058      -0.009
  OO_wo~2005 |     -0.001      0.001   -2.374   0.018       -0.003      -0.000
  Have_~2005 |     -0.010      0.012   -0.854   0.393       -0.034       0.013
  OO_D_~2005 |     -0.037      0.012   -3.000   0.003       -0.060      -0.013
  OO_ra~2005 |      0.007      0.018    0.413   0.680       -0.027       0.042
       _cons |      0.127      0.025    5.147   0.000        0.079       0.176
  -----------+----------------------------------------------------------------
  td~2006 <- |
  PO_ge~2004 |      0.010      0.015    0.709   0.478       -0.018       0.039
  LnAss~2006 |      0.015      0.002   10.030   0.000        0.012       0.018
  Home_~2006 |     -0.034      0.013   -2.672   0.008       -0.058      -0.009
  OO_wo~2006 |     -0.001      0.001   -2.374   0.018       -0.003      -0.000
  Have_~2006 |     -0.010      0.012   -0.854   0.393       -0.034       0.013
  OO_D_~2006 |     -0.037      0.012   -3.000   0.003       -0.060      -0.013
  OO_ra~2006 |      0.007      0.018    0.413   0.680       -0.027       0.042
       _cons |      0.127      0.025    5.147   0.000        0.079       0.176
  -----------+----------------------------------------------------------------
  td~2007 <- |
  PO_ge~2004 |      0.010      0.015    0.709   0.478       -0.018       0.039
  LnAss~2007 |      0.015      0.002   10.030   0.000        0.012       0.018
  Home_~2007 |     -0.034      0.013   -2.672   0.008       -0.058      -0.009
  OO_wo~2007 |     -0.001      0.001   -2.374   0.018       -0.003      -0.000
  Have_~2007 |     -0.010      0.012   -0.854   0.393       -0.034       0.013
  OO_D_~2007 |     -0.037      0.012   -3.000   0.003       -0.060      -0.013
  OO_ra~2007 |      0.007      0.018    0.413   0.680       -0.027       0.042
       _cons |      0.127      0.025    5.147   0.000        0.079       0.176
  -----------+----------------------------------------------------------------
  td~2008 <- |
  PO_ge~2004 |      0.010      0.015    0.709   0.478       -0.018       0.039
  LnAss~2008 |      0.015      0.002   10.030   0.000        0.012       0.018
  Home_~2008 |     -0.034      0.013   -2.672   0.008       -0.058      -0.009
  OO_wo~2008 |     -0.001      0.001   -2.374   0.018       -0.003      -0.000
  Have_~2008 |     -0.010      0.012   -0.854   0.393       -0.034       0.013
```

```
  OO_D_~2008 |     -0.037      0.012   -3.000   0.003       -0.060      -0.013
  OO_ra~2008 |      0.007      0.018    0.413   0.680       -0.027       0.042
       _cons |      0.127      0.025    5.147   0.000        0.079       0.176
  -----------+------------------------------------------------------------------
  td~2009 <- |
  PO_ge~2004 |      0.010      0.015    0.709   0.478       -0.018       0.039
  LnAss~2009 |      0.015      0.002   10.030   0.000        0.012       0.018
  Home_~2009 |     -0.034      0.013   -2.672   0.008       -0.058      -0.009
  OO_wo~2009 |     -0.001      0.001   -2.374   0.018       -0.003      -0.000
  Have_~2009 |     -0.010      0.012   -0.854   0.393       -0.034       0.013
  OO_D_~2009 |     -0.037      0.012   -3.000   0.003       -0.060      -0.013
  OO_ra~2009 |      0.007      0.018    0.413   0.680       -0.027       0.042
       _cons |      0.127      0.025    5.147   0.000        0.079       0.176
  -----------+------------------------------------------------------------------
  td~2010 <- |
  PO_ge~2004 |      0.010      0.015    0.709   0.478       -0.018       0.039
  LnAss~2010 |      0.015      0.002   10.030   0.000        0.012       0.018
  Home_~2010 |     -0.034      0.013   -2.672   0.008       -0.058      -0.009
  OO_wo~2010 |     -0.001      0.001   -2.374   0.018       -0.003      -0.000
  Have_~2010 |     -0.010      0.012   -0.854   0.393       -0.034       0.013
  OO_D_~2010 |     -0.037      0.012   -3.000   0.003       -0.060      -0.013
  OO_ra~2010 |      0.007      0.018    0.413   0.680       -0.027       0.042
       _cons |      0.127      0.025    5.147   0.000        0.079       0.176
  -----------+------------------------------------------------------------------
  td~2011 <- |
  PO_ge~2004 |      0.010      0.015    0.709   0.478       -0.018       0.039
  LnAss~2011 |      0.015      0.002   10.030   0.000        0.012       0.018
  Home_~2011 |     -0.034      0.013   -2.672   0.008       -0.058      -0.009
  OO_wo~2011 |     -0.001      0.001   -2.374   0.018       -0.003      -0.000
  Have_~2011 |     -0.010      0.012   -0.854   0.393       -0.034       0.013
  OO_D_~2011 |     -0.037      0.012   -3.000   0.003       -0.060      -0.013
  OO_ra~2011 |      0.007      0.018    0.413   0.680       -0.027       0.042
       _cons |      0.127      0.025    5.147   0.000        0.079       0.176
-------------+------------------------------------------------------------------
Variance     |
  e.tdca2004 |      0.108      0.003                        0.103       0.113
  e.tdca2005 |      0.108      0.003                        0.103       0.113
  e.tdca2006 |      0.108      0.003                        0.103       0.113
  e.tdca2007 |      0.108      0.003                        0.103       0.113
  e.tdca2008 |      0.108      0.003                        0.103       0.113
  e.tdca2009 |      0.108      0.003                        0.103       0.113
  e.tdca2010 |      0.108      0.003                        0.103       0.113
  e.tdca2011 |      0.108      0.003                        0.103       0.113
------------------------------------------------------------------------------
```

Examples 6.33 Fixed Effects using SEM

```
mi estimate :xtreg tdca   LnAssets Home_Based PO_gender OO_work_exp_owner Have_IP
OO_D_education_owner  OO_race_white_owner    [pweight=wgt_7_long],  fe  i(mprid)
vce(robust)
mi estimate,    cformat(%6.3f)  sformat(%6.3f)   nolstretch


Multiple-imputation estimates                   Imputations        =         5
Fixed-effects (within) regression               Number of obs      =     13040

Group variable: mprid                           Number of groups   =      1630
                                                Obs per group: min =         8
                                                               avg =       8.0
                                                               max =         8

                                                Average RVI        =  232.3401
                                                Largest FMI        =    0.2149
                                                Complete DF        =      1629
DF adjustment:   Small sample                   DF:     min        =     93.50
                                                        avg        =    836.95
                                                        max        =   1594.90
Model F test:       Equal FMI                   F(   6, 1192.4)    =      3.24
Within VCE type:       Robust                   Prob > F           =    0.0037

                            (Within VCE adjusted for 1630 clusters in mprid)
------------------------------------------------------------------------------
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      0.006      0.002    3.877   0.000        0.003       0.009
  Home_Based |     -0.017      0.024   -0.709   0.478       -0.063       0.030
   PO_gender |      0.000  (omitted)
OO_work_ex~r |      0.001      0.002    0.402   0.688       -0.003       0.005
     Have_IP |      0.015      0.013    1.154   0.249       -0.010       0.039
OO_D_educa~r |     -0.017      0.024   -0.718   0.473       -0.064       0.030
OO_race_wh~r |     -0.041      0.067   -0.618   0.538       -0.173       0.091
       _cons |      0.217      0.067    3.236   0.002        0.084       0.351
-------------+----------------------------------------------------------------
     sigma_u |      0.214
     sigma_e |      0.270
         rho |      0.384   (fraction of variance due to u_i)
------------------------------------------------------------------------------
Note: sigma_u and sigma_e are combined in the original metric.




mi xtset,clear
keep   tdca        LnAssets   Home_Based   PO_gender   OO_work_exp_owner   Have_IP
OO_D_education_owner OO_race_white_owner ///
wgt_7_long mprid year sampleinfo_samplestrata _mi_m _mi_id _mi_miss

mi reshape wide   tdca   LnAssets Home_Based PO_gender OO_work_exp_owner Have_IP
OO_D_education_owner OO_race_white_owner , i(mprid) j(year)


 mi svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
  set matsize 11000
  mi estimate ,cmdok post:  sem        ///
(      tdca2004     <-     L@1     _cons@c     LnAssets2004 @b1
       Home_Based2004       @b2          OO_work_exp_owner2004@b3
```

```
       OO_D_education_owner2004 @b6     Have_IP2004                          @b5
       OO_race_white_owner2004 @b7       ) ///
(      tdca2005     <-     L@1     _cons@c     LnAssets2005 @b1
       Home_Based2005       @b2          OO_work_exp_owner2005@b3
       OO_D_education_owner2005 @b6     Have_IP2005                          @b5
       OO_race_white_owner2005 @b7       ) ///
(      tdca2006     <-     L@1     _cons@c     LnAssets2006 @b1
       Home_Based2006       @b2          OO_work_exp_owner2006@b3
       OO_D_education_owner2006 @b6     Have_IP2006                          @b5
       OO_race_white_owner2006 @b7       ) ///
(      tdca2007     <-     L@1     _cons@c     LnAssets2007 @b1
       Home_Based2007       @b2          OO_work_exp_owner2007@b3
       OO_D_education_owner2007 @b6     Have_IP2007                          @b5
       OO_race_white_owner2007 @b7       ) ///
(      tdca2008     <-     L@1     _cons@c     LnAssets2008 @b1
       Home_Based2008       @b2          OO_work_exp_owner2008@b3
       OO_D_education_owner2008 @b6     Have_IP2008                          @b5
       OO_race_white_owner2008 @b7       ) ///
(      tdca2009     <-     L@1     _cons@c     LnAssets2009 @b1
       Home_Based2009       @b2          OO_work_exp_owner2009@b3
       OO_D_education_owner2009 @b6     Have_IP2009                          @b5
       OO_race_white_owner2009 @b7       ) ///
(      tdca2010     <-     L@1     _cons@c     LnAssets2010 @b1
       Home_Based2010       @b2          OO_work_exp_owner2010@b3
       OO_D_education_owner2010 @b6     Have_IP2010                          @b5
       OO_race_white_owner2010 @b7       ) ///
(      tdca2011     <-     L@1     _cons@c     LnAssets2011 @b1
       Home_Based2011       @b2          OO_work_exp_owner2011@b3
       OO_D_education_owner2011 @b6     Have_IP2011                          @b5
       OO_race_white_owner2011 @b7       ) ///
 [pweight=wgt_7_long]    ,latent(L  )      cov(e._OEn@S_e  )    cov(_OEx*_OEx    )
nocapslatent   vce(robust)   notable

 mi estimate,    cformat(%6.3f)  sformat(%6.3f)   nolstretch


------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  td~2004 <- |
  LnAss~2004 |      0.006      0.002    3.878   0.000        0.003       0.009
  Home_~2004 |     -0.017      0.024   -0.709   0.478       -0.063       0.030
  OO_wo~2004 |      0.001      0.002    0.402   0.688       -0.003       0.005
  OO_D_~2004 |     -0.017      0.024   -0.718   0.473       -0.064       0.030
  Have_~2004 |      0.015      0.013    1.154   0.249       -0.010       0.039
  OO_ra~2004 |     -0.041      0.067   -0.618   0.538       -0.173       0.090
           L |      1.000      0.000  7.3e+15   0.000        1.000       1.000
       _cons |      0.217      0.067    3.231   0.002        0.084       0.351
  -----------+----------------------------------------------------------------
  td~2005 <- |
  LnAss~2005 |      0.006      0.002    3.878   0.000        0.003       0.009
  Home_~2005 |     -0.017      0.024   -0.709   0.478       -0.063       0.030
  OO_wo~2005 |      0.001      0.002    0.402   0.688       -0.003       0.005
  OO_D_~2005 |     -0.017      0.024   -0.718   0.473       -0.064       0.030
  Have_~2005 |      0.015      0.013    1.154   0.249       -0.010       0.039
  OO_ra~2005 |     -0.041      0.067   -0.618   0.538       -0.173       0.090
           L |      1.000  (constrained)
       _cons |      0.217      0.067    3.231   0.002        0.084       0.351
  -----------+----------------------------------------------------------------
  td~2006 <- |
  LnAss~2006 |      0.006      0.002    3.878   0.000        0.003       0.009
```

```
  Home_~2006 |     -0.017      0.024   -0.709   0.478       -0.063       0.030
  OO_wo~2006 |      0.001      0.002    0.402   0.688       -0.003       0.005
  OO_D_~2006 |     -0.017      0.024   -0.718   0.473       -0.064       0.030
  Have_~2006 |      0.015      0.013    1.154   0.249       -0.010       0.039
  OO_ra~2006 |     -0.041      0.067   -0.618   0.538       -0.173       0.090
           L |      1.000      0.000  1.6e+42   0.000        1.000       1.000
       _cons |      0.217      0.067    3.231   0.002        0.084       0.351
  -----------+------------------------------------------------------------------
  td~2007 <- |
  LnAss~2007 |      0.006      0.002    3.878   0.000        0.003       0.009
  Home_~2007 |     -0.017      0.024   -0.709   0.478       -0.063       0.030
  OO_wo~2007 |      0.001      0.002    0.402   0.688       -0.003       0.005
  OO_D_~2007 |     -0.017      0.024   -0.718   0.473       -0.064       0.030
  Have_~2007 |      0.015      0.013    1.154   0.249       -0.010       0.039
  OO_ra~2007 |     -0.041      0.067   -0.618   0.538       -0.173       0.090
           L |      1.000      0.000  7.4e+15   0.000        1.000       1.000
       _cons |      0.217      0.067    3.231   0.002        0.084       0.351
  -----------+------------------------------------------------------------------
  td~2008 <- |
  LnAss~2008 |      0.006      0.002    3.878   0.000        0.003       0.009
  Home_~2008 |     -0.017      0.024   -0.709   0.478       -0.063       0.030
  OO_wo~2008 |      0.001      0.002    0.402   0.688       -0.003       0.005
  OO_D_~2008 |     -0.017      0.024   -0.718   0.473       -0.064       0.030
  Have_~2008 |      0.015      0.013    1.154   0.249       -0.010       0.039
  OO_ra~2008 |     -0.041      0.067   -0.618   0.538       -0.173       0.090
           L |      1.000      0.000  1.4e+43   0.000        1.000       1.000
       _cons |      0.217      0.067    3.231   0.002        0.084       0.351
  -----------+------------------------------------------------------------------
  td~2009 <- |
  LnAss~2009 |      0.006      0.002    3.878   0.000        0.003       0.009
  Home_~2009 |     -0.017      0.024   -0.709   0.478       -0.063       0.030
  OO_wo~2009 |      0.001      0.002    0.402   0.688       -0.003       0.005
  OO_D_~2009 |     -0.017      0.024   -0.718   0.473       -0.064       0.030
  Have_~2009 |      0.015      0.013    1.154   0.249       -0.010       0.039
  OO_ra~2009 |     -0.041      0.067   -0.618   0.538       -0.173       0.090
           L |      1.000      0.000  2.3e+42   0.000        1.000       1.000
       _cons |      0.217      0.067    3.231   0.002        0.084       0.351
  -----------+------------------------------------------------------------------
  td~2010 <- |
  LnAss~2010 |      0.006      0.002    3.878   0.000        0.003       0.009
  Home_~2010 |     -0.017      0.024   -0.709   0.478       -0.063       0.030
  OO_wo~2010 |      0.001      0.002    0.402   0.688       -0.003       0.005
  OO_D_~2010 |     -0.017      0.024   -0.718   0.473       -0.064       0.030
  Have_~2010 |      0.015      0.013    1.154   0.249       -0.010       0.039
  OO_ra~2010 |     -0.041      0.067   -0.618   0.538       -0.173       0.090
           L |      1.000      0.000  1.1e+43   0.000        1.000       1.000
       _cons |      0.217      0.067    3.231   0.002        0.084       0.351
  -----------+------------------------------------------------------------------
  td~2011 <- |
  LnAss~2011 |      0.006      0.002    3.878   0.000        0.003       0.009
  Home_~2011 |     -0.017      0.024   -0.709   0.478       -0.063       0.030
  OO_wo~2011 |      0.001      0.002    0.402   0.688       -0.003       0.005
  OO_D_~2011 |     -0.017      0.024   -0.718   0.473       -0.064       0.030
  Have_~2011 |      0.015      0.013    1.154   0.249       -0.010       0.039
  OO_ra~2011 |     -0.041      0.067   -0.618   0.538       -0.173       0.090
           L |      1.000  (constrained)
       _cons |      0.217      0.067    3.231   0.002        0.084       0.351
-------------+------------------------------------------------------------------
Variance     |
  e.tdca2004 |      0.073      0.002                        0.069       0.077
  e.tdca2005 |      0.073      0.002                        0.069       0.077
  e.tdca2006 |      0.073      0.002                        0.069       0.077
```

e.tdca2007	0.073	0.002			0.069	0.077
e.tdca2008	0.073	0.002			0.069	0.077
e.tdca2009	0.073	0.002			0.069	0.077
e.tdca2010	0.073	0.002			0.069	0.077
e.tdca2011	0.073	0.002			0.069	0.077
L	0.037	0.003			0.032	0.042
Covariance						
LnAss~2004						
L	0.091	0.023	4.003	0.000	0.046	0.136
Home_~2004						
L	-0.005	0.006	-0.913	0.361	-0.017	0.006
OO_wo~2004						
L	-0.209	0.219	-0.952	0.341	-0.639	0.221
OO_D_~2004						
L	-0.004	0.006	-0.610	0.542	-0.016	0.008
Have_~2004						
L	-0.003	0.003	-1.266	0.206	-0.008	0.002
OO_ra~2004						
L	0.007	0.008	0.786	0.433	-0.010	0.023
LnAss~2005						
L	0.116	0.021	5.523	0.000	0.075	0.158
Home_~2005						
L	-0.007	0.006	-1.220	0.223	-0.019	0.004
OO_wo~2005						
L	-0.219	0.217	-1.012	0.312	-0.644	0.205
OO_D_~2005						
L	-0.004	0.006	-0.684	0.494	-0.017	0.008
Have_~2005						
L	0.001	0.003	0.305	0.760	-0.005	0.006
OO_ra~2005						
L	0.006	0.008	0.756	0.451	-0.010	0.023
LnAss~2006						
L	0.124	0.024	5.162	0.000	0.077	0.171
Home_~2006						
L	-0.006	0.006	-1.047	0.295	-0.018	0.006
OO_wo~2006						
L	-0.201	0.219	-0.917	0.359	-0.630	0.229
OO_D_~2006						
L	-0.004	0.006	-0.582	0.561	-0.016	0.009
Have_~2006						
L	-0.005	0.003	-1.899	0.058	-0.011	0.000
OO_ra~2006						
L	0.006	0.008	0.775	0.439	-0.010	0.023

```
LnAss~2007 |
         L |      0.101      0.024     4.255   0.000        0.055       0.148
-----------+------------------------------------------------------------------
Home_~2007 |
         L |     -0.008      0.006    -1.342   0.180       -0.020       0.004
-----------+------------------------------------------------------------------
OO_wo~2007 |
         L |     -0.191      0.219    -0.875   0.382       -0.620       0.238
-----------+------------------------------------------------------------------
OO_D_~2007 |
         L |     -0.004      0.006    -0.712   0.476       -0.017       0.008
-----------+------------------------------------------------------------------
Have_~2007 |
         L |     -0.004      0.003    -1.528   0.127       -0.009       0.001
-----------+------------------------------------------------------------------
OO_ra~2007 |
         L |      0.007      0.009     0.825   0.410       -0.010       0.024
-----------+------------------------------------------------------------------
LnAss~2008 |
         L |      0.101      0.024     4.166   0.000        0.053       0.148
-----------+------------------------------------------------------------------
Home_~2008 |
         L |     -0.009      0.006    -1.406   0.160       -0.021       0.003
-----------+------------------------------------------------------------------
OO_wo~2008 |
         L |     -0.202      0.219    -0.924   0.356       -0.631       0.227
-----------+------------------------------------------------------------------
OO_D_~2008 |
         L |     -0.006      0.006    -0.998   0.318       -0.018       0.006
-----------+------------------------------------------------------------------
Have_~2008 |
         L |     -0.003      0.002    -1.111   0.266       -0.008       0.002
-----------+------------------------------------------------------------------
OO_ra~2008 |
         L |      0.007      0.009     0.822   0.412       -0.010       0.024
-----------+------------------------------------------------------------------
LnAss~2009 |
         L |      0.099      0.026     3.744   0.000        0.047       0.151
-----------+------------------------------------------------------------------
Home_~2009 |
         L |     -0.010      0.006    -1.618   0.106       -0.022       0.002
-----------+------------------------------------------------------------------
OO_wo~2009 |
         L |     -0.200      0.216    -0.927   0.354       -0.624       0.224
-----------+------------------------------------------------------------------
OO_D_~2009 |
         L |     -0.005      0.006    -0.772   0.440       -0.017       0.007
-----------+------------------------------------------------------------------
Have_~2009 |
         L |     -0.005      0.002    -2.023   0.043       -0.010      -0.000
-----------+------------------------------------------------------------------
OO_ra~2009 |
         L |      0.007      0.009     0.810   0.419       -0.010       0.024
-----------+------------------------------------------------------------------
LnAss~2010 |
         L |      0.115      0.025     4.546   0.000        0.065       0.165
-----------+------------------------------------------------------------------
Home_~2010 |
         L |     -0.009      0.006    -1.473   0.141       -0.021       0.003
-----------+------------------------------------------------------------------
OO_wo~2010 |
         L |     -0.203      0.217    -0.931   0.352       -0.629       0.224
```

```
  -----------+------------------------------------------------------------------
  OO_D_~2010 |
           L |     -0.006      0.006   -0.891   0.373       -0.018       0.007
  -----------+------------------------------------------------------------------
  Have_~2010 |
           L |     -0.004      0.003   -1.516   0.129       -0.009       0.001
  -----------+------------------------------------------------------------------
  OO_ra~2010 |
           L |      0.006      0.008    0.745   0.457       -0.010       0.023
  -----------+------------------------------------------------------------------
  LnAss~2011 |
           L |      0.107      0.030    3.581   0.000        0.048       0.166
  -----------+------------------------------------------------------------------
  Home_~2011 |
           L |     -0.009      0.006   -1.399   0.162       -0.021       0.003
  -----------+------------------------------------------------------------------
  OO_wo~2011 |
           L |     -0.203      0.219   -0.929   0.353       -0.632       0.226
  -----------+------------------------------------------------------------------
  OO_D_~2011 |
           L |     -0.005      0.006   -0.766   0.444       -0.017       0.008
  -----------+------------------------------------------------------------------
  Have_~2011 |
           L |     -0.005      0.002   -2.202   0.028       -0.010      -0.001
  -----------+------------------------------------------------------------------
  OO_ra~2011 |
           L |      0.007      0.008    0.777   0.438       -0.010       0.023
------------------------------------------------------------------------------

mi svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)

  mi estimate ,cmdok post:svy:  sem        ///
(      tdca2004     <-     L@1     _cons@c     LnAssets2004 @b1
       Home_Based2004       @b2          OO_work_exp_owner2004@b3
       OO_D_education_owner2004 @b6      Have_IP2004                          @b5
       OO_race_white_owner2004 @b7        ) ///
(      tdca2005     <-     L@1     _cons@c     LnAssets2005 @b1
       Home_Based2005       @b2          OO_work_exp_owner2005@b3
       OO_D_education_owner2005 @b6      Have_IP2005                          @b5
       OO_race_white_owner2005 @b7        ) ///
(      tdca2006     <-     L@1     _cons@c     LnAssets2006 @b1
       Home_Based2006       @b2          OO_work_exp_owner2006@b3
       OO_D_education_owner2006 @b6      Have_IP2006                          @b5
       OO_race_white_owner2006 @b7        ) ///
(      tdca2007     <-     L@1     _cons@c     LnAssets2007 @b1
       Home_Based2007       @b2          OO_work_exp_owner2007@b3
       OO_D_education_owner2007 @b6      Have_IP2007                          @b5
       OO_race_white_owner2007 @b7        ) ///
(      tdca2008     <-     L@1     _cons@c     LnAssets2008 @b1
       Home_Based2008       @b2          OO_work_exp_owner2008@b3
       OO_D_education_owner2008 @b6      Have_IP2008                          @b5
       OO_race_white_owner2008 @b7        ) ///
(      tdca2009     <-     L@1     _cons@c     LnAssets2009 @b1
       Home_Based2009       @b2          OO_work_exp_owner2009@b3
       OO_D_education_owner2009 @b6      Have_IP2009                          @b5
       OO_race_white_owner2009 @b7        ) ///
(      tdca2010     <-     L@1     _cons@c     LnAssets2010 @b1
       Home_Based2010       @b2          OO_work_exp_owner2010@b3
       OO_D_education_owner2010 @b6      Have_IP2010                          @b5
       OO_race_white_owner2010 @b7        ) ///
(      tdca2011     <-     L@1     _cons@c     LnAssets2011 @b1
       Home_Based2011       @b2          OO_work_exp_owner2011@b3
```

```
       OO_D_education_owner2011 @b6     Have_IP2011                          @b5
       OO_race_white_owner2011 @b7       ) ///
  ,latent(L )   cov(e._OEn@S_e )   nocapslatent       notable
 mi estimate,    cformat(%6.3f)  sformat(%6.3f)   nolstretch

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  td~2004 <- |
  LnAss~2004 |      0.006      0.002    3.877   0.000        0.003       0.009
  Home_~2004 |     -0.017      0.024   -0.709   0.478       -0.063       0.030
  OO_wo~2004 |      0.001      0.002    0.402   0.688       -0.003       0.005
  OO_D_~2004 |     -0.017      0.024   -0.717   0.473       -0.064       0.030
  Have_~2004 |      0.015      0.013    1.153   0.249       -0.010       0.039
  OO_ra~2004 |     -0.041      0.067   -0.618   0.538       -0.173       0.091
           L |      1.000      0.000  7.3e+15   0.000        1.000       1.000
       _cons |      0.217      0.067    3.231   0.002        0.084       0.351
  -----------+----------------------------------------------------------------
  td~2005 <- |
  LnAss~2005 |      0.006      0.002    3.877   0.000        0.003       0.009
  Home_~2005 |     -0.017      0.024   -0.709   0.478       -0.063       0.030
  OO_wo~2005 |      0.001      0.002    0.402   0.688       -0.003       0.005
  OO_D_~2005 |     -0.017      0.024   -0.717   0.473       -0.064       0.030
  Have_~2005 |      0.015      0.013    1.153   0.249       -0.010       0.039
  OO_ra~2005 |     -0.041      0.067   -0.618   0.538       -0.173       0.091
           L |      1.000  (constrained)
       _cons |      0.217      0.067    3.231   0.002        0.084       0.351
  -----------+----------------------------------------------------------------
  td~2006 <- |
  LnAss~2006 |      0.006      0.002    3.877   0.000        0.003       0.009
  Home_~2006 |     -0.017      0.024   -0.709   0.478       -0.063       0.030
  OO_wo~2006 |      0.001      0.002    0.402   0.688       -0.003       0.005
  OO_D_~2006 |     -0.017      0.024   -0.717   0.473       -0.064       0.030
  Have_~2006 |      0.015      0.013    1.153   0.249       -0.010       0.039
  OO_ra~2006 |     -0.041      0.067   -0.618   0.538       -0.173       0.091
           L |      1.000      0.000  1.6e+42   0.000        1.000       1.000
       _cons |      0.217      0.067    3.231   0.002        0.084       0.351
  -----------+----------------------------------------------------------------
  td~2007 <- |
  LnAss~2007 |      0.006      0.002    3.877   0.000        0.003       0.009
  Home_~2007 |     -0.017      0.024   -0.709   0.478       -0.063       0.030
  OO_wo~2007 |      0.001      0.002    0.402   0.688       -0.003       0.005
  OO_D_~2007 |     -0.017      0.024   -0.717   0.473       -0.064       0.030
  Have_~2007 |      0.015      0.013    1.153   0.249       -0.010       0.039
  OO_ra~2007 |     -0.041      0.067   -0.618   0.538       -0.173       0.091
           L |      1.000      0.000  7.4e+15   0.000        1.000       1.000
       _cons |      0.217      0.067    3.231   0.002        0.084       0.351
  -----------+----------------------------------------------------------------
  td~2008 <- |
  LnAss~2008 |      0.006      0.002    3.877   0.000        0.003       0.009
  Home_~2008 |     -0.017      0.024   -0.709   0.478       -0.063       0.030
  OO_wo~2008 |      0.001      0.002    0.402   0.688       -0.003       0.005
  OO_D_~2008 |     -0.017      0.024   -0.717   0.473       -0.064       0.030
  Have_~2008 |      0.015      0.013    1.153   0.249       -0.010       0.039
  OO_ra~2008 |     -0.041      0.067   -0.618   0.538       -0.173       0.091
           L |      1.000      0.000  1.4e+43   0.000        1.000       1.000
       _cons |      0.217      0.067    3.231   0.002        0.084       0.351
  -----------+----------------------------------------------------------------
  td~2009 <- |
  LnAss~2009 |      0.006      0.002    3.877   0.000        0.003       0.009
  Home_~2009 |     -0.017      0.024   -0.709   0.478       -0.063       0.030
```

```
  OO_wo~2009 |      0.001      0.002    0.402   0.688       -0.003       0.005
  OO_D_~2009 |     -0.017      0.024   -0.717   0.473       -0.064       0.030
  Have_~2009 |      0.015      0.013    1.153   0.249       -0.010       0.039
  OO_ra~2009 |     -0.041      0.067   -0.618   0.538       -0.173       0.091
           L |      1.000      0.000  2.3e+42   0.000        1.000       1.000
       _cons |      0.217      0.067    3.231   0.002        0.084       0.351
  -----------+------------------------------------------------------------------
  td~2010 <- |
  LnAss~2010 |      0.006      0.002    3.877   0.000        0.003       0.009
  Home_~2010 |     -0.017      0.024   -0.709   0.478       -0.063       0.030
  OO_wo~2010 |      0.001      0.002    0.402   0.688       -0.003       0.005
  OO_D_~2010 |     -0.017      0.024   -0.717   0.473       -0.064       0.030
  Have_~2010 |      0.015      0.013    1.153   0.249       -0.010       0.039
  OO_ra~2010 |     -0.041      0.067   -0.618   0.538       -0.173       0.091
           L |      1.000      0.000  1.1e+43   0.000        1.000       1.000
       _cons |      0.217      0.067    3.231   0.002        0.084       0.351
  -----------+------------------------------------------------------------------
  td~2011 <- |
  LnAss~2011 |      0.006      0.002    3.877   0.000        0.003       0.009
  Home_~2011 |     -0.017      0.024   -0.709   0.478       -0.063       0.030
  OO_wo~2011 |      0.001      0.002    0.402   0.688       -0.003       0.005
  OO_D_~2011 |     -0.017      0.024   -0.717   0.473       -0.064       0.030
  Have_~2011 |      0.015      0.013    1.153   0.249       -0.010       0.039
  OO_ra~2011 |     -0.041      0.067   -0.618   0.538       -0.173       0.091
           L |      1.000  (constrained)
       _cons |      0.217      0.067    3.231   0.002        0.084       0.351
-------------+------------------------------------------------------------------
Variance     |
  e.tdca2004 |      0.073      0.002                         0.069       0.077
  e.tdca2005 |      0.073      0.002                         0.069       0.077
  e.tdca2006 |      0.073      0.002                         0.069       0.077
  e.tdca2007 |      0.073      0.002                         0.069       0.077
  e.tdca2008 |      0.073      0.002                         0.069       0.077
  e.tdca2009 |      0.073      0.002                         0.069       0.077
  e.tdca2010 |      0.073      0.002                         0.069       0.077
  e.tdca2011 |      0.073      0.002                         0.069       0.077
           L |      0.037      0.003                         0.032       0.042
-------------+------------------------------------------------------------------
Covariance   |
  LnAss~2004 |
           L |      0.091      0.023    4.001   0.000        0.046       0.136
  -----------+------------------------------------------------------------------
  Home_~2004 |
           L |     -0.005      0.006   -0.913   0.361       -0.017       0.006
  -----------+------------------------------------------------------------------
  OO_wo~2004 |
           L |     -0.209      0.220   -0.951   0.342       -0.640       0.222
  -----------+------------------------------------------------------------------
  OO_D_~2004 |
           L |     -0.004      0.006   -0.610   0.542       -0.016       0.009
  -----------+------------------------------------------------------------------
  Have_~2004 |
           L |     -0.003      0.003   -1.268   0.205       -0.008       0.002
  -----------+------------------------------------------------------------------
  OO_ra~2004 |
           L |      0.007      0.008    0.786   0.433       -0.010       0.023
  -----------+------------------------------------------------------------------
  LnAss~2005 |
           L |      0.116      0.021    5.521   0.000        0.075       0.158
  -----------+------------------------------------------------------------------
  Home_~2005 |
           L |     -0.007      0.006   -1.219   0.223       -0.019       0.004
```

```
-----------+------------------------------------------------------------------
OO_wo~2005 |
         L |     -0.219      0.217   -1.011   0.312       -0.644       0.206
-----------+------------------------------------------------------------------
OO_D_~2005 |
         L |     -0.004      0.006   -0.684   0.494       -0.017       0.008
-----------+------------------------------------------------------------------
Have_~2005 |
         L |      0.001      0.003    0.305   0.760       -0.005       0.006
-----------+------------------------------------------------------------------
OO_ra~2005 |
         L |      0.006      0.008    0.756   0.451       -0.010       0.023
-----------+------------------------------------------------------------------
LnAss~2006 |
         L |      0.124      0.024    5.161   0.000        0.077       0.171
-----------+------------------------------------------------------------------
Home_~2006 |
         L |     -0.006      0.006   -1.046   0.296       -0.018       0.006
-----------+------------------------------------------------------------------
OO_wo~2006 |
         L |     -0.201      0.219   -0.916   0.360       -0.630       0.229
-----------+------------------------------------------------------------------
OO_D_~2006 |
         L |     -0.004      0.006   -0.581   0.561       -0.016       0.009
-----------+------------------------------------------------------------------
Have_~2006 |
         L |     -0.005      0.003   -1.901   0.058       -0.011       0.000
-----------+------------------------------------------------------------------
OO_ra~2006 |
         L |      0.006      0.008    0.775   0.440       -0.010       0.023
-----------+------------------------------------------------------------------
LnAss~2007 |
         L |      0.101      0.024    4.253   0.000        0.055       0.148
-----------+------------------------------------------------------------------
Home_~2007 |
         L |     -0.008      0.006   -1.341   0.180       -0.020       0.004
-----------+------------------------------------------------------------------
OO_wo~2007 |
         L |     -0.191      0.219   -0.874   0.382       -0.621       0.238
-----------+------------------------------------------------------------------
OO_D_~2007 |
         L |     -0.004      0.006   -0.711   0.477       -0.017       0.008
-----------+------------------------------------------------------------------
Have_~2007 |
         L |     -0.004      0.003   -1.527   0.127       -0.009       0.001
-----------+------------------------------------------------------------------
OO_ra~2007 |
         L |      0.007      0.009    0.825   0.411       -0.010       0.024
-----------+------------------------------------------------------------------
LnAss~2008 |
         L |      0.101      0.024    4.163   0.000        0.053       0.148
-----------+------------------------------------------------------------------
Home_~2008 |
         L |     -0.009      0.006   -1.406   0.160       -0.021       0.003
-----------+------------------------------------------------------------------
OO_wo~2008 |
         L |     -0.202      0.219   -0.923   0.356       -0.632       0.228
-----------+------------------------------------------------------------------
OO_D_~2008 |
         L |     -0.006      0.006   -0.997   0.319       -0.018       0.006
-----------+------------------------------------------------------------------
Have_~2008 |
```

```
           L |     -0.003      0.002   -1.110   0.267       -0.008       0.002
-------------+------------------------------------------------------------------
  OO_ra~2008 |
           L |      0.007      0.009    0.822   0.412       -0.010       0.024
-------------+------------------------------------------------------------------
  LnAss~2009 |
           L |      0.099      0.026    3.742   0.000        0.047       0.151
-------------+------------------------------------------------------------------
  Home_~2009 |
           L |     -0.010      0.006   -1.617   0.106       -0.022       0.002
-------------+------------------------------------------------------------------
  OO_wo~2009 |
           L |     -0.200      0.216   -0.926   0.355       -0.625       0.224
-------------+------------------------------------------------------------------
  OO_D_~2009 |
           L |     -0.005      0.006   -0.771   0.441       -0.017       0.007
-------------+------------------------------------------------------------------
  Have_~2009 |
           L |     -0.005      0.002   -2.022   0.043       -0.010      -0.000
-------------+------------------------------------------------------------------
  OO_ra~2009 |
           L |      0.007      0.009    0.810   0.419       -0.010       0.024
-------------+------------------------------------------------------------------
  LnAss~2010 |
           L |      0.115      0.025    4.543   0.000        0.065       0.165
-------------+------------------------------------------------------------------
  Home_~2010 |
           L |     -0.009      0.006   -1.473   0.141       -0.021       0.003
-------------+------------------------------------------------------------------
  OO_wo~2010 |
           L |     -0.203      0.218   -0.931   0.352       -0.630       0.225
-------------+------------------------------------------------------------------
  OO_D_~2010 |
           L |     -0.006      0.006   -0.890   0.373       -0.018       0.007
-------------+------------------------------------------------------------------
  Have_~2010 |
           L |     -0.004      0.003   -1.515   0.130       -0.009       0.001
-------------+------------------------------------------------------------------
  OO_ra~2010 |
           L |      0.006      0.008    0.745   0.458       -0.010       0.023
-------------+------------------------------------------------------------------
  LnAss~2011 |
           L |      0.107      0.030    3.581   0.000        0.048       0.166
-------------+------------------------------------------------------------------
  Home_~2011 |
           L |     -0.009      0.006   -1.399   0.162       -0.021       0.003
-------------+------------------------------------------------------------------
  OO_wo~2011 |
           L |     -0.203      0.219   -0.928   0.353       -0.633       0.226
-------------+------------------------------------------------------------------
  OO_D_~2011 |
           L |     -0.005      0.006   -0.765   0.444       -0.017       0.008
-------------+------------------------------------------------------------------
  Have_~2011 |
           L |     -0.005      0.002   -2.204   0.028       -0.010      -0.001
-------------+------------------------------------------------------------------
  OO_ra~2011 |
           L |      0.007      0.008    0.777   0.438       -0.010       0.023
------------------------------------------------------------------------------
```

Examples 6.35 Basic Growth Model

Latent growth modeling is a statistical technique that uses the SEM framework to estimate growth trajectories. A measure advantage of using latent growth models is to investigate systematic change, inter-individual variability in this change, and the correlation of the growth parameters (endowments "initial status" and growth rate) with time varying and nontime varying covariates.

The latent growth curve model is represented by the following set of formulas:
level-1 equation (measurement model):

$$y_{it} = \pi_{0i} + \pi_{1i} Time_t + \epsilon_{it} \quad \text{, for } i = 1,2,..n \text{ and } t = 1,2,..T.$$

where y_{it} is the response variable for firm i at time t. π_{0i} is a latent variable that represents the level-1 intercept (endowments "initial status"), π_{1i} is a latent variable that represents the growth trajectory (growth rate).

More traditionally, the structural model would be represented by

$$\boldsymbol{\Upsilon} = \boldsymbol{\tau_y} + \boldsymbol{\Lambda_y} \boldsymbol{\eta} + \boldsymbol{\varepsilon}$$

$$\begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_T \end{bmatrix} \begin{bmatrix} \pi_{0i} \\ \pi_{1i} \end{bmatrix} + \begin{bmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{iT} \end{bmatrix}$$

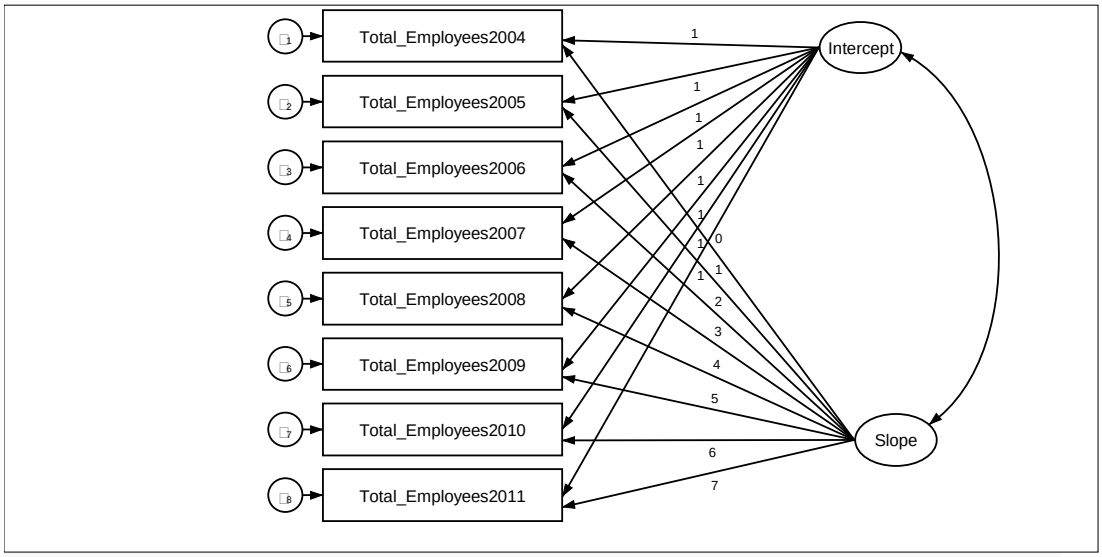
level-2 equations (structural model):

$$\pi_{0i} = \gamma_{00} + \gamma_{01} x_i + \zeta_{0i}$$
$$\pi_{1i} = \gamma_{10} + \gamma_{11} x_i + \zeta_{1i}$$

x_i is a time varying (or nontime varying) predictor(s) of the intercept and (or) slope variables. In the level-2 equations, γ_{00} and γ_{10} are the intercepts or average value of π_{0i} and γ_{10}, respectively, and ζ_{0i} and ζ_{1i} are error terms.

A growth curve requires us to have a model and we should draw this before writing the Stata program; some background in Structural Equation Modeling is assumed.

For illustrative purposes, let us consider studying the employment by start-up firms using KFS data. For simplicity, no level-2 predictors are presented in our model.



- The key variables are the two latent variables labeled the Intercept and the Slope.
- The intercept represents the initial level. We expect substantial variance as some firms have a higher or lower starting number of employees.
- The slope is identified by fixing the values of the paths to each of the Total_Employees variable. We fix the paths at 0, 1, 2, 3, 4, 5, 6, and 7 where the first year is the base year or year zero. We expect there would be substantial variance as some firms will increase (or decrease) their Total_Employees at a different rate than the average growth rate.
- The ε_i terms represent individual error terms for each year.

This is an equivalent to studying growth using a multilevel approach (Random-Coefficient Models) where each firm will have its own intercept and slope (random effects).

```
keep if Duration==8

gen female=1 if PO_gender==0
replace female=0 if PO_gender==1

mi xtset,clear
keep Total_Employees tdca LnAssets Home_Based PO_gender OO_work_exp_owner
Have_IP OO_D_education_owner OO_race_white_owner ///
wgt_7_long mprid year sampleinfo_samplestrata _mi_m _mi_id _mi_miss master
Duration time female
mi reshape wide Total_Employees tdca LnAssets Home_Based PO_gender
OO_work_exp_owner Have_IP OO_D_education_owner OO_race_white_owner time female,
i(mprid) j(year)
mi svyset mprid [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)

 mi svyset mprid [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
 ******************************************************************************
* Basic Growth Model
```

```
*************************************************************************
  mi estimate ,cmdok post:svy:  sem        ///
(      Total_Employees2004 <- Intercept@1 Slope@0       _cons@0)  ///
(      Total_Employees2005 <- Intercept@1 Slope@1       _cons@0)  ///
(      Total_Employees2006 <- Intercept@1 Slope@2       _cons@0)  ///
(      Total_Employees2007 <- Intercept@1 Slope@3       _cons@0)  ///
(      Total_Employees2008 <- Intercept@1 Slope@4       _cons@0)  ///
(      Total_Employees2009 <- Intercept@1 Slope@5       _cons@0)  ///
(      Total_Employees2010 <- Intercept@1 Slope@6       _cons@0)  ///
(      Total_Employees2011 <- Intercept@1 Slope@7       _cons@0)  ///
,var(e._OEn)   latent(Intercept Slope )    nocapslatent   means(Intercept Slope)


mi estimate,    cformat(%6.3f)  sformat(%6.3f)   nolstretch

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Measurement  |
  To~2004 <- |
   Intercept |      1.000  (constrained)
        _cons |      0.000  (constrained)
  -----------+----------------------------------------------------------------
  To~2005 <- |
   Intercept |      1.000  (constrained)
       Slope |      1.000  (constrained)
        _cons |      0.000  (constrained)
  -----------+----------------------------------------------------------------
  To~2006 <- |
   Intercept |      1.000  (constrained)
       Slope |      2.000  (constrained)
        _cons |      0.000  (constrained)
  -----------+----------------------------------------------------------------
  To~2007 <- |
   Intercept |      1.000  (constrained)
       Slope |      3.000  (constrained)
        _cons |      0.000  (constrained)
  -----------+----------------------------------------------------------------
  To~2008 <- |
   Intercept |      1.000  (constrained)
       Slope |      4.000  (constrained)
        _cons |      0.000  (constrained)
  -----------+----------------------------------------------------------------
  To~2009 <- |
   Intercept |      1.000  (constrained)
       Slope |      5.000  (constrained)
        _cons |      0.000  (constrained)
  -----------+----------------------------------------------------------------
  To~2010 <- |
   Intercept |      1.000  (constrained)
       Slope |      6.000  (constrained)
        _cons |      0.000  (constrained)
  -----------+----------------------------------------------------------------
  To~2011 <- |
   Intercept |      1.000  (constrained)
       Slope |      7.000  (constrained)
        _cons |      0.000  (constrained)
-------------+----------------------------------------------------------------
Mean         |
   Intercept |      3.417      0.228   14.987   0.000        2.970       3.864
       Slope |      0.201      0.053    3.784   0.000        0.097       0.305
-------------+----------------------------------------------------------------
```

```
Variance     |
  e.Tot~2004 |     17.890      5.205                      10.109      31.662
  e.Tot~2005 |      5.809      2.624                       2.395      14.091
  e.Tot~2006 |      7.546      2.126                       4.343      13.112
  e.Tot~2007 |     53.710     43.335                      11.035     261.422
  e.Tot~2008 |     17.534      5.576                       9.396      32.719
  e.Tot~2009 |      8.752      3.126                       4.343      17.635
  e.Tot~2010 |     14.056     11.273                       2.915      67.773
  e.Tot~2011 |      9.965      7.145                       2.441      40.672
   Intercept |     34.445      6.926                      23.218      51.100
       Slope |      2.488      0.743                       1.385       4.470
-------------+------------------------------------------------------------------
Covariance   |
  Intercept  |
       Slope |      1.053      1.188    0.887   0.375       -1.276       3.383
------------------------------------------------------------------------------

*Goodness of fit : not suppored by MI, so let us look at each data standalone
       forval i = 1/5 {

       tempname cd
       tempname srmr
             display as txt %33s "  "
   display as txt %33s " coefficient of determination"            ///
       as txt %1s "  "                   ///
       as txt %7s "Standardized root mean squared residual"



cap:     mi xeq `i': svy: sem        ///
(      Total_Employees2004 <- Intercept@1 Slope@0      _cons@0)  ///
(      Total_Employees2005 <- Intercept@1 Slope@1      _cons@0)  ///
(      Total_Employees2006 <- Intercept@1 Slope@2      _cons@0)  ///
(      Total_Employees2007 <- Intercept@1 Slope@3      _cons@0)  ///
(      Total_Employees2008 <- Intercept@1 Slope@4      _cons@0)  ///
(      Total_Employees2009 <- Intercept@1 Slope@5      _cons@0)  ///
(      Total_Employees2010 <- Intercept@1 Slope@6      _cons@0)  ///
(      Total_Employees2011 <- Intercept@1 Slope@7      _cons@0)  ///
,var(e._OEn)   latent(Intercept Slope )    nocapslatent   means(Intercept Slope)

cap:estat gof, stats(all)
             scalar `cd' =   r(cd)
         scalar `srmr' =  r(srmr)
             display         as res %12.4f  `cd' "                        "
as txt %50s  as res %12.4f  `srmr'

       display as txt %33s "  "
       estat eqgof, format(%8.4f)
}

     coefficient of determination  Standardized root mean squared residual
      0.9967                              0.0342
```

```
Equation-level goodness of fit

------------------------------------------------------------------------------
             |             Variance             |
     depvars |    fitted  predicted   residual | R-squared        mc      mc2
-------------+----------------------------------+------------------------------
observed     |                                  |
Total_E~2004 |   51.3895    34.0891    17.3004 |    0.6633    0.8145    0.6633
Total_E~2005 |   44.6503    38.7746     5.8757 |    0.8684    0.9319    0.8684
Total_E~2006 |   56.0967    48.4267     7.6699 |    0.8633    0.9291    0.8633
Total_E~2007 |  116.8241    63.0456    53.7786 |    0.5397    0.7346    0.5397
Total_E~2008 |  100.1972    82.6310    17.5662 |    0.8247    0.9081    0.8247
Total_E~2009 |  115.9322   107.1831     8.7492 |    0.9245    0.9615    0.9245
Total_E~2010 |  150.7732   136.7018    14.0714 |    0.9067    0.9522    0.9067
Total_E~2011 |  181.1782   171.1872     9.9910 |    0.9449    0.9720    0.9449
-------------+----------------------------------+------------------------------
     overall |                                  |    0.9967
------------------------------------------------------------------------------
mc  = correlation between depvar and its prediction
mc2 = mc^2 is the Bentler-Raykov squared multiple correlation coefficient

     coefficient of determination  Standardized root mean squared residual
      0.9967                              0.0343


Equation-level goodness of fit

------------------------------------------------------------------------------
             |             Variance             |
     depvars |    fitted  predicted   residual | R-squared        mc      mc2
-------------+----------------------------------+------------------------------
observed     |                                  |
Total_E~2004 |   51.1655    34.0747    17.0908 |    0.6660    0.8161    0.6660
Total_E~2005 |   44.6013    38.7541     5.8471 |    0.8689    0.9321    0.8689
Total_E~2006 |   56.0761    48.4040     7.6721 |    0.8632    0.9291    0.8632
Total_E~2007 |  116.8267    63.0244    53.8024 |    0.5395    0.7345    0.5395
Total_E~2008 |  100.1701    82.6152    17.5549 |    0.8247    0.9082    0.8247
Total_E~2009 |  115.9327   107.1765     8.7562 |    0.9245    0.9615    0.9245
Total_E~2010 |  150.7945   136.7082    14.0863 |    0.9066    0.9521    0.9066
Total_E~2011 |  181.2068   171.2104     9.9965 |    0.9448    0.9720    0.9448
-------------+----------------------------------+------------------------------
     overall |                                  |    0.9967
------------------------------------------------------------------------------
mc  = correlation between depvar and its prediction
mc2 = mc^2 is the Bentler-Raykov squared multiple correlation coefficient

     coefficient of determination  Standardized root mean squared residual
      0.9967                              0.0339
```

```
Equation-level goodness of fit
------------------------------------------------------------------------------
             |             Variance             |
     depvars |    fitted  predicted   residual | R-squared        mc      mc2
-------------+----------------------------------+-----------------------------
observed     |                                  |
Total_E~2004 |   51.8762    34.3270    17.5492 |    0.6617    0.8135    0.6617
Total_E~2005 |   44.7599    38.9372     5.8226 |    0.8699    0.9327    0.8699
Total_E~2006 |   56.0315    48.5372     7.4943 |    0.8662    0.9307    0.8662
Total_E~2007 |  116.8038    63.1269    53.6769 |    0.5405    0.7352    0.5405
Total_E~2008 |   99.9704    82.7064    17.2639 |    0.8273    0.9096    0.8273
Total_E~2009 |  115.9948   107.2757     8.7191 |    0.9248    0.9617    0.9248
Total_E~2010 |  150.9386   136.8348    14.1038 |    0.9066    0.9521    0.9066
Total_E~2011 |  181.3930   171.3836    10.0094 |    0.9448    0.9720    0.9448
-------------+----------------------------------+-----------------------------
     overall |                                  |    0.9967
------------------------------------------------------------------------------
mc  = correlation between depvar and its prediction
mc2 = mc^2 is the Bentler-Raykov squared multiple correlation coefficient
     coefficient of determination  Standardized root mean squared residual
      0.9967                              0.0341
Equation-level goodness of fit
------------------------------------------------------------------------------
             |             Variance             |
     depvars |    fitted  predicted   residual | R-squared        mc      mc2
-------------+----------------------------------+-----------------------------
observed     |                                  |
Total_E~2004 |   53.3556    34.7459    18.6097 |    0.6512    0.8070    0.6512
Total_E~2005 |   45.0481    39.2759     5.7722 |    0.8719    0.9337    0.8719
Total_E~2006 |   56.2216    48.7839     7.4377 |    0.8677    0.9315    0.8677
Total_E~2007 |  116.8718    63.2699    53.6019 |    0.5414    0.7358    0.5414
Total_E~2008 |  100.2899    82.7339    17.5560 |    0.8249    0.9083    0.8249
Total_E~2009 |  115.9376   107.1760     8.7616 |    0.9244    0.9615    0.9244
Total_E~2010 |  150.6244   136.5960    14.0283 |    0.9069    0.9523    0.9069
Total_E~2011 |  180.9270   170.9941     9.9329 |    0.9451    0.9722    0.9451
-------------+----------------------------------+-----------------------------
     overall |                                  |    0.9967
------------------------------------------------------------------------------
mc  = correlation between depvar and its prediction
mc2 = mc^2 is the Bentler-Raykov squared multiple correlation coefficient

     coefficient of determination  Standardized root mean squared residual
      0.9968                              0.0336
 Equation-level goodness of fit
------------------------------------------------------------------------------
             |             Variance             |
     depvars |    fitted  predicted   residual | R-squared        mc      mc2
-------------+----------------------------------+-----------------------------
observed     |                                  |
Total_E~2004 |   53.8903    34.9889    18.9013 |    0.6493    0.8058    0.6493
Total_E~2005 |   45.1873    39.4595     5.7278 |    0.8732    0.9345    0.8732
Total_E~2006 |   56.3651    48.9089     7.4562 |    0.8677    0.9315    0.8677
Total_E~2007 |  117.0262    63.3371    53.6891 |    0.5412    0.7357    0.5412
Total_E~2008 |  100.4732    82.7441    17.7290 |    0.8235    0.9075    0.8235
Total_E~2009 |  115.9034   107.1300     8.7733 |    0.9243    0.9614    0.9243
Total_E~2010 |  150.4831   136.4947    13.9884 |    0.9070    0.9524    0.9070
Total_E~2011 |  180.7318   170.8383     9.8935 |    0.9453    0.9722    0.9453
-------------+----------------------------------+-----------------------------
     overall |                                  |    0.9968
------------------------------------------------------------------------------
mc  = correlation between depvar and its prediction
mc2 = mc^2 is the Bentler-Raykov squared multiple correlation coefficient
```

```
/*
A perfect fit corresponds to an SRMR of 0.  A good fit is a small value, considered
by some to be limited to 0.08.  SRMR is calculated using the first and second
moments unless sem option nomeans was specified or implied, in which case
SRMR is calculated based on second moments only.  Some software packages ignore
the first moments even when available.  See Hancock and Mueller (2006, 157).
For CD, a perfect fit corresponds to a CD of 1.  CD is like R-squared for the
whole model.
*/


*************************************************************************
* Quadratic
*************************************************************************


  mi estimate ,cmdok post:svy:  sem        ///
(      Total_Employees2004 <- Intercept@1 Slope@0     Quadratic@0  _cons@0)  ///
(      Total_Employees2005 <- Intercept@1 Slope@1     Quadratic@1  _cons@0)  ///
(      Total_Employees2006 <- Intercept@1 Slope@2     Quadratic@4  _cons@0)  ///
(      Total_Employees2007 <- Intercept@1 Slope@3     Quadratic@9  _cons@0)  ///
(      Total_Employees2008 <- Intercept@1 Slope@4     Quadratic@16  _cons@0)  ///
(      Total_Employees2009 <- Intercept@1 Slope@5     Quadratic@25  _cons@0)  ///
(      Total_Employees2010 <- Intercept@1 Slope@6     Quadratic@36  _cons@0)  ///
(      Total_Employees2011 <- Intercept@1 Slope@7     Quadratic@49  _cons@0)  ///
,var(e._OEn)   latent(Intercept Slope Quadratic)    nocapslatent   means(Intercept
Slope Quadratic)


mi estimate,    cformat(%6.3f)  sformat(%6.3f)   nolstretch

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Measurement  |
  To~2004 <- |
   Intercept |      1.000  (constrained)
       _cons |      0.000  (constrained)
  -----------+----------------------------------------------------------------
  To~2005 <- |
   Intercept |      1.000  (constrained)
       Slope |      1.000  (constrained)
   Quadratic |      1.000  (constrained)
       _cons |      0.000  (constrained)
  -----------+----------------------------------------------------------------
  To~2006 <- |
   Intercept |      1.000  (constrained)
       Slope |      2.000  (constrained)
   Quadratic |      4.000  (constrained)
       _cons |      0.000  (constrained)
  -----------+----------------------------------------------------------------
  To~2007 <- |
   Intercept |      1.000  (constrained)
       Slope |      3.000  (constrained)
   Quadratic |      9.000  (constrained)
       _cons |      0.000  (constrained)
  -----------+----------------------------------------------------------------
  To~2008 <- |
   Intercept |      1.000  (constrained)
       Slope |      4.000  (constrained)
   Quadratic |     16.000  (constrained)
       _cons |      0.000  (constrained)
```

```
  -----------+------------------------------------------------------------------
  To~2009 <- |
   Intercept |      1.000  (constrained)
       Slope |      5.000  (constrained)
   Quadratic |     25.000  (constrained)
       _cons |      0.000  (constrained)
  -----------+------------------------------------------------------------------
  To~2010 <- |
   Intercept |      1.000  (constrained)
       Slope |      6.000  (constrained)
   Quadratic |     36.000  (constrained)
       _cons |      0.000  (constrained)
  -----------+------------------------------------------------------------------
  To~2011 <- |
   Intercept |      1.000  (constrained)
       Slope |      7.000  (constrained)
   Quadratic |     49.000  (constrained)
       _cons |      0.000  (constrained)
-------------+------------------------------------------------------------------
Mean         |
   Intercept |      2.798      0.234   11.982   0.000        2.340       3.256
       Slope |      0.672      0.107    6.261   0.000        0.462       0.883
   Quadratic |     -0.059      0.011   -5.279   0.000       -0.081      -0.037
-------------+------------------------------------------------------------------
Variance     |
  e.Tot~2004 |      8.353      3.340                        3.807      18.324
  e.Tot~2005 |      8.171      3.366                        3.642      18.333
  e.Tot~2006 |      7.119      2.488                        3.586      14.129
  e.Tot~2007 |     48.138     42.856                        8.397     275.971
  e.Tot~2008 |     12.202      4.409                        6.006      24.787
  e.Tot~2009 |      7.377      2.198                        4.112      13.235
  e.Tot~2010 |     16.115     13.314                        3.188      81.472
  e.Tot~2011 |      2.510      6.534                        0.015     414.329
   Intercept |     23.783      5.842                       14.690      38.506
       Slope |      7.228      3.491                        2.803      18.642
   Quadratic |      0.089      0.043                        0.035       0.229
-------------+------------------------------------------------------------------
Covariance   |
  Intercept  |
       Slope |      2.734      1.300    2.103   0.036        0.184       5.284
   Quadratic |     -0.172      0.136   -1.267   0.205       -0.439       0.094
  -----------+------------------------------------------------------------------
  Slope      |
   Quadratic |     -0.642      0.344   -1.869   0.062       -1.316       0.032
------------------------------------------------------------------------------
```

```
*Goodness of fit : not suppored by MI, so let us look at each data standalone
       forval i = 1/5 {

       tempname cd
       tempname srmr
             display as txt %33s "  "
   display as txt %33s " coefficient of determination"            ///
       as txt %1s "  "                  ///
       as txt %7s "Standardized root mean squared residual"



cap:     mi xeq `i': svy: sem        ///
(      Total_Employees2004 <- Intercept@1 Slope@0     Quadratic@0  _cons@0)  ///
(      Total_Employees2005 <- Intercept@1 Slope@1     Quadratic@1  _cons@0)  ///
(      Total_Employees2006 <- Intercept@1 Slope@2     Quadratic@4  _cons@0)  ///
(      Total_Employees2007 <- Intercept@1 Slope@3     Quadratic@9  _cons@0)  ///
(      Total_Employees2008 <- Intercept@1 Slope@4     Quadratic@16  _cons@0)  ///
(      Total_Employees2009 <- Intercept@1 Slope@5     Quadratic@25  _cons@0)  ///
(      Total_Employees2010 <- Intercept@1 Slope@6     Quadratic@36  _cons@0)  ///
(      Total_Employees2011 <- Intercept@1 Slope@7     Quadratic@49  _cons@0)  ///
,var(e._OEn)   latent(Intercept Slope Quadratic)    nocapslatent   means(Intercept
Slope Quadratic)

cap:estat gof, stats(all)
             scalar `cd' =   r(cd)
        scalar `srmr' =  r(srmr)
             display         as res %12.4f  `cd' "                      "
as txt %50s  as res %12.4f  `srmr'

       display as txt %33s "  "
       estat eqgof, format(%8.4f)
}


     coefficient of determination  Standardized root mean squared residual
      0.9995                              0.0272


Equation-level goodness of fit

------------------------------------------------------------------------------
             |             Variance             |
     depvars |    fitted  predicted   residual | R-squared        mc      mc2
-------------+----------------------------------+-----------------------------
observed     |                                  |
Total_E~2004 |   31.1802    23.3939     7.7863 |    0.7503    0.8662    0.7503
Total_E~2005 |   43.0014    34.7116     8.2898 |    0.8072    0.8985    0.8072
Total_E~2006 |   60.5453    53.3174     7.2279 |    0.8806    0.9384    0.8806
Total_E~2007 |  122.8632    74.7002    48.1629 |    0.6080    0.7797    0.6080
Total_E~2008 |  108.7005    96.4909    12.2097 |    0.8877    0.9422    0.8877
Total_E~2009 |  125.8349   118.4614     7.3735 |    0.9414    0.9703    0.9414
Total_E~2010 |  158.6474   142.5254    16.1220 |    0.8984    0.9478    0.8984
Total_E~2011 |  175.2757   172.7380     2.5377 |    0.9855    0.9927    0.9855
-------------+----------------------------------+-----------------------------
     overall |                                  |    0.9995
------------------------------------------------------------------------------
mc  = correlation between depvar and its prediction
mc2 = mc^2 is the Bentler-Raykov squared multiple correlation coefficient

     coefficient of determination  Standardized root mean squared residual
      0.9995                              0.0273
```

```
Equation-level goodness of fit

------------------------------------------------------------------------------
             |             Variance             |
     depvars |    fitted  predicted   residual | R-squared        mc      mc2
-------------+----------------------------------+------------------------------
observed     |                                  |
Total_E~2004 |   30.9402    23.3515     7.5886 |    0.7547    0.8688    0.7547
Total_E~2005 |   42.9865    34.7079     8.2786 |    0.8074    0.8986    0.8074
Total_E~2006 |   60.5510    53.3151     7.2359 |    0.8805    0.9383    0.8805
Total_E~2007 |  122.8471    74.6839    48.1632 |    0.6079    0.7797    0.6079
Total_E~2008 |  108.6809    96.4602    12.2207 |    0.8876    0.9421    0.8876
Total_E~2009 |  125.8215   118.4250     7.3965 |    0.9412    0.9702    0.9412
Total_E~2010 |  158.6219   142.4942    16.1276 |    0.8983    0.9478    0.8983
Total_E~2011 |  175.2757   172.7191     2.5566 |    0.9854    0.9927    0.9854
-------------+----------------------------------+------------------------------
     overall |                                  |    0.9995
------------------------------------------------------------------------------
mc  = correlation between depvar and its prediction
mc2 = mc^2 is the Bentler-Raykov squared multiple correlation coefficient

     coefficient of determination  Standardized root mean squared residual
      0.9995                              0.0272


Equation-level goodness of fit

------------------------------------------------------------------------------
             |             Variance             |
     depvars |    fitted  predicted   residual | R-squared        mc      mc2
-------------+----------------------------------+------------------------------
observed     |                                  |
Total_E~2004 |   31.3741    23.4543     7.9198 |    0.7476    0.8646    0.7476
Total_E~2005 |   43.0138    34.7573     8.2565 |    0.8081    0.8989    0.8081
Total_E~2006 |   60.5120    53.4009     7.1110 |    0.8825    0.9394    0.8825
Total_E~2007 |  122.9619    74.8396    48.1223 |    0.6086    0.7802    0.6086
Total_E~2008 |  108.6128    96.6687    11.9442 |    0.8900    0.9434    0.8900
Total_E~2009 |  125.9503   118.6248     7.3256 |    0.9418    0.9705    0.9418
Total_E~2010 |  158.7005   142.5855    16.1150 |    0.8985    0.9479    0.8985
Total_E~2011 |  175.1929   172.5695     2.6234 |    0.9850    0.9925    0.9850
-------------+----------------------------------+------------------------------
     overall |                                  |    0.9995
------------------------------------------------------------------------------
mc  = correlation between depvar and its prediction
mc2 = mc^2 is the Bentler-Raykov squared multiple correlation coefficient

     coefficient of determination  Standardized root mean squared residual
      0.9995                              0.0271
```

Equation-level goodness of fit

depvars	Variance fitted	Variance predicted	Variance residual	R-squared	mc	mc2
observed						
Total_E~2004	32.9936	23.9761	9.0175	0.7267	0.8525	0.7267
Total_E~2005	43.1908	35.0935	8.0973	0.8125	0.9014	0.8125
Total_E~2006	60.4862	53.4678	7.0184	0.8840	0.9402	0.8840
Total_E~2007	122.7121	74.6480	48.0641	0.6083	0.7799	0.6083
Total_E~2008	108.5852	96.3119	12.2734	0.8870	0.9418	0.8870
Total_E~2009	125.6850	118.2666	7.4184	0.9410	0.9700	0.9410
Total_E~2010	158.5490	142.4483	16.1006	0.8985	0.9479	0.8985
Total_E~2011	175.3849	172.9226	2.4623	0.9860	0.9930	0.9860
overall				0.9995		

mc = correlation between depvar and its prediction
mc2 = mc^2 is the Bentler-Raykov squared multiple correlation coefficient

coefficient of determination	Standardized root mean squared residual
0.9995	0.0266

Equation-level goodness of fit

depvars	Variance fitted	Variance predicted	Variance residual	R-squared	mc	mc2
observed						
Total_E~2004	34.1903	24.7395	9.4508	0.7236	0.8506	0.7236
Total_E~2005	43.3597	35.4277	7.9320	0.8171	0.9039	0.8171
Total_E~2006	60.5338	53.5345	6.9993	0.8844	0.9404	0.8844
Total_E~2007	122.7465	74.5694	48.1771	0.6075	0.7794	0.6075
Total_E~2008	108.5423	96.1827	12.3596	0.8861	0.9413	0.8861
Total_E~2009	125.5378	118.1653	7.3724	0.9413	0.9702	0.9413
Total_E~2010	158.5598	142.4494	16.1104	0.8984	0.9478	0.8984
Total_E~2011	175.4769	173.1077	2.3693	0.9865	0.9932	0.9865
overall				0.9995		

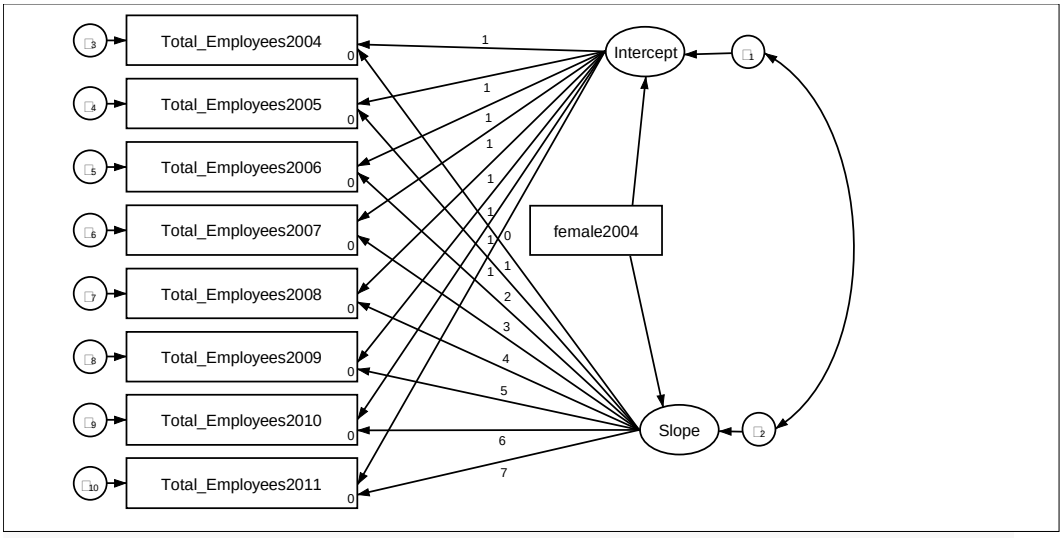
mc = correlation between depvar and its prediction
mc2 = mc^2 is the Bentler-Raykov squared multiple correlation coefficient

$$y_{it} = \mathbf{3.417} + \mathbf{0.228}\, Time_t$$

$$y_{it} = \mathbf{2.798} + \mathbf{0.672} Time_t - \mathbf{0.059}\, Time_t^2$$

Examples 6.36 Basic Growth Model with Time Invariant Covariate

$$y_{it} = \pi_{0i} + \pi_{1i} Time_t + \epsilon_{it}$$
$$\pi_{0i} = \gamma_{00} + \gamma_{01} Female + \zeta_{0i}$$
$$\pi_{1i} = \gamma_{10} + \gamma_{11} Female + \zeta_{1i}$$




```
  mi estimate ,cmdok post:svy:  sem        ///
(      Total_Employees2004 <- Intercept@1 Slope@0       _cons@0)  ///
(      Total_Employees2005 <- Intercept@1 Slope@1       _cons@0)  ///
(      Total_Employees2006 <- Intercept@1 Slope@2       _cons@0)  ///
(      Total_Employees2007 <- Intercept@1 Slope@3       _cons@0)  ///
(      Total_Employees2008 <- Intercept@1 Slope@4       _cons@0)  ///
(      Total_Employees2009 <- Intercept@1 Slope@5       _cons@0)  ///
(      Total_Employees2010 <- Intercept@1 Slope@6       _cons@0)  ///
(      Total_Employees2011 <- Intercept@1 Slope@7       _cons@0)  ///
    (Intercept Slope <- female2004 _cons) , ///
   cov(e.Intercept*e.Slope)  latent(Intercept Slope )    nocapslatent

mi estimate,    cformat(%6.3f)  sformat(%6.3f)   nolstretch



------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  Inter~t <- |
  female2004 |     -1.219      0.333   -3.659   0.000       -1.872      -0.566
       _cons |      3.761      0.278   13.519   0.000        3.215       4.306
  -----------+----------------------------------------------------------------
  Slope <-   |
  female2004 |     -0.195      0.074   -2.634   0.009       -0.339      -0.050
       _cons |      0.255      0.068    3.769   0.000        0.123       0.388
-------------+----------------------------------------------------------------
Measurement  |
  To~2004 <- |
   Intercept |      1.000  (constrained)
       _cons |      0.000  (constrained)
```

```
-----------+------------------------------------------------------------------
To~2005 <- |
 Intercept |      1.000  (constrained)
     Slope |      1.000  (constrained)
     _cons |      0.000  (constrained)
-----------+------------------------------------------------------------------
To~2006 <- |
 Intercept |      1.000  (constrained)
     Slope |      2.000  (constrained)
     _cons |      0.000  (constrained)
-----------+------------------------------------------------------------------
To~2007 <- |
 Intercept |      1.000  (constrained)
     Slope |      3.000  (constrained)
     _cons |      0.000  (constrained)
-----------+------------------------------------------------------------------
To~2008 <- |
 Intercept |      1.000  (constrained)
     Slope |      4.000  (constrained)
     _cons |      0.000  (constrained)
-----------+------------------------------------------------------------------
To~2009 <- |
 Intercept |      1.000  (constrained)
     Slope |      5.000  (constrained)
     _cons |      0.000  (constrained)
-----------+------------------------------------------------------------------
To~2010 <- |
 Intercept |      1.000  (constrained)
     Slope |      6.000  (constrained)
     _cons |      0.000  (constrained)
-----------+------------------------------------------------------------------
To~2011 <- |
 Intercept |      1.000  (constrained)
     Slope |      7.000  (constrained)
     _cons |      0.000  (constrained)
-----------+------------------------------------------------------------------
Variance   |
  e.Tot~2004 |     17.899      5.202                      10.120      31.658
  e.Tot~2005 |      5.795      2.623                       2.385      14.078
  e.Tot~2006 |      7.552      2.124                       4.349      13.112
  e.Tot~2007 |     53.708     43.329                      11.036     261.372
  e.Tot~2008 |     17.543      5.580                       9.400      32.737
  e.Tot~2009 |      8.757      3.129                       4.344      17.651
  e.Tot~2010 |     14.052     11.267                       2.916      67.726
  e.Tot~2011 |      9.958      7.140                       2.440      40.639
  e.Interc~t |     34.152      6.851                      23.043      50.617
     e.Slope |      2.481      0.739                       1.383       4.450
-----------+------------------------------------------------------------------
Covariance |
  e.Interc~t |
     e.Slope |      1.004      1.178    0.852   0.394      -1.307       3.315
------------------------------------------------------------------------------
```

$$\pi_{0i} = \mathbf{3.761} - \mathbf{1.219}\ Female$$

$$\pi_{1i} = \mathbf{0.255} - \mathbf{0.195}\ Female$$

$$y_{it} = \mathbf{3.761} + \mathbf{0.255}\ Time_t : \text{Male}$$

$$y_{it} = \mathbf{2.542} + \mathbf{0.060}\ Time_t : \text{Female}$$

Female owned start-ups have lower initial employment and lower employment growth rate.

Examples 6.37 Basic Growth Model with Time Invariant and Time Varying Covariates

$$y_{it} = \pi_{0i} + \pi_{1i} Time_t + \pi_{2i} Assets_{it} + \epsilon_{it}$$
$$\pi_{0i} = \gamma_{00} + \gamma_{01} Female + \zeta_{0i}$$
$$\pi_{1i} = \gamma_{10} + \gamma_{11} Female + \zeta_{1i}$$

```
  mi estimate ,cmdok post:svy:  sem        ///
(      Total_Employees2004 <- Intercept@1 Slope@0     LnAssets2004@b      _cons@0)
///
(      Total_Employees2005 <- Intercept@1 Slope@1     LnAssets2005@b      _cons@0)
///
(      Total_Employees2006 <- Intercept@1 Slope@2     LnAssets2006@b      _cons@0)
///
(      Total_Employees2007 <- Intercept@1 Slope@3     LnAssets2007@b      _cons@0)
///
(      Total_Employees2008 <- Intercept@1 Slope@4     LnAssets2008@b      _cons@0)
///
(      Total_Employees2009 <- Intercept@1 Slope@5     LnAssets2009@b      _cons@0)
///
(      Total_Employees2010 <- Intercept@1 Slope@6     LnAssets2010@b      _cons@0)
///
(      Total_Employees2011 <- Intercept@1 Slope@7     LnAssets2011@b      _cons@0)
///
    (Intercept Slope <- female2004 _cons) , ///
   cov(e.Intercept*e.Slope)  latent(Intercept Slope )    nocapslatent

mi estimate,    cformat(%6.3f)  sformat(%6.3f)   nolstretch

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  To~2004 <- |
   Intercept |      1.000  (constrained)
  LnAss~2004 |      0.189      0.031    6.100   0.000        0.128       0.250
       _cons |      0.000  (constrained)
  -----------+----------------------------------------------------------------
  To~2005 <- |
   Intercept |      1.000  (constrained)
       Slope |      1.000  (constrained)
  LnAss~2005 |      0.189      0.031    6.100   0.000        0.128       0.250
       _cons |      0.000  (constrained)
  -----------+----------------------------------------------------------------
  To~2006 <- |
   Intercept |      1.000      0.000  6.4e+49   0.000        1.000       1.000
       Slope |      2.000      0.000  1.9e+50   0.000        2.000       2.000
  LnAss~2006 |      0.189      0.031    6.100   0.000        0.128       0.250
       _cons |      0.000  (constrained)
  -----------+----------------------------------------------------------------
  To~2007 <- |
   Intercept |      1.000      0.000  6.9e+49   0.000        1.000       1.000
       Slope |      3.000      0.000  7.3e+50   0.000        3.000       3.000
  LnAss~2007 |      0.189      0.031    6.100   0.000        0.128       0.250
       _cons |      0.000  (constrained)
  -----------+----------------------------------------------------------------
  To~2008 <- |
   Intercept |      1.000      0.000  7.4e+15   0.000        1.000       1.000
       Slope |      4.000      0.000  3.6e+50   0.000        4.000       4.000
  LnAss~2008 |      0.189      0.031    6.100   0.000        0.128       0.250
       _cons |      0.000  (constrained)
```

```
-----------+------------------------------------------------------------------
To~2009 <- |
 Intercept |      1.000      0.000  1.2e+51   0.000        1.000       1.000
     Slope |      5.000      0.000  6.8e+50   0.000        5.000       5.000
LnAss~2009 |      0.189      0.031    6.100   0.000        0.128       0.250
     _cons |      0.000  (constrained)
-----------+------------------------------------------------------------------
To~2010 <- |
 Intercept |      1.000      0.000  7.6e+49   0.000        1.000       1.000
     Slope |      6.000      0.000  1.7e+51   0.000        6.000       6.000
LnAss~2010 |      0.189      0.031    6.100   0.000        0.128       0.250
     _cons |      0.000  (constrained)
-----------+------------------------------------------------------------------
To~2011 <- |
 Intercept |      1.000  (constrained)
     Slope |      7.000  (constrained)
LnAss~2011 |      0.189      0.031    6.100   0.000        0.128       0.250
     _cons |      0.000  (constrained)
-----------+------------------------------------------------------------------
Inter~t <- |
female2004 |     -1.012      0.329   -3.079   0.002       -1.656      -0.367
     _cons |      1.775      0.430    4.123   0.000        0.931       2.619
-----------+------------------------------------------------------------------
Slope <-   |
female2004 |     -0.199      0.073   -2.705   0.007       -0.343      -0.055
     _cons |      0.256      0.067    3.850   0.000        0.126       0.387
-------------+----------------------------------------------------------------
Variance     |
  e.Tot~2004 |     16.789      5.130                        9.218      30.579
  e.Tot~2005 |      5.984      2.698                        2.471      14.491
  e.Tot~2006 |      7.691      2.175                        4.416      13.395
  e.Tot~2007 |     53.251     42.712                       11.043     256.792
  e.Tot~2008 |     17.504      5.544                        9.405      32.578
  e.Tot~2009 |      8.807      3.142                        4.374      17.732
  e.Tot~2010 |     14.157     11.286                        2.964      67.620
  e.Tot~2011 |     10.048      7.201                        2.463      40.980
  e.Interc~t |     32.150      6.772                       21.270      48.597
     e.Slope |      2.443      0.732                        1.357       4.398
-------------+----------------------------------------------------------------
Covariance   |
  e.Interc~t |
     e.Slope |      0.987      1.150    0.858   0.391       -1.269       3.243
------------------------------------------------------------------------------
```

$$Total\_Employees = \pi_{0i} + \pi_{1i} Time_t + 0.189\ Assets_{it}$$

$$\pi_{0i} = 1.775 - 1.012 Female$$

$$\pi_{1i} = 0.256 - 0.199\ Female$$

The start-up size explains some of the gap in the initial employment by female owned start-ups, but female owned start-ups have lower employment growth rate.

Examples 6.38 Multivariate Regression Using SEM

Unlike multiple regression, in multivariate regression, several dependent variables are jointly regressed on the same independent variables. Being a joint estimator, multivariate regression estimates the between-equation covariance and can allow to test coefficients across equations.

Assume that we have G linear equations:

$$y_{1i} = \mathbf{x_{1i}}\boldsymbol{\beta_1} + \varepsilon_{1i}$$
$$y_{2i} = \mathbf{x_{2i}}\boldsymbol{\beta_2} + \varepsilon_{2i}$$
$$\vdots$$
$$y_{Gi} = \mathbf{x_{Gi}}\boldsymbol{\beta_G} + \varepsilon_{Gi}$$

where $\mathbf{x_g}$ is the $n \times K_g$ matrix of explanatory variables and $\mathbf{x_g}$ is the same for all G.

For G=2 we will have

$$\mathbf{y_1} = \mathbf{x_1}\boldsymbol{\beta_1} + \boldsymbol{\varepsilon_1}$$
$$\mathbf{y_2} = \mathbf{x_2}\boldsymbol{\beta_2} + \boldsymbol{\varepsilon_2}$$

where Var $(\mathbf{y_1})$= $\boldsymbol{\sigma_1^2 I_n}$, Var $(\mathbf{y_2})$= $\boldsymbol{\sigma_2^2 I_n}$ and Cov $(\mathbf{y_1}, \mathbf{y_2})$= $\boldsymbol{\sigma_{12} I_n}$

$$\mathbf{Var}(\boldsymbol{\varepsilon}) = \begin{bmatrix} \boldsymbol{\sigma_1^2 I_n} & \boldsymbol{\sigma_{12} I_n} \\ \boldsymbol{\sigma_{12} I_n} & \boldsymbol{\sigma_2^2 I_n} \end{bmatrix}$$

Running a separate OLS regression for each equation will not count for $\boldsymbol{\sigma_{12}}$.

Multivariate regression is the same as the seemingly unrelated regression, but in multivariate regression, we have the same set of independent variables used in each equation.

```
use Longitudinal_Long_MI_Long_L2,clear
merge m:1 mprid   using Kfs8_enclave_14oct13, keepusing(wgt_ini_0 )
keep if _merge==3
*Declare data to be panel data
mi xtset mprid year
gen LnAssets=ln( Assets+1)
*Declare survey design for dataset
mi svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)
* Debt injections  to capital injections
* Farhat, Cotei 2014, draft
egen capital_injection=rowtotal(Debt  Equity )
gen Debt_inj= Debt / capital_injection
gen Equity_inj= Equity / capital_injection
recode Debt_inj Equity_inj (.=0) if master!=0 & capital_injection==0

egen capital=rowtotal(Equity_AllYrs Debt_Owed )
gen tdca= Debt_Owed / capital
recode tdca (.=.a) if  Debt_Owed==.a & Equity_AllYrs==.a
recode tdca (.=0) if master!=0



gen tdcaBus= Bus_Debt_Owed / capital
recode tdcaBus (.=.a) if  Bus_Debt_Owed==.a & Equity_AllYrs==.a
recode tdcaBus (.=0) if master!=0
```

```
gen tdcaPer= Debt_Owed_Owner_Operators / capital
recode tdcaPer (.=.a) if  Debt_Owed_Owner_Operators==.a & Equity_AllYrs==.a
recode tdcaPer (.=0) if master!=0

mi estimate: svy: mean tdcaBus tdcaPer tdca if year==2004
* Replace few missing in PO by OO
bysort master mprid (year):replace            PO_hours     =
      OO_hours_owner[1]   if     PO_hours      ==.   &      OO_hours_owner[1]
      <.
bysort master mprid (year):replace            PO_work_exp  =
      OO_work_exp_owner[1]      if     PO_work_exp   ==.   &
      OO_work_exp_owner   <.
bysort master mprid (year):replace            PO_age_owner =
      OO_age_owner[1]     if     PO_age_owner  ==.   &      OO_age_owner <.
bysort master mprid (year):replace            PO_emp       =
      round(OO_emp_owner[1] ,1)  if     PO_emp        ==.   &
      round(OO_emp_owner[1] ,1)  <.
bysort master mprid (year):replace            PO_oth_bus_owner    =
      round(OO_oth_bus_owner[1] ,1)     if     PO_oth_bus_owner     ==.   &
      round(OO_oth_bus_owner[1] ,1)     <.
bysort master mprid (year):replace            PO_hisp_origin      =
      round(OO_hisp_origin_owner ,1)    if     PO_hisp_origin       ==.   &
      round(OO_hisp_origin_owner[1] ,1)<.
bysort master mprid (year):replace            PO_race_amind_owner       =
      round(OO_race_amind_owner[1] ,1) if     PO_race_amind_owner        ==.
      &      round(OO_race_amind_owner[1] ,1)  <.
bysort master mprid (year):replace            PO_race_asian_owner       =
      round(OO_race_asian_owner[1] ,1) if     PO_race_asian_owner        ==.
      &      round(OO_race_asian_owner[1] ,1)  <.
bysort master mprid (year):replace            PO_race_black_owner       =
      round(OO_race_black_owner[1] ,1) if     PO_race_black_owner        ==.
      &      round(OO_race_black_owner[1] ,1)  <.
bysort master mprid (year):replace            PO_race_nathaw_owner      =
      round(OO_race_nathaw_owner[1] ,1)if     PO_race_nathaw_owner       ==.
      &      round(OO_race_nathaw_owner[1] ,1)<.
bysort master mprid (year):replace            PO_race_other_owner       =
      round(OO_race_other_owner[1] ,1) if     PO_race_other_owner        ==.
      &      round(OO_race_other_owner[1] ,1)  <.
bysort master mprid (year):replace            PO_race_white_owner       =
      round(OO_race_white_owner[1] ,1) if     PO_race_white_owner        ==.
      &      round(OO_race_white_owner[1] ,1)  <.
bysort master mprid (year):replace            PO_native_born      =
      round(OO_native_born_owner[1] ,1)if     PO_native_born       ==.   &
      round(OO_native_born_owner[1] ,1)<.
bysort master mprid (year):replace            PO_us_cit    =
      round(OO_us_cit_owner[1] ,1)      if     PO_us_cit     ==.   &
      round(OO_us_cit_owner[1] ,1)      <.
bysort master mprid (year):replace            PO_education =
      round(OO_education_owner[1] ,1)   if     PO_education  ==.   &
      round(OO_education_owner[1] ,1)   <.
bysort master mprid (year):replace            PO_gender    =
      round(OO_gender_owner[1],1)       if     PO_gender     ==.   &
      round(OO_gender_owner[1],1)       <.
keep if Duration==8
mi xtset,clear
keep  Total_Employees    tdca    tdcaPer  tdcaBus  LnAssets  Home_Based  PO_gender
OO_work_exp_owner Have_IP OO_D_education_owner OO_race_white_owner ///
wgt_7_long  mprid  year  sampleinfo_samplestrata  _mi_m  _mi_id  _mi_miss  master
Duration
```

```
mi reshape wide  Total_Employees  tdca  tdcaPer tdcaBus   LnAssets Home_Based
PO_gender OO_work_exp_owner Have_IP OO_D_education_owner OO_race_white_owner  ,
i(mprid) j(year)
mi svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)

 mi svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)

 save sem,replace

 use sem,clear

  mi svyset  mprid  [pweight=wgt_7_long] , strata(sampleinfo_samplestrata)

  mi estimate ,cmdok post:   sem        ///
(      tdcaPer2004  <-      LnAssets2004 @e1    Home_Based2004      @e2
PO_gender2004        @e3    OO_work_exp_owner2004      @e4
OO_D_education_owner2004    @e5    OO_race_white_owner2004    @e6
Total_Employees2004  @e7   ) ///
(      tdcaPer2005  <-      LnAssets2005 @e1    Home_Based2005      @e2
PO_gender2004        @e3    OO_work_exp_owner2005      @e4
OO_D_education_owner2005    @e5    OO_race_white_owner2005    @e6
Total_Employees2005  @e7   ) ///
(      tdcaPer2006  <-      LnAssets2006 @e1    Home_Based2006      @e2
PO_gender2004        @e3    OO_work_exp_owner2006      @e4
OO_D_education_owner2006    @e5    OO_race_white_owner2006    @e6
Total_Employees2006  @e7   ) ///
(      tdcaPer2007  <-      LnAssets2007 @e1    Home_Based2007      @e2
PO_gender2004        @e3    OO_work_exp_owner2007      @e4
OO_D_education_owner2007    @e5    OO_race_white_owner2007    @e6
Total_Employees2007  @e7   ) ///
(      tdcaPer2008  <-      LnAssets2008 @e1    Home_Based2008      @e2
PO_gender2004        @e3    OO_work_exp_owner2008      @e4
OO_D_education_owner2008    @e5    OO_race_white_owner2008    @e6
Total_Employees2008  @e7   ) ///
(      tdcaPer2009  <-      LnAssets2009 @e1    Home_Based2009      @e2
PO_gender2004        @e3    OO_work_exp_owner2009      @e4
OO_D_education_owner2009    @e5    OO_race_white_owner2009    @e6
Total_Employees2009  @e7   ) ///
(      tdcaPer2010  <-      LnAssets2010 @e1    Home_Based2010      @e2
PO_gender2004        @e3    OO_work_exp_owner2010      @e4
OO_D_education_owner2010    @e5    OO_race_white_owner2010    @e6
Total_Employees2010  @e7   ) ///
(      tdcaPer2011  <-      LnAssets2011 @e1    Home_Based2011      @e2
PO_gender2004        @e3    OO_work_exp_owner2011      @e4
OO_D_education_owner2011    @e5    OO_race_white_owner2011    @e6
Total_Employees2011  @e7   ) ///
(      tdcaBus2004  <-      LnAssets2004 @b1    Home_Based2004      @b2
PO_gender2004        @b3    OO_work_exp_owner2004      @b4
OO_D_education_owner2004    @b5    OO_race_white_owner2004    @b6
Total_Employees2004  @b7   ) ///
(      tdcaBus2005  <-      LnAssets2005 @b1    Home_Based2005      @b2
PO_gender2004        @b3    OO_work_exp_owner2005      @b4
OO_D_education_owner2005    @b5    OO_race_white_owner2005    @b6
Total_Employees2005  @b7   ) ///
(      tdcaBus2006  <-      LnAssets2006 @b1    Home_Based2006      @b2
PO_gender2004        @b3    OO_work_exp_owner2006      @b4
OO_D_education_owner2006    @b5    OO_race_white_owner2006    @b6
Total_Employees2006  @b7   ) ///
(      tdcaBus2007  <-      LnAssets2007 @b1    Home_Based2007      @b2
PO_gender2004        @b3    OO_work_exp_owner2007      @b4
OO_D_education_owner2007    @b5    OO_race_white_owner2007    @b6
Total_Employees2007  @b7   ) ///
```

```
(      tdcaBus2008  <-      LnAssets2008 @b1    Home_Based2008      @b2
PO_gender2004        @b3    OO_work_exp_owner2008      @b4
OO_D_education_owner2008    @b5    OO_race_white_owner2008    @b6
Total_Employees2008  @b7   ) ///
(      tdcaBus2009  <-      LnAssets2009 @b1    Home_Based2009      @b2
PO_gender2004        @b3    OO_work_exp_owner2009      @b4
OO_D_education_owner2009    @b5    OO_race_white_owner2009    @b6
Total_Employees2009  @b7   ) ///
(      tdcaBus2010  <-      LnAssets2010 @b1    Home_Based2010      @b2
PO_gender2004        @b3    OO_work_exp_owner2010      @b4
OO_D_education_owner2010    @b5    OO_race_white_owner2010    @b6
Total_Employees2010  @b7   ) ///
(      tdcaBus2011  <-      LnAssets2011 @b1    Home_Based2011      @b2
PO_gender2004        @b3    OO_work_exp_owner2011      @b4
OO_D_education_owner2011    @b5    OO_race_white_owner2011    @b6
Total_Employees2011  @b7   ) ///
 [pweight=wgt_7_long] , cov(     e.tdcaPer2004*      e.tdcaBus2004 ///
       e.tdcaPer2005*      e.tdcaBus2005 ///
       e.tdcaPer2006*      e.tdcaBus2006 ///
       e.tdcaPer2007*      e.tdcaBus2007 ///
       e.tdcaPer2008*      e.tdcaBus2008 ///
       e.tdcaPer2009*      e.tdcaBus2009 ///
       e.tdcaPer2010*      e.tdcaBus2010 ///
       e.tdcaPer2011*      e.tdcaBus2011)  ///
       var(e.tdcaPer2004    @v1   e.tdcaBus2004 @v2    ///
e.tdcaPer2005 @v1   e.tdcaBus2005 @v2    ///
e.tdcaPer2006 @v1   e.tdcaBus2006 @v2    ///
e.tdcaPer2007 @v1   e.tdcaBus2007 @v2    ///
e.tdcaPer2008 @v1   e.tdcaBus2008 @v2    ///
e.tdcaPer2009 @v1   e.tdcaBus2009 @v2    ///
e.tdcaPer2010 @v1   e.tdcaBus2010 @v2    ///
e.tdcaPer2011 @v1   e.tdcaBus2011 @v2   )  vce(cluster mprid)  nocapslatent

mi estimate,    cformat(%6.3f)  sformat(%6.3f)   nolstretch
```

```
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  tdcaP~4 <- |
  LnAss~2004 |      0.007      0.001    6.132   0.000        0.004       0.009
  Home_~2004 |      0.002      0.009    0.274   0.784       -0.015       0.020
  PO_ge~2004 |     -0.002      0.010   -0.240   0.810       -0.022       0.017
  OO_wo~2004 |     -0.002      0.000   -3.923   0.000       -0.002      -0.001
  OO_D_~2004 |     -0.028      0.008   -3.379   0.001       -0.045      -0.012
  OO_ra~2004 |      0.002      0.012    0.153   0.879       -0.022       0.026
  Total~2004 |     -0.001      0.000   -3.969   0.000       -0.002      -0.001
       _cons |      0.212      0.019   10.969   0.000        0.174       0.250
  -----------+----------------------------------------------------------------
  tdcaP~5 <- |
  PO_ge~2004 |     -0.002      0.010   -0.240   0.810       -0.022       0.017
  LnAss~2005 |      0.007      0.001    6.132   0.000        0.004       0.009
  Home_~2005 |      0.002      0.009    0.274   0.784       -0.015       0.020
  OO_wo~2005 |     -0.002      0.000   -3.923   0.000       -0.002      -0.001
  OO_D_~2005 |     -0.028      0.008   -3.379   0.001       -0.045      -0.012
  OO_ra~2005 |      0.002      0.012    0.153   0.879       -0.022       0.026
  Total~2005 |     -0.001      0.000   -3.969   0.000       -0.002      -0.001
       _cons |      0.121      0.018    6.698   0.000        0.086       0.157
  -----------+----------------------------------------------------------------
```

```
tdcaP~6 <- |
PO_ge~2004 |     -0.002      0.010   -0.240   0.810       -0.022       0.017
LnAss~2006 |      0.007      0.001    6.132   0.000        0.004       0.009
Home_~2006 |      0.002      0.009    0.274   0.784       -0.015       0.020
OO_wo~2006 |     -0.002      0.000   -3.923   0.000       -0.002      -0.001
OO_D_~2006 |     -0.028      0.008   -3.379   0.001       -0.045      -0.012
OO_ra~2006 |      0.002      0.012    0.153   0.879       -0.022       0.026
Total~2006 |     -0.001      0.000   -3.969   0.000       -0.002      -0.001
     _cons |      0.120      0.018    6.531   0.000        0.084       0.156
-----------+------------------------------------------------------------------
tdcaP~7 <- |
PO_ge~2004 |     -0.002      0.010   -0.240   0.810       -0.022       0.017
LnAss~2007 |      0.007      0.001    6.132   0.000        0.004       0.009
Home_~2007 |      0.002      0.009    0.274   0.784       -0.015       0.020
OO_wo~2007 |     -0.002      0.000   -3.923   0.000       -0.002      -0.001
OO_D_~2007 |     -0.028      0.008   -3.379   0.001       -0.045      -0.012
OO_ra~2007 |      0.002      0.012    0.153   0.879       -0.022       0.026
Total~2007 |     -0.001      0.000   -3.969   0.000       -0.002      -0.001
     _cons |      0.123      0.019    6.641   0.000        0.087       0.160
-----------+------------------------------------------------------------------
tdcaP~8 <- |
PO_ge~2004 |     -0.002      0.010   -0.240   0.810       -0.022       0.017
LnAss~2008 |      0.007      0.001    6.132   0.000        0.004       0.009
Home_~2008 |      0.002      0.009    0.274   0.784       -0.015       0.020
OO_wo~2008 |     -0.002      0.000   -3.923   0.000       -0.002      -0.001
OO_D_~2008 |     -0.028      0.008   -3.379   0.001       -0.045      -0.012
OO_ra~2008 |      0.002      0.012    0.153   0.879       -0.022       0.026
Total~2008 |     -0.001      0.000   -3.969   0.000       -0.002      -0.001
     _cons |      0.111      0.018    6.104   0.000        0.075       0.146
-----------+------------------------------------------------------------------
tdcaP~9 <- |
PO_ge~2004 |     -0.002      0.010   -0.240   0.810       -0.022       0.017
LnAss~2009 |      0.007      0.001    6.132   0.000        0.004       0.009
Home_~2009 |      0.002      0.009    0.274   0.784       -0.015       0.020
OO_wo~2009 |     -0.002      0.000   -3.923   0.000       -0.002      -0.001
OO_D_~2009 |     -0.028      0.008   -3.379   0.001       -0.045      -0.012
OO_ra~2009 |      0.002      0.012    0.153   0.879       -0.022       0.026
Total~2009 |     -0.001      0.000   -3.969   0.000       -0.002      -0.001
     _cons |      0.099      0.019    5.240   0.000        0.062       0.135
-----------+------------------------------------------------------------------
tdcaP~0 <- |
PO_ge~2004 |     -0.002      0.010   -0.240   0.810       -0.022       0.017
LnAss~2010 |      0.007      0.001    6.132   0.000        0.004       0.009
Home_~2010 |      0.002      0.009    0.274   0.784       -0.015       0.020
OO_wo~2010 |     -0.002      0.000   -3.923   0.000       -0.002      -0.001
OO_D_~2010 |     -0.028      0.008   -3.379   0.001       -0.045      -0.012
OO_ra~2010 |      0.002      0.012    0.153   0.879       -0.022       0.026
Total~2010 |     -0.001      0.000   -3.969   0.000       -0.002      -0.001
     _cons |      0.080      0.018    4.356   0.000        0.044       0.116
-----------+------------------------------------------------------------------
tdcaP~1 <- |
PO_ge~2004 |     -0.002      0.010   -0.240   0.810       -0.022       0.017
LnAss~2011 |      0.007      0.001    6.132   0.000        0.004       0.009
Home_~2011 |      0.002      0.009    0.274   0.784       -0.015       0.020
OO_wo~2011 |     -0.002      0.000   -3.923   0.000       -0.002      -0.001
OO_D_~2011 |     -0.028      0.008   -3.379   0.001       -0.045      -0.012
OO_ra~2011 |      0.002      0.012    0.153   0.879       -0.022       0.026
Total~2011 |     -0.001      0.000   -3.969   0.000       -0.002      -0.001
     _cons |      0.061      0.018    3.381   0.001        0.026       0.096
-----------+------------------------------------------------------------------
tdcaB~4 <- |
LnAss~2004 |      0.010      0.001    9.549   0.000        0.008       0.012
```

```
Home_~2004 |     -0.035      0.008   -4.298   0.000       -0.051      -0.019
PO_ge~2004 |      0.011      0.009    1.251   0.211       -0.006       0.029
OO_wo~2004 |      0.000      0.000    0.176   0.860       -0.001       0.001
OO_D_~2004 |     -0.010      0.008   -1.258   0.208       -0.025       0.005
OO_ra~2004 |      0.005      0.010    0.461   0.645       -0.015       0.024
Total~2004 |      0.001      0.001    2.313   0.021        0.000       0.003
     _cons |      0.019      0.015    1.233   0.218       -0.011       0.048
-----------+------------------------------------------------------------------
tdcaB~5 <- |
PO_ge~2004 |      0.011      0.009    1.251   0.211       -0.006       0.029
LnAss~2005 |      0.010      0.001    9.549   0.000        0.008       0.012
Home_~2005 |     -0.035      0.008   -4.298   0.000       -0.051      -0.019
OO_wo~2005 |      0.000      0.000    0.176   0.860       -0.001       0.001
OO_D_~2005 |     -0.010      0.008   -1.258   0.208       -0.025       0.005
OO_ra~2005 |      0.005      0.010    0.461   0.645       -0.015       0.024
Total~2005 |      0.001      0.001    2.313   0.021        0.000       0.003
     _cons |     -0.002      0.015   -0.134   0.894       -0.032       0.028
-----------+------------------------------------------------------------------
tdcaB~6 <- |
PO_ge~2004 |      0.011      0.009    1.251   0.211       -0.006       0.029
LnAss~2006 |      0.010      0.001    9.549   0.000        0.008       0.012
Home_~2006 |     -0.035      0.008   -4.298   0.000       -0.051      -0.019
OO_wo~2006 |      0.000      0.000    0.176   0.860       -0.001       0.001
OO_D_~2006 |     -0.010      0.008   -1.258   0.208       -0.025       0.005
OO_ra~2006 |      0.005      0.010    0.461   0.645       -0.015       0.024
Total~2006 |      0.001      0.001    2.313   0.021        0.000       0.003
     _cons |      0.011      0.015    0.699   0.484       -0.020       0.041
-----------+------------------------------------------------------------------
tdcaB~7 <- |
PO_ge~2004 |      0.011      0.009    1.251   0.211       -0.006       0.029
LnAss~2007 |      0.010      0.001    9.549   0.000        0.008       0.012
Home_~2007 |     -0.035      0.008   -4.298   0.000       -0.051      -0.019
OO_wo~2007 |      0.000      0.000    0.176   0.860       -0.001       0.001
OO_D_~2007 |     -0.010      0.008   -1.258   0.208       -0.025       0.005
OO_ra~2007 |      0.005      0.010    0.461   0.645       -0.015       0.024
Total~2007 |      0.001      0.001    2.313   0.021        0.000       0.003
     _cons |      0.012      0.015    0.772   0.440       -0.018       0.041
-----------+------------------------------------------------------------------
tdcaB~8 <- |
PO_ge~2004 |      0.011      0.009    1.251   0.211       -0.006       0.029
LnAss~2008 |      0.010      0.001    9.549   0.000        0.008       0.012
Home_~2008 |     -0.035      0.008   -4.298   0.000       -0.051      -0.019
OO_wo~2008 |      0.000      0.000    0.176   0.860       -0.001       0.001
OO_D_~2008 |     -0.010      0.008   -1.258   0.208       -0.025       0.005
OO_ra~2008 |      0.005      0.010    0.461   0.645       -0.015       0.024
Total~2008 |      0.001      0.001    2.313   0.021        0.000       0.003
     _cons |      0.012      0.016    0.764   0.445       -0.019       0.043
-----------+------------------------------------------------------------------
tdcaB~9 <- |
PO_ge~2004 |      0.011      0.009    1.251   0.211       -0.006       0.029
LnAss~2009 |      0.010      0.001    9.549   0.000        0.008       0.012
Home_~2009 |     -0.035      0.008   -4.298   0.000       -0.051      -0.019
OO_wo~2009 |      0.000      0.000    0.176   0.860       -0.001       0.001
OO_D_~2009 |     -0.010      0.008   -1.258   0.208       -0.025       0.005
OO_ra~2009 |      0.005      0.010    0.461   0.645       -0.015       0.024
Total~2009 |      0.001      0.001    2.313   0.021        0.000       0.003
     _cons |     -0.006      0.015   -0.427   0.670       -0.035       0.022
-----------+------------------------------------------------------------------
tdcaB~0 <- |
PO_ge~2004 |      0.011      0.009    1.251   0.211       -0.006       0.029
LnAss~2010 |      0.010      0.001    9.549   0.000        0.008       0.012
Home_~2010 |     -0.035      0.008   -4.298   0.000       -0.051      -0.019
```

```
  OO_wo~2010 |      0.000      0.000    0.176   0.860       -0.001       0.001
  OO_D_~2010 |     -0.010      0.008   -1.258   0.208       -0.025       0.005
  OO_ra~2010 |      0.005      0.010    0.461   0.645       -0.015       0.024
  Total~2010 |      0.001      0.001    2.313   0.021        0.000       0.003
       _cons |     -0.010      0.015   -0.661   0.509       -0.039       0.020
  ------------+------------------------------------------------------------------
  tdcaB~1 <- |
  PO_ge~2004 |      0.011      0.009    1.251   0.211       -0.006       0.029
  LnAss~2011 |      0.010      0.001    9.549   0.000        0.008       0.012
  Home_~2011 |     -0.035      0.008   -4.298   0.000       -0.051      -0.019
  OO_wo~2011 |      0.000      0.000    0.176   0.860       -0.001       0.001
  OO_D_~2011 |     -0.010      0.008   -1.258   0.208       -0.025       0.005
  OO_ra~2011 |      0.005      0.010    0.461   0.645       -0.015       0.024
  Total~2011 |      0.001      0.001    2.313   0.021        0.000       0.003
       _cons |     -0.012      0.015   -0.797   0.425       -0.042       0.018
-------------+------------------------------------------------------------------
Variance     |
  e.tdcaPe~4 |      0.061      0.002                        0.057       0.065
  e.tdcaPe~5 |      0.061      0.002                        0.057       0.065
  e.tdcaPe~6 |      0.061      0.002                        0.057       0.065
  e.tdcaPe~7 |      0.061      0.002                        0.057       0.065
  e.tdcaPe~8 |      0.061      0.002                        0.057       0.065
  e.tdcaPe~9 |      0.061      0.002                        0.057       0.065
  e.tdcaPe~0 |      0.061      0.002                        0.057       0.065
  e.tdcaPe~1 |      0.061      0.002                        0.057       0.065
  e.tdcaBu~4 |      0.047      0.002                        0.043       0.052
  e.tdcaBu~5 |      0.047      0.002                        0.043       0.052
  e.tdcaBu~6 |      0.047      0.002                        0.043       0.052
  e.tdcaBu~7 |      0.047      0.002                        0.043       0.052
  e.tdcaBu~8 |      0.047      0.002                        0.043       0.052
  e.tdcaBu~9 |      0.047      0.002                        0.043       0.052
  e.tdcaBu~0 |      0.047      0.002                        0.043       0.052
  e.tdcaBu~1 |      0.047      0.002                        0.043       0.052
-------------+------------------------------------------------------------------
Covariance   |
  e.tdcaPe~4 |
  e.tdcaBu~4 |     -0.004      0.001   -4.092   0.000       -0.006      -0.002
  -----------+------------------------------------------------------------------
  e.tdcaPe~5 |
  e.tdcaBu~5 |      0.000      0.001    0.196   0.845       -0.002       0.003
  -----------+------------------------------------------------------------------
  e.tdcaPe~6 |
  e.tdcaBu~6 |     -0.001      0.001   -1.235   0.217       -0.004       0.001
  -----------+------------------------------------------------------------------
  e.tdcaPe~7 |
  e.tdcaBu~7 |     -0.002      0.001   -2.264   0.024       -0.004      -0.000
  -----------+------------------------------------------------------------------
  e.tdcaPe~8 |
  e.tdcaBu~8 |     -0.001      0.001   -1.174   0.240       -0.004       0.001
  -----------+------------------------------------------------------------------
  e.tdcaPe~9 |
  e.tdcaBu~9 |      0.001      0.001    0.438   0.661       -0.002       0.004
  -----------+------------------------------------------------------------------
  e.tdcaPe~0 |
  e.tdcaBu~0 |      0.001      0.001    0.493   0.622       -0.002       0.004
  -----------+------------------------------------------------------------------
  e.tdcaPe~1 |
  e.tdcaBu~1 |      0.002      0.002    0.968   0.333       -0.002       0.006
------------------------------------------------------------------------------
```

Examples 6.39 Seemingly Unrelated Regressions Using SEM

Assume that we have G linear equations:

$$y_{1i} = \mathbf{x_{1i}}\boldsymbol{\beta_1} + \varepsilon_{1i}$$
$$y_{2i} = \mathbf{x_{2i}}\boldsymbol{\beta_2} + \varepsilon_{2i}$$
$$\vdots$$
$$y_{Gi} = \mathbf{x_{Gi}}\boldsymbol{\beta_G} + \varepsilon_{Gi}$$

where $\mathbf{x_g}$ is the $n \times K_g$ matrix of explanatory variables, the elements and the dimension of $\mathbf{x_g}$ are allowed to vary across equations.

For G=2 we will have

$$\mathbf{y_1} = \mathbf{x_1}\boldsymbol{\beta_1} + \boldsymbol{\varepsilon_1}$$
$$\mathbf{y_2} = \mathbf{x_2}\boldsymbol{\beta_2} + \boldsymbol{\varepsilon_2}$$

where Var $(\mathbf{y_1})$= $\boldsymbol{\sigma_1^2 I_n}$, Var $(\mathbf{y_2})$= $\boldsymbol{\sigma_2^2 I_n}$ and Cov $(\mathbf{y_1}, \mathbf{y_2})$= $\boldsymbol{\sigma_{12} I_n}$

$$\mathbf{Var}(\boldsymbol{\varepsilon}) = \begin{bmatrix} \boldsymbol{\sigma_1^2 I_n} & \boldsymbol{\sigma_{12} I_n} \\ \boldsymbol{\sigma_{12} I_n} & \boldsymbol{\sigma_2^2 I_n} \end{bmatrix}$$

```
  mi estimate ,cmdok post:   sem        ///
(      tdcaPer2004  <-            Home_Based2004      @e2    PO_gender2004
@e3     OO_work_exp_owner2004     @e4    OO_D_education_owner2004   @e5
OO_race_white_owner2004     @e6          ) ///
(      tdcaPer2005  <-            Home_Based2005      @e2    PO_gender2004
@e3     OO_work_exp_owner2005     @e4    OO_D_education_owner2005   @e5
OO_race_white_owner2005     @e6          ) ///
(      tdcaPer2006  <-            Home_Based2006      @e2    PO_gender2004
@e3     OO_work_exp_owner2006     @e4    OO_D_education_owner2006   @e5
OO_race_white_owner2006     @e6          ) ///
(      tdcaPer2007  <-            Home_Based2007      @e2    PO_gender2004
@e3     OO_work_exp_owner2007     @e4    OO_D_education_owner2007   @e5
OO_race_white_owner2007     @e6          ) ///
(      tdcaPer2008  <-            Home_Based2008      @e2    PO_gender2004
@e3     OO_work_exp_owner2008     @e4    OO_D_education_owner2008   @e5
OO_race_white_owner2008     @e6          ) ///
(      tdcaPer2009  <-            Home_Based2009      @e2    PO_gender2004
@e3     OO_work_exp_owner2009     @e4    OO_D_education_owner2009   @e5
OO_race_white_owner2009     @e6          ) ///
(      tdcaPer2010  <-            Home_Based2010      @e2    PO_gender2004
@e3     OO_work_exp_owner2010     @e4    OO_D_education_owner2010   @e5
OO_race_white_owner2010     @e6          ) ///
(      tdcaPer2011  <-            Home_Based2011      @e2    PO_gender2004
@e3     OO_work_exp_owner2011     @e4    OO_D_education_owner2011   @e5
OO_race_white_owner2011     @e6          ) ///
(      tdcaBus2004  <-      LnAssets2004 @b1    Home_Based2004      @b2
PO_gender2004        @b3    OO_work_exp_owner2004      @b4
OO_D_education_owner2004    @b5    OO_race_white_owner2004    @b6
Total_Employees2004  @b7   ) ///
(      tdcaBus2005  <-      LnAssets2005 @b1    Home_Based2005      @b2
PO_gender2004        @b3    OO_work_exp_owner2005      @b4
OO_D_education_owner2005    @b5    OO_race_white_owner2005    @b6
Total_Employees2005  @b7   ) ///
(      tdcaBus2006  <-      LnAssets2006 @b1    Home_Based2006      @b2
PO_gender2004        @b3    OO_work_exp_owner2006      @b4
OO_D_education_owner2006    @b5    OO_race_white_owner2006    @b6
Total_Employees2006  @b7   ) ///
```

```
(      tdcaBus2007  <-      LnAssets2007 @b1    Home_Based2007      @b2
PO_gender2004        @b3    OO_work_exp_owner2007      @b4
OO_D_education_owner2007    @b5    OO_race_white_owner2007    @b6
Total_Employees2007  @b7   ) ///
(      tdcaBus2008  <-      LnAssets2008 @b1    Home_Based2008      @b2
PO_gender2004        @b3    OO_work_exp_owner2008      @b4
OO_D_education_owner2008    @b5    OO_race_white_owner2008    @b6
Total_Employees2008  @b7   ) ///
(      tdcaBus2009  <-      LnAssets2009 @b1    Home_Based2009      @b2
PO_gender2004        @b3    OO_work_exp_owner2009      @b4
OO_D_education_owner2009    @b5    OO_race_white_owner2009    @b6
Total_Employees2009  @b7   ) ///
(      tdcaBus2010  <-      LnAssets2010 @b1    Home_Based2010      @b2
PO_gender2004        @b3    OO_work_exp_owner2010      @b4
OO_D_education_owner2010    @b5    OO_race_white_owner2010    @b6
Total_Employees2010  @b7   ) ///
(      tdcaBus2011  <-      LnAssets2011 @b1    Home_Based2011      @b2
PO_gender2004        @b3    OO_work_exp_owner2011      @b4
OO_D_education_owner2011    @b5    OO_race_white_owner2011    @b6
Total_Employees2011  @b7   ) ///
 [pweight=wgt_7_long] , cov(      e.tdcaPer2004 *      e.tdcaBus2004 ///
       e.tdcaPer2005 *      e.tdcaBus2005 ///
       e.tdcaPer2006 *      e.tdcaBus2006 ///
       e.tdcaPer2007 *      e.tdcaBus2007 ///
       e.tdcaPer2008 *      e.tdcaBus2008 ///
       e.tdcaPer2009 *      e.tdcaBus2009 ///
       e.tdcaPer2010 *      e.tdcaBus2010 ///
       e.tdcaPer2011 *      e.tdcaBus2011 )  ///
       var(e.tdcaPer2004    @v1   e.tdcaBus2004 @v2    ///
e.tdcaPer2005 @v1   e.tdcaBus2005 @v2    ///
e.tdcaPer2006 @v1   e.tdcaBus2006 @v2    ///
e.tdcaPer2007 @v1   e.tdcaBus2007 @v2    ///
e.tdcaPer2008 @v1   e.tdcaBus2008 @v2    ///
e.tdcaPer2009 @v1   e.tdcaBus2009 @v2    ///
e.tdcaPer2010 @v1   e.tdcaBus2010 @v2    ///
e.tdcaPer2011 @v1   e.tdcaBus2011 @v2   )  vce(cluster mprid)  nocapslatent

mi estimate,    cformat(%6.3f)  sformat(%6.3f)   nolstretch

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  tdcaP~4 <- |
  Home_~2004 |     -0.004      0.009   -0.423   0.672       -0.020       0.013
  PO_ge~2004 |      0.002      0.010    0.151   0.880       -0.018       0.021
  OO_wo~2004 |     -0.002      0.000   -3.765   0.000       -0.002      -0.001
  OO_D_~2004 |     -0.027      0.008   -3.223   0.001       -0.044      -0.011
  OO_ra~2004 |      0.009      0.012    0.712   0.477       -0.015       0.033
       _cons |      0.265      0.019   14.269   0.000        0.228       0.301
  -----------+----------------------------------------------------------------
  tdcaP~5 <- |
  PO_ge~2004 |      0.002      0.010    0.151   0.880       -0.018       0.021
  Home_~2005 |     -0.004      0.009   -0.423   0.672       -0.020       0.013
  OO_wo~2005 |     -0.002      0.000   -3.765   0.000       -0.002      -0.001
  OO_D_~2005 |     -0.027      0.008   -3.223   0.001       -0.044      -0.011
  OO_ra~2005 |      0.009      0.012    0.712   0.477       -0.015       0.033
       _cons |      0.178      0.017   10.586   0.000        0.145       0.211
  -----------+----------------------------------------------------------------
  tdcaP~6 <- |
  PO_ge~2004 |      0.002      0.010    0.151   0.880       -0.018       0.021
  Home_~2006 |     -0.004      0.009   -0.423   0.672       -0.020       0.013
```

```
OO_wo~2006 |     -0.002      0.000   -3.765   0.000       -0.002      -0.001
OO_D_~2006 |     -0.027      0.008   -3.223   0.001       -0.044      -0.011
OO_ra~2006 |      0.009      0.012    0.712   0.477       -0.015       0.033
     _cons |      0.175      0.017   10.348   0.000        0.142       0.208
-----------+------------------------------------------------------------------
tdcaP~7 <- |
PO_ge~2004 |      0.002      0.010    0.151   0.880       -0.018       0.021
Home_~2007 |     -0.004      0.009   -0.423   0.672       -0.020       0.013
OO_wo~2007 |     -0.002      0.000   -3.765   0.000       -0.002      -0.001
OO_D_~2007 |     -0.027      0.008   -3.223   0.001       -0.044      -0.011
OO_ra~2007 |      0.009      0.012    0.712   0.477       -0.015       0.033
     _cons |      0.179      0.017   10.285   0.000        0.145       0.213
-----------+------------------------------------------------------------------
tdcaP~8 <- |
PO_ge~2004 |      0.002      0.010    0.151   0.880       -0.018       0.021
Home_~2008 |     -0.004      0.009   -0.423   0.672       -0.020       0.013
OO_wo~2008 |     -0.002      0.000   -3.765   0.000       -0.002      -0.001
OO_D_~2008 |     -0.027      0.008   -3.223   0.001       -0.044      -0.011
OO_ra~2008 |      0.009      0.012    0.712   0.477       -0.015       0.033
     _cons |      0.167      0.016   10.111   0.000        0.134       0.199
-----------+------------------------------------------------------------------
tdcaP~9 <- |
PO_ge~2004 |      0.002      0.010    0.151   0.880       -0.018       0.021
Home_~2009 |     -0.004      0.009   -0.423   0.672       -0.020       0.013
OO_wo~2009 |     -0.002      0.000   -3.765   0.000       -0.002      -0.001
OO_D_~2009 |     -0.027      0.008   -3.223   0.001       -0.044      -0.011
OO_ra~2009 |      0.009      0.012    0.712   0.477       -0.015       0.033
     _cons |      0.154      0.018    8.761   0.000        0.120       0.189
-----------+------------------------------------------------------------------
tdcaP~0 <- |
PO_ge~2004 |      0.002      0.010    0.151   0.880       -0.018       0.021
Home_~2010 |     -0.004      0.009   -0.423   0.672       -0.020       0.013
OO_wo~2010 |     -0.002      0.000   -3.765   0.000       -0.002      -0.001
OO_D_~2010 |     -0.027      0.008   -3.223   0.001       -0.044      -0.011
OO_ra~2010 |      0.009      0.012    0.712   0.477       -0.015       0.033
     _cons |      0.135      0.017    8.129   0.000        0.103       0.168
-----------+------------------------------------------------------------------
tdcaP~1 <- |
PO_ge~2004 |      0.002      0.010    0.151   0.880       -0.018       0.021
Home_~2011 |     -0.004      0.009   -0.423   0.672       -0.020       0.013
OO_wo~2011 |     -0.002      0.000   -3.765   0.000       -0.002      -0.001
OO_D_~2011 |     -0.027      0.008   -3.223   0.001       -0.044      -0.011
OO_ra~2011 |      0.009      0.012    0.712   0.477       -0.015       0.033
     _cons |      0.115      0.017    6.936   0.000        0.082       0.147
-----------+------------------------------------------------------------------
tdcaB~4 <- |
Home_~2004 |     -0.035      0.008   -4.290   0.000       -0.051      -0.019
PO_ge~2004 |      0.011      0.009    1.245   0.213       -0.006       0.029
OO_wo~2004 |      0.000      0.000    0.176   0.860       -0.001       0.001
OO_D_~2004 |     -0.010      0.008   -1.256   0.209       -0.025       0.005
OO_ra~2004 |      0.004      0.010    0.452   0.651       -0.015       0.024
LnAss~2004 |      0.010      0.001    9.338   0.000        0.008       0.012
Total~2004 |      0.001      0.001    2.304   0.021        0.000       0.003
     _cons |      0.018      0.015    1.184   0.237       -0.012       0.048
-----------+------------------------------------------------------------------
tdcaB~5 <- |
PO_ge~2004 |      0.011      0.009    1.245   0.213       -0.006       0.029
Home_~2005 |     -0.035      0.008   -4.290   0.000       -0.051      -0.019
OO_wo~2005 |      0.000      0.000    0.176   0.860       -0.001       0.001
OO_D_~2005 |     -0.010      0.008   -1.256   0.209       -0.025       0.005
OO_ra~2005 |      0.004      0.010    0.452   0.651       -0.015       0.024
LnAss~2005 |      0.010      0.001    9.338   0.000        0.008       0.012
```

```
Total~2005 |      0.001      0.001     2.304   0.021        0.000       0.003
     _cons |     -0.003      0.016    -0.179   0.858       -0.033       0.028
-----------+------------------------------------------------------------------
tdcaB~6 <- |
PO_ge~2004 |      0.011      0.009     1.245   0.213       -0.006       0.029
Home_~2006 |     -0.035      0.008    -4.290   0.000       -0.051      -0.019
OO_wo~2006 |      0.000      0.000     0.176   0.860       -0.001       0.001
OO_D_~2006 |     -0.010      0.008    -1.256   0.209       -0.025       0.005
OO_ra~2006 |      0.004      0.010     0.452   0.651       -0.015       0.024
LnAss~2006 |      0.010      0.001     9.338   0.000        0.008       0.012
Total~2006 |      0.001      0.001     2.304   0.021        0.000       0.003
     _cons |      0.010      0.016     0.650   0.516       -0.020       0.041
-----------+------------------------------------------------------------------
tdcaB~7 <- |
PO_ge~2004 |      0.011      0.009     1.245   0.213       -0.006       0.029
Home_~2007 |     -0.035      0.008    -4.290   0.000       -0.051      -0.019
OO_wo~2007 |      0.000      0.000     0.176   0.860       -0.001       0.001
OO_D_~2007 |     -0.010      0.008    -1.256   0.209       -0.025       0.005
OO_ra~2007 |      0.004      0.010     0.452   0.651       -0.015       0.024
LnAss~2007 |      0.010      0.001     9.338   0.000        0.008       0.012
Total~2007 |      0.001      0.001     2.304   0.021        0.000       0.003
     _cons |      0.011      0.015     0.722   0.470       -0.019       0.041
-----------+------------------------------------------------------------------
tdcaB~8 <- |
PO_ge~2004 |      0.011      0.009     1.245   0.213       -0.006       0.029
Home_~2008 |     -0.035      0.008    -4.290   0.000       -0.051      -0.019
OO_wo~2008 |      0.000      0.000     0.176   0.860       -0.001       0.001
OO_D_~2008 |     -0.010      0.008    -1.256   0.209       -0.025       0.005
OO_ra~2008 |      0.004      0.010     0.452   0.651       -0.015       0.024
LnAss~2008 |      0.010      0.001     9.338   0.000        0.008       0.012
Total~2008 |      0.001      0.001     2.304   0.021        0.000       0.003
     _cons |      0.011      0.016     0.716   0.474       -0.020       0.042
-----------+------------------------------------------------------------------
tdcaB~9 <- |
PO_ge~2004 |      0.011      0.009     1.245   0.213       -0.006       0.029
Home_~2009 |     -0.035      0.008    -4.290   0.000       -0.051      -0.019
OO_wo~2009 |      0.000      0.000     0.176   0.860       -0.001       0.001
OO_D_~2009 |     -0.010      0.008    -1.256   0.209       -0.025       0.005
OO_ra~2009 |      0.004      0.010     0.452   0.651       -0.015       0.024
LnAss~2009 |      0.010      0.001     9.338   0.000        0.008       0.012
Total~2009 |      0.001      0.001     2.304   0.021        0.000       0.003
     _cons |     -0.007      0.015    -0.472   0.637       -0.036       0.022
-----------+------------------------------------------------------------------
tdcaB~0 <- |
PO_ge~2004 |      0.011      0.009     1.245   0.213       -0.006       0.029
Home_~2010 |     -0.035      0.008    -4.290   0.000       -0.051      -0.019
OO_wo~2010 |      0.000      0.000     0.176   0.860       -0.001       0.001
OO_D_~2010 |     -0.010      0.008    -1.256   0.209       -0.025       0.005
OO_ra~2010 |      0.004      0.010     0.452   0.651       -0.015       0.024
LnAss~2010 |      0.010      0.001     9.338   0.000        0.008       0.012
Total~2010 |      0.001      0.001     2.304   0.021        0.000       0.003
     _cons |     -0.011      0.015    -0.703   0.482       -0.040       0.019
-----------+------------------------------------------------------------------
tdcaB~1 <- |
PO_ge~2004 |      0.011      0.009     1.245   0.213       -0.006       0.029
Home_~2011 |     -0.035      0.008    -4.290   0.000       -0.051      -0.019
OO_wo~2011 |      0.000      0.000     0.176   0.860       -0.001       0.001
OO_D_~2011 |     -0.010      0.008    -1.256   0.209       -0.025       0.005
OO_ra~2011 |      0.004      0.010     0.452   0.651       -0.015       0.024
LnAss~2011 |      0.010      0.001     9.338   0.000        0.008       0.012
Total~2011 |      0.001      0.001     2.304   0.021        0.000       0.003
     _cons |     -0.013      0.015    -0.838   0.402       -0.043       0.017
```

```
-------------+------------------------------------------------------------
Variance     |
  e.tdcaPe~4 |      0.061      0.002                          0.057       0.065
  e.tdcaPe~5 |      0.061      0.002                          0.057       0.065
  e.tdcaPe~6 |      0.061      0.002                          0.057       0.065
  e.tdcaPe~7 |      0.061      0.002                          0.057       0.065
  e.tdcaPe~8 |      0.061      0.002                          0.057       0.065
  e.tdcaPe~9 |      0.061      0.002                          0.057       0.065
  e.tdcaPe~0 |      0.061      0.002                          0.057       0.065
  e.tdcaPe~1 |      0.061      0.002                          0.057       0.065
  e.tdcaBu~4 |      0.047      0.002                          0.043       0.052
  e.tdcaBu~5 |      0.047      0.002                          0.043       0.052
  e.tdcaBu~6 |      0.047      0.002                          0.043       0.052
  e.tdcaBu~7 |      0.047      0.002                          0.043       0.052
  e.tdcaBu~8 |      0.047      0.002                          0.043       0.052
  e.tdcaBu~9 |      0.047      0.002                          0.043       0.052
  e.tdcaBu~0 |      0.047      0.002                          0.043       0.052
  e.tdcaBu~1 |      0.047      0.002                          0.043       0.052
-------------+------------------------------------------------------------
Covariance   |
  e.tdcaPe~4 |
  e.tdcaBu~4 |     -0.004      0.001   -4.070   0.000       -0.006      -0.002
  -----------+------------------------------------------------------------
  e.tdcaPe~5 |
  e.tdcaBu~5 |      0.000      0.001    0.207   0.836       -0.002       0.003
  -----------+------------------------------------------------------------
  e.tdcaPe~6 |
  e.tdcaBu~6 |     -0.001      0.001   -1.209   0.227       -0.004       0.001
  -----------+------------------------------------------------------------
  e.tdcaPe~7 |
  e.tdcaBu~7 |     -0.002      0.001   -2.149   0.032       -0.004      -0.000
  -----------+------------------------------------------------------------
  e.tdcaPe~8 |
  e.tdcaBu~8 |     -0.001      0.001   -1.174   0.240       -0.004       0.001
  -----------+------------------------------------------------------------
  e.tdcaPe~9 |
  e.tdcaBu~9 |      0.000      0.001    0.333   0.739       -0.002       0.003
  -----------+------------------------------------------------------------
  e.tdcaPe~0 |
  e.tdcaBu~0 |      0.001      0.001    0.379   0.705       -0.002       0.003
  -----------+------------------------------------------------------------
  e.tdcaPe~1 |
  e.tdcaBu~1 |      0.002      0.002    0.862   0.389       -0.002       0.006
-------------------------------------------------------------------------------
```

6.9 Working with Unbalanced Panel Data with Gaps

So far all our analyses of the panel data focused on using the unbalanced panel (3,140 observations) and utilizing the seventh survey longitudinal weights (wgt_7_long).

```
*Describe pattern of xt data
mi xeq 0: xtdescribe

     Freq.  Percent    Cum. |  Pattern
 -------------------------+----------
     1630     51.91   51.91 |  11111111
      303      9.65   61.56 |  1.......
      283      9.01   70.57 |  11......
      238      7.58   78.15 |  1111....
      224      7.13   85.29 |  111.....
      164      5.22   90.51 |  11111...
      153      4.87   95.38 |  111111..
      145      4.62  100.00 |  1111111.
 -------------------------+----------
     3140    100.00         |  XXXXXXXX
```

A natural question will regard use the full sample (4,928 observations) for panel analysis utilizing the base line survey weights (wgt_final_0). The problem one will face in this case is gaps (missing observations) in the data.

Traditional panel analysis methods of survey data are valid only when the missing completely at random (MCAR) assumption is satisfied. MCAR means that missing values (gaps) are not correlated with any other observed or unobserved variables in the data. Given that the KFS data is self-reported and based on a complex sample design, the MCAR assumption is seldom met, and violations can result in incorrect estimates and decreased efficiency.

On the other hand, the seventh survey longitudinal weight allows completers to mimic a pseudo random sample from the baseline population (something that the base line survey weight cannot do).

```
use Cross_Sectional_Long_MI_Long_L2,clear
merge m:1 mprid   using Kfs8_enclave_14oct13, keepusing(wgt_final_0 )
keep if _merge==3
mi xtset mprid year
drop if cswgt_final==.
keep if classf==6
mi xeq 0: xtdescribe, patterns(300)

     Freq.  Percent    Cum. |  Pattern
 -------------------------+----------
     1575     31.96   31.96 |  11111111
      698     14.16   46.12 |  1.......
      524     10.63   56.76 |  11......
      443      8.99   65.75 |  111.....
      286      5.80   71.55 |  1111....
      213      4.32   75.87 |  11111...
      201      4.08   79.95 |  111111..
      180      3.65   83.60 |  1111111.
```

```
        51      1.03   84.64 |  111.1111
        48      0.97   85.61 |  111111.1
        41      0.83   86.44 |  11.1....
        35      0.71   87.16 |  1111.111
        31      0.63   87.78 |  11111.11
        26      0.53   88.31 |  1.1.....
        25      0.51   88.82 |  1.111111
        25      0.51   89.33 |  11.11111
        22      0.45   89.77 |  1..1....
        21      0.43   90.20 |  111.1...
        19      0.39   90.58 |  1111.1..
        17      0.34   90.93 |  1.11....
        17      0.34   91.27 |  111.11..
        17      0.34   91.62 |  11111..1
        15      0.30   91.92 |  111..111
        14      0.28   92.21 |  11.11...
        14      0.28   92.49 |  1111...1
        13      0.26   92.76 |  1..11111
        13      0.26   93.02 |  1111..11
        13      0.26   93.28 |  11111.1.
        12      0.24   93.53 |  1......1
        12      0.24   93.77 |  11..1111
        12      0.24   94.01 |  11.111..
        12      0.24   94.26 |  111.111.
        11      0.22   94.48 |  11.....1
        10      0.20   94.68 |  11..1...
        10      0.20   94.89 |  1111.11.
         9      0.18   95.07 |  1....1..
         9      0.18   95.25 |  1...1...
         9      0.18   95.43 |  1..111..
         9      0.18   95.62 |  111....1
         9      0.18   95.80 |  111..1..
         8      0.16   95.96 |  11...111
         8      0.16   96.12 |  111.1.11
         8      0.16   96.29 |  1111..1.
         7      0.14   96.43 |  1.1111..
         7      0.14   96.57 |  11..11..
         7      0.14   96.71 |  111.11.1
         6      0.12   96.83 |  1....111
         6      0.12   96.96 |  1...1111
         6      0.12   97.08 |  1..11...
         6      0.12   97.20 |  11...1..
         6      0.12   97.32 |  11...11.
         6      0.12   97.44 |  1111.1.1
         5      0.10   97.54 |  1.11111.
         5      0.10   97.65 |  11.1.111
         5      0.10   97.75 |  11.111.1
         5      0.10   97.85 |  11.1111.
         5      0.10   97.95 |  111..11.
         4      0.08   98.03 |  1....1.1
         4      0.08   98.11 |  1...11..
         4      0.08   98.19 |  11...1.1
         4      0.08   98.28 |  111...1.
         4      0.08   98.36 |  111...11
         4      0.08   98.44 |  111.1.1.
         3      0.06   98.50 |  1....11.
         3      0.06   98.56 |  1..1...1
         3      0.06   98.62 |  1..1..1.
         3      0.06   98.68 |  1.1.1...
         3      0.06   98.74 |  1.111...
         3      0.06   98.80 |  11.1..11
         3      0.06   98.86 |  11.1.11.
```

```
        3       0.06    98.92 |  111.1..1
        2       0.04    98.97 |  1.....1.
        2       0.04    99.01 |  1.....11
        2       0.04    99.05 |  1...111.
        2       0.04    99.09 |  1..11..1
        2       0.04    99.13 |  1..1111.
        2       0.04    99.17 |  1.1...1.
        2       0.04    99.21 |  1.1..1..
        2       0.04    99.25 |  1.1..1.1
        2       0.04    99.29 |  1.1..11.
        2       0.04    99.33 |  1.1.11..
        2       0.04    99.37 |  1.1.1111
        2       0.04    99.41 |  1.11.1..
        2       0.04    99.45 |  1.1111.1
        2       0.04    99.49 |  11....1.
        2       0.04    99.53 |  11....11
        2       0.04    99.57 |  11..1.11
        2       0.04    99.61 |  11..111.
        2       0.04    99.66 |  11.1...1
        2       0.04    99.70 |  11.1.1..
        1       0.02    99.72 |  1...11.1
        1       0.02    99.74 |  1..1.11.
        1       0.02    99.76 |  1..1.111
        1       0.02    99.78 |  1..11.1.
        1       0.02    99.80 |  1..11.11
        1       0.02    99.82 |  1.1..111
        1       0.02    99.84 |  1.1.1..1
        1       0.02    99.86 |  1.1.1.11
        1       0.02    99.88 |  1.11...1
        1       0.02    99.90 |  1.11.1.1
        1       0.02    99.92 |  1.111.11
        1       0.02    99.94 |  11..1..1
        1       0.02    99.96 |  11..11.1
        1       0.02    99.98 |  11.11.11
        1       0.02   100.00 |  111..1.1
  --------------------------+----------
     4928    100.00         |  XXXXXXXX
```

6.10 Working with Cross-Sectional Surveys

The KFS is a multipurpose survey that can be used to measure gross change (also known as micro/individual/internal change) to measure current levels (also known as snap shot/static/point in time/one-shot/single-period estimate) and measure net change (also known as macro/mean/external change).

Measuring current levels involves using one follow-up survey. Meanwhile, measuring net change involves using more than one survey. Because the KFS can be treated as a repeated cross-sectional survey with observations overlapping, the assumption that observations are independent will not hold; thus, we should take into account the covariance structure to draw valid statistical inferences.

This section will focus on two general goals and will provide examples for each when making inferences from the KFS cross-sectional surveys. The first goal is estimating a change (net change) in a characteristic between two points of time. The second goal is showing how to use some of the Stata regression commands with a single-period cross-sectional analysis.

6.10.1 Net Change in a Characteristic between Two Points of Time

Survey estimates like these should use the cross-sectional weight and the count of employees for the respective years. It is important to note that if there are missing data for either of the years, the survey estimate may be biased. This bias can be either positive or negative and the magnitude of the bias is unknown. To avoid such a bias, we will use MI data.

Examples 6.40 Net Change in Employment

Total_Employees	The KFS Multiply Imputed Data		The KFS Non-Imputed Data	
Year	Mean	N	Mean	N
2004	2.51	4,928	2.51	4,797
2005	3.90	3,998	3.86	3,885
2006	4.25	3,390	4.23	3,325
2007	4.42	2,915	4.45	2,879
2008	4.45	2,606	4.45	2,591
2009	4.44	2,408	4.39	2,390
2010	4.67	2,126	4.68	2,118
2011	5.29	2,007	5.31	1,995

```
*Employment Growth Relative to the Baseline Year 2004
use Cross_Sectional_Long_MI_Long_L2,clear
*Declare survey design for dataset
mi svyset  mprid  [pweight=cswgt_final] , strata(sampleinfo_samplestrata)
mi estimate :svy: reg Total_Employees i.year
mi estimate,  cformat(%6.2f)  sformat(%6.2f)   nolstretch

Multiple-imputation estimates                   Imputations        =         5
Survey: Linear regression                       Number of obs      =     24378

Number of strata  =         6                   Population size    = 402372.87
Number of PSUs    =      4928
                                                Average RVI        =    0.0015
                                                Largest FMI        =    0.0057
                                                Complete DF        =      4922
DF adjustment:   Small sample                   DF:     min        =   4704.63
                                                        avg        =   4857.60
                                                        max        =   4915.61
Model F test:       Equal FMI                   F(   7, 4918.9)    =     26.36
Within VCE type:   Linearized                   Prob > F           =    0.0000
------------------------------------------------------------------------------
Total_Empl~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |
       2005  |       1.39       0.12    11.34   0.000         1.15        1.63
       2006  |       1.75       0.14    12.29   0.000         1.47        2.02
       2007  |       1.92       0.22     8.79   0.000         1.49        2.35
       2008  |       1.94       0.24     7.97   0.000         1.46        2.42
       2009  |       1.93       0.27     7.21   0.000         1.41        2.46
       2010  |       2.17       0.31     6.98   0.000         1.56        2.78
       2011  |       2.79       0.46     6.09   0.000         1.89        3.69
       _cons |       2.51       0.09    26.73   0.000         2.32        2.69
------------------------------------------------------------------------------
```

```
*Employment Growth 2004-2005

use Cross_Sectional_Long_MI_Long_L2,clear
*Declare survey design for dataset
mi svyset  mprid  [pweight=cswgt_final] , strata(sampleinfo_samplestrata)
mi estimate :svy: reg Total_Employees i.year if year ==2004 | year==2005
mi estimate,  cformat(%6.2f)  sformat(%6.2f)   nolstretch

Multiple-imputation estimates                   Imputations        =         5
Survey: Linear regression                       Number of obs      =      8926

Number of strata  =         6                   Population size    = 140230.77
Number of PSUs    =      4928
                                                Average RVI        =    0.0046
                                                Largest FMI        =    0.0057
                                                Complete DF        =      4922
DF adjustment:   Small sample                   DF:     min        =   4704.63
                                                        avg        =   4744.79
                                                        max        =   4784.94
Model F test:       Equal FMI                   F(   1, 4784.9)    =    128.70
Within VCE type:   Linearized                   Prob > F           =    0.0000

------------------------------------------------------------------------------
Total_Empl~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |
       2005  |       1.39       0.12    11.34   0.000         1.15        1.63
             |
       _cons |       2.51       0.09    26.73   0.000         2.32        2.69
------------------------------------------------------------------------------

*Employment Growth 2006-2010

use Cross_Sectional_Long_MI_Long_L2,clear
*Declare survey design for dataset
mi svyset  mprid  [pweight=cswgt_final] , strata(sampleinfo_samplestrata)
mi estimate :svy: reg Total_Employees i.year if year ==2006 | year==2010
mi estimate,  cformat(%6.2f)  sformat(%6.2f)   nolstretch

Multiple-imputation estimates                   Imputations        =         5
Survey: Linear regression                       Number of obs      =      5516

Number of strata  =         6                   Population size    = 93635.946
Number of PSUs    =      3509
                                                Average RVI        =    0.0001
                                                Largest FMI        =    0.0001
                                                Complete DF        =      3503
DF adjustment:   Small sample                   DF:     min        =   3500.49
                                                        avg        =   3500.61
                                                        max        =   3500.74
Model F test:       Equal FMI                   F(   1, 3500.7)    =      2.29
Within VCE type:   Linearized                   Prob > F           =    0.1300
------------------------------------------------------------------------------
Total_Empl~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |
       2010  |       0.42       0.28     1.51   0.130        -0.12        0.97
             |
       _cons |       4.25       0.16    26.25   0.000         3.93        4.57
------------------------------------------------------------------------------
```

```
*Employment Growth Relative to Financial Crisis of 2007
use Cross_Sectional_Long_MI_Long_L2,clear
*Declare survey design for dataset
mi svyset  mprid  [pweight=cswgt_final] , strata(sampleinfo_samplestrata)
mi estimate :svy: reg Total_Employees i.year if year >2006
mi estimate,  cformat(%6.2f)  sformat(%6.2f)   nolstretch

Multiple-imputation estimates                   Imputations        =         5
Survey: Linear regression                       Number of obs      =     12062

Number of strata  =         6                   Population size    = 204187.82
Number of PSUs    =      3237
                                                Average RVI        =    0.0002
                                                Largest FMI        =    0.0004
                                                Complete DF        =      3231
DF adjustment:   Small sample                   DF:     min        =   3227.12
                                                        avg        =   3227.68
                                                        max        =   3228.36
Model F test:       Equal FMI                   F(   4, 3229.0)    =      1.36
Within VCE type:   Linearized                   Prob > F           =    0.2442
------------------------------------------------------------------------------
Total_Empl~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |
       2008  |       0.02       0.21     0.11   0.914        -0.40        0.44
       2009  |       0.01       0.26     0.06   0.954        -0.49        0.52
       2010  |       0.25       0.26     0.97   0.330        -0.25        0.75
       2011  |       0.87       0.42     2.06   0.040         0.04        1.70
             |
       _cons |       4.42       0.23    19.14   0.000         3.97        4.88
------------------------------------------------------------------------------
*Percentage of Firms with More Than One Employee in 2004 Versus 2011
* 0-1 indicator variable is to denote whether the firm has more than one employee
use Cross_Sectional_Long_MI_Long_L2,clear
*Declare survey design for dataset
mi svyset  mprid  [pweight=cswgt_final] , strata(sampleinfo_samplestrata)
gen MultipleE=(Total_Employees>1  )
replace MultipleE=.a  if   Total_Employees==.a
mi estimate :svy: reg MultipleE i.year if year ==2004 | year==2011
mi estimate,  cformat(%6.2f)  sformat(%6.2f)   nolstretch

Multiple-imputation estimates                   Imputations        =         5
Survey: Linear regression                       Number of obs      =      6935

Number of strata  =         6                   Population size    = 105959.76
Number of PSUs    =      4928
                                                Average RVI        =    0.0027
                                                Largest FMI        =    0.0035
                                                Complete DF        =      4922
DF adjustment:   Small sample                   DF:     min        =   4828.57
                                                        avg        =   4850.90
                                                        max        =   4873.22
Model F test:       Equal FMI                   F(   1, 4828.6)    =     65.06
Within VCE type:   Linearized                   Prob > F           =    0.0000
------------------------------------------------------------------------------
   MultipleE |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |
       2011  |       0.11       0.01     8.07   0.000         0.08        0.14
             |
       _cons |       0.40       0.01    49.13   0.000         0.38        0.41
------------------------------------------------------------------------------
```

```
*Percentage of Firms with More Than Five Employee in 2004 Versus 2011

* 0-1 indicator variable is to denote whether the firm has more than 5 employee
use Cross_Sectional_Long_MI_Long_L2,clear
*Declare survey design for dataset
mi svyset  mprid  [pweight=cswgt_final] , strata(sampleinfo_samplestrata)
gen MultipleE=(Total_Employees>4  )
replace MultipleE=.a  if   Total_Employees==.a
mi estimate :svy: reg MultipleE i.year if year ==2004 | year==2011
mi estimate,  cformat(%6.2f)  sformat(%6.2f)   nolstretch



Multiple-imputation estimates                   Imputations        =         5
Survey: Linear regression                       Number of obs      =      6935
Number of strata  =         6                   Population size    = 105959.76
Number of PSUs    =      4928
                                                Average RVI        =    0.0101
                                                Largest FMI        =    0.0169
                                                Complete DF        =      4922
DF adjustment:   Small sample                   DF:     min        =   3615.87
                                                        avg        =   4157.97
                                                        max        =   4700.06
Model F test:       Equal FMI                   F(   1, 4700.1)    =     99.38
Within VCE type:   Linearized                   Prob > F           =    0.0000
------------------------------------------------------------------------------
   MultipleE |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |
       2011  |       0.11       0.01     9.97   0.000         0.09        0.13
             |
       _cons |       0.13       0.01    23.37   0.000         0.12        0.14
------------------------------------------------------------------------------

*Employment Growth Relative to the Baseline Year 2004 by Gender

use Cross_Sectional_Long_MI_Long_L2,clear
*Declare survey design for dataset
mi svyset  mprid  [pweight=cswgt_final] , strata(sampleinfo_samplestrata)
gen Female=PO_gender==0
replace Female=.a  if      PO_gender==.a

mi estimate :svy,subpop(Female): reg Total_Employees i.year
mi estimate,  cformat(%6.2f)  sformat(%6.2f)   nolstretch
```

```
Multiple-imputation estimates                   Imputations        =         5
Survey: Linear regression                       Number of obs      =     24378

Number of strata  =         6                   Population size    = 402372.87
Number of PSUs    =      4928                   Subpop. no. of obs =      6002
                                                Subpop. size       =  117986.5
                                                Average RVI        =    0.0099
                                                Largest FMI        =    0.0372
                                                Complete DF        =      4922
DF adjustment:   Small sample                   DF:     min        =   1833.68
                                                        avg        =   4136.89
                                                        max        =   4905.44
Model F test:       Equal FMI                   F(   7, 4872.4)    =      7.49
Within VCE type:   Linearized                   Prob > F           =    0.0000
------------------------------------------------------------------------------
Total_Empl~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |
       2005  |       1.08       0.19     5.68   0.000         0.71        1.45
       2006  |       1.22       0.20     6.10   0.000         0.83        1.61
       2007  |       1.36       0.34     3.96   0.000         0.69        2.04
       2008  |       1.68       0.46     3.65   0.000         0.78        2.58
       2009  |       1.36       0.52     2.61   0.009         0.34        2.38
       2010  |       2.01       0.69     2.90   0.004         0.65        3.36
       2011  |       1.96       0.74     2.66   0.008         0.52        3.41
             |
       _cons |       1.84       0.12    14.74   0.000         1.60        2.09
------------------------------------------------------------------------------
mi estimate :svy,subpop(if Female==0): reg Total_Employees i.year
mi estimate,  cformat(%6.2f)  sformat(%6.2f)   nolstretch

Multiple-imputation estimates                   Imputations        =         5
Survey: Linear regression                       Number of obs      =     34284

Number of strata  =         6                   Population size    = 586227.52
Number of PSUs    =      4928                   Subpop. no. of obs =     18376
                                                Subpop. size       = 284386.37
                                                Average RVI        =    0.0005
                                                Largest FMI        =    0.0014
                                                Complete DF        =      4922
DF adjustment:   Small sample                   DF:     min        =   4901.87
                                                        avg        =   4915.91
                                                        max        =   4919.70
Model F test:       Equal FMI                   F(   7, 4919.8)    =     21.01
Within VCE type:   Linearized                   Prob > F           =    0.0000

------------------------------------------------------------------------------
Total_Empl~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |
       2005  |       1.52       0.15     9.80   0.000         1.21        1.82
       2006  |       1.94       0.18    10.70   0.000         1.58        2.29
       2007  |       2.13       0.27     7.82   0.000         1.59        2.66
       2008  |       2.05       0.29     7.14   0.000         1.49        2.61
       2009  |       2.16       0.31     6.95   0.000         1.55        2.77
       2010  |       2.22       0.33     6.65   0.000         1.57        2.87
       2011  |       3.11       0.57     5.46   0.000         1.99        4.23
             |
       _cons |       2.79       0.12    22.76   0.000         2.55        3.03
------------------------------------------------------------------------------
```

```
*Employment Growth Relative to the Baseline Year 2004 by Gender×Race
use Cross_Sectional_Long_MI_Long_L2,clear
*Declare survey design for dataset
mi svyset  mprid  [pweight=cswgt_final] , strata(sampleinfo_samplestrata)
gen Female=PO_gender==0
replace Female=.a  if      PO_gender==.a

mi   estimate   :svy,subpop(if   Female==1   &   PO_race_white_owner==0):   reg
Total_Employees i.year
mi estimate,  cformat(%6.2f)  sformat(%6.2f)   nolstretch


Multiple-imputation estimates                   Imputations        =         5
Survey: Linear regression                       Number of obs      =     34284

Number of strata  =         6                   Population size    = 586227.52
Number of PSUs    =      4928                   Subpop. no. of obs =      1134
                                                Subpop. size       = 23216.433
                                                Average RVI        =    0.0025
                                                Largest FMI        =    0.0131
                                                Complete DF        =      4922
DF adjustment:   Small sample                   DF:     min        =   4031.46
                                                        avg        =   4783.02
                                                        max        =   4919.76
Model F test:       Equal FMI                   F(   7, 4917.6)    =      3.41
Within VCE type:   Linearized                   Prob > F           =    0.0012
------------------------------------------------------------------------------
Total_Empl~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |
       2005  |       1.55       0.45     3.47   0.001         0.68        2.43
       2006  |       1.88       0.55     3.42   0.001         0.80        2.95
       2007  |       2.98       1.50     1.99   0.046         0.05        5.91
       2008  |       2.50       1.33     1.87   0.061        -0.12        5.11
       2009  |       0.85       0.36     2.35   0.019         0.14        1.57
       2010  |       4.36       2.90     1.51   0.132        -1.31       10.04
       2011  |       4.23       3.13     1.35   0.177        -1.92       10.37
             |
       _cons |       1.69       0.19     8.95   0.000         1.32        2.06
------------------------------------------------------------------------------
```

```
mi   estimate   :svy,subpop(if   Female==1   &   PO_race_white_owner==1):   reg
Total_Employees i.year
mi estimate,  cformat(%6.2f)  sformat(%6.2f)   nolstretch

Multiple-imputation estimates                   Imputations        =         5
Survey: Linear regression                       Number of obs      =     34284

Number of strata  =         6                   Population size    = 586227.52
Number of PSUs    =      4928                   Subpop. no. of obs =      4868
                                                Subpop. size       = 94770.067
                                                Average RVI        =    0.0143
                                                Largest FMI        =    0.0449
                                                Complete DF        =      4922
DF adjustment:   Small sample                   DF:     min        =   1436.68
                                                        avg        =   3572.88
                                                        max        =   4881.67
Model F test:       Equal FMI                   F(   7, 4799.0)    =      5.00
Within VCE type:   Linearized                   Prob > F           =    0.0000

------------------------------------------------------------------------------
Total_Empl~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |
       2005  |       0.96       0.21     4.56   0.000         0.54        1.37
       2006  |       1.05       0.21     5.01   0.000         0.64        1.46
       2007  |       0.99       0.26     3.87   0.000         0.49        1.49
       2008  |       1.48       0.47     3.13   0.002         0.55        2.41
       2009  |       1.47       0.63     2.32   0.021         0.23        2.71
       2010  |       1.46       0.53     2.74   0.006         0.41        2.51
       2011  |       1.47       0.60     2.46   0.014         0.30        2.64
             |
       _cons |       1.88       0.15    12.48   0.000         1.59        2.18
------------------------------------------------------------------------------
```

6.10.2 Single-Period Cross Sectional Analysis

In this section, we will show how to use some of the Stata regression commands with a single-period cross-sectional analysis.

Examples 6.41 Bivariate Probit Regression

Bivariate Probit Regression fits maximum-likelihood two-equation probit models. The model setup is as follows:

$$y_{1i}^{*} = \alpha x_{1i} + \varepsilon_{1i} \ \{ y_{1i} \ = 1 \ if \ y_{1i}^{*} > 0 \ , 0 \ otherwise$$
$$y_{2i}^{*} = \beta x_{2i} + \varepsilon_{2i} \ \{ y_{2i} \ = 1 \ if \ y_{2i}^{*} > 0 \ , 0 \ otherwise$$

where x_{1i} , x_{2i} denote vectors of exogenous variables (for the ith firm) and the α and β are vectors of parameters. The parameter ρ_{12} estimates the correlation between the error terms of the Bivariate Probit equations. If the MLE estimate of the correlation coefficient ρ_{12} is significant, then the bivariate probit estimation is more efficient than that of independent probit equations.

Assume that y_1 equal 1 if the firm used equity financing in 2004 and 0 otherwise. Assume that y_2 equal 1 if the firm used debt financing in 2004 and 0 otherwise.

```
use Cross_Sectional_Long_MI_Long_L2,clear

keep if year==2004

egen capital_injection=rowtotal(Debt  Equity )
gen Debt_inj= Debt / capital_injection
gen Equity_inj= Equity / capital_injection
recode Debt_inj Equity_inj (.=0) if master!=0 & capital_injection==0
gen Eq=0
replace Eq=1 if Equity_inj>0 & Equity_inj<.
gen De=0
replace De=1 if Debt_inj>0  & Debt_inj<.
mi svyset  mprid  [pweight=cswgt_final] , strata(sampleinfo_samplestrata)
mi estimate,cmdok :svy: tab Eq De

Multiple-imputation estimates                   Imputations        =         5
                                                Number of obs      =      4928
Number of strata  =         6
Number of PSUs    =      4928                   Population size    = 73278.441

                                                Average RVI        =    0.0112
                                                Largest FMI        =    0.0194
                                                Complete DF        =      4922
DF adjustment:   Small sample                   DF:     min        =   3339.86
                                                        avg        =   4254.85
Within VCE type:   Linearized                           max        =   4863.42
------------------------------------------------------------------------------
        Mean |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         p11 |   .0796751   .0044351    17.96   0.000     .0709794    .0883709
         p12 |   .0974686   .0049525    19.68   0.000     .0877591    .1071782
         p21 |   .3275389   .0077254    42.40   0.000     .3123936    .3426841
         p22 |   .4953174   .0082768    59.84   0.000     .4790909    .5115438
------------------------------------------------------------------------------
```

```
------------------------------
          |         De
       Eq |     0      1  Total
----------+-------------------
        0 | .0794  .0973  .1767
        1 | .3274  .4959  .8233
          |
    Total | .4068  .5932      1
------------------------------

gen LnAssets=ln( Assets+1)
mi estimate,cmdok :svy: biprobit Eq De LnAssets   Home_Based  OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner OO_gender  Comp_advantage
mi estimate,cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                   Imputations       =         5
Survey: Bivariate probit regression             Number of obs     =      4928

Number of strata  =         6                   Population size   = 73278.441
Number of PSUs    =      4928
                                                Average RVI       =    0.0114
                                                Largest FMI       =    0.0423
                                                Complete DF       =      4922
DF adjustment:   Small sample                   DF:     min       =   1560.05
                                                        avg       =   4178.34
                                                        max       =   4904.94
Model F test:       Equal FMI                   F(  20, 4880.0)   =     18.74
Within VCE type:   Linearized                   Prob > F          =    0.0000
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Eq           |
    LnAssets |      0.047      0.007    6.781   0.000        0.033       0.060
  Home_Based |      0.015      0.051    0.294   0.768       -0.085       0.115
OO_D_educa~r |      0.190      0.050    3.777   0.000        0.091       0.289
OO_work_ex~r |     -0.005      0.002   -2.128   0.033       -0.010      -0.000
OO_race_wh~r |     -0.063      0.066   -0.950   0.342       -0.193       0.067
OO_gender_~r |     -0.012      0.061   -0.202   0.840       -0.131       0.107
Comp_advan~e |      0.134      0.053    2.544   0.011        0.031       0.237
    hightech |      0.061      0.077    0.791   0.429       -0.090       0.211
d1a_provid~e |     -0.003      0.082   -0.036   0.971       -0.164       0.159
d1b_provid~t |      0.221      0.055    3.981   0.000        0.112       0.329
       _cons |      0.351      0.129    2.716   0.007        0.098       0.604
-------------+----------------------------------------------------------------
De           |
    LnAssets |      0.078      0.007   11.750   0.000        0.065       0.091
  Home_Based |     -0.215      0.044   -4.856   0.000       -0.302      -0.128
OO_D_educa~r |     -0.103      0.044   -2.342   0.019       -0.189      -0.017
OO_work_ex~r |     -0.014      0.002   -6.259   0.000       -0.018      -0.009
OO_race_wh~r |     -0.005      0.059   -0.090   0.928       -0.120       0.109
OO_gender_~r |      0.025      0.054    0.470   0.638       -0.080       0.130
Comp_advan~e |      0.111      0.046    2.415   0.016        0.021       0.202
    hightech |     -0.145      0.067   -2.151   0.031       -0.277      -0.013
d1a_provid~e |     -0.017      0.069   -0.247   0.805       -0.152       0.118
d1b_provid~t |      0.064      0.048    1.349   0.178       -0.029       0.158
       _cons |     -0.257      0.115   -2.233   0.026       -0.483      -0.031
-------------+----------------------------------------------------------------
     /athrho |      0.000      0.032    0.013   0.989       -0.062       0.062
-------------+----------------------------------------------------------------
         rho |      0.000      0.032                        -0.061       0.062
------------------------------------------------------------------------------
```

Examples 6.42 Probit Model with Sample Selection

In general, a selected sample is a term used to describe a nonrandom sample. The probit model with sample selection assumes that

$$y_{1i}^{*} = \alpha x_{1i} + \varepsilon_{1i} \ \{ y_{1i} \ = 1 \ if \ y_{1i}^{*} > 0 \ , 0 \ otherwise$$

where the dependent variable (y_{1i}^{*}) for observation *i* is observed only if $y_{2i}^{*} > 0$, in the selection equation:

$$y_{2i}^{*} = \beta x_{2i} + \varepsilon_{2i} \ \{ y_{2i} \ = 1 \ if \ y_{2i}^{*} > 0 \ , 0 \ otherwise$$

The parameter $\rho_{12} = Corr(\varepsilon_1, \varepsilon_2)$ estimates the correlation between the error terms of the Bivariate Probit equations when $\rho_{12} \neq 0$, standard probit techniques applied to the first equation yield biased results.

Consider studying loans approved/denied (question number f14e_approved_denied: Were these applications always approved, sometimes approved and sometimes denied, or always denied?) in the year 2008. This selected sample is not random. Also, we only observe data for f14e_approved_denied if the answer to f14d_new_loans (Did [NAME BUSINESS] make any applications for new or renewed loans or lines of credit in calendar year YYYY?) is yes.

```
use Cross_Sectional_Long_MI_Long_L2,clear
gen LnAssets=ln( Assets+1)
keep if year==2008
drop if cswgt_final>=.
mi svyset  mprid  [pweight=cswgt_final] , strata(sampleinfo_samplestrata)
recode f14e_approved_denied (2=0) (3=0)
mi  estimate,cmdok  esampvaryok:svy:  heckprob  f14e_approved_denied     LnAssets
OO_D_education_owner      OO_work_exp_owner      OO_race_white_owner   OO_gender
Comp_advantage , ///
select(f14d_new_loans =LnAssets Home_Based    )      difficult
mi estimate,cformat(%6.3f)  sformat(%6.3f)   nolstretch
```

```
Multiple-imputation estimates                   Imputations        =         5
Survey: Probit model with sample selection      Number of obs      =      2603

Number of strata  =         6                   Population size    = 44583.199
Number of PSUs    =      2603
                                                Average RVI        =    0.0018
                                                Largest FMI        =    0.0064
                                                Complete DF        =      2597
                                                DF:     min        =   2513.54
                                                        avg        =   2577.95
DF adjustment:   Small sample                           max        =   2594.57
                                                F(   6,      .)    =         .
Within VCE type:   Linearized                   Prob > F           =         .

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
f14e_appro~d |
    LnAssets |      0.131      0.031    4.188   0.000        0.070       0.193
OO_D_educa~r |      0.065      0.117    0.557   0.578       -0.165       0.295
OO_work_ex~r |      0.009      0.008    1.174   0.240       -0.006       0.024
OO_race_wh~r |      0.395      0.224    1.759   0.079       -0.045       0.835
OO_gender_~r |      0.193      0.167    1.160   0.246       -0.133       0.520
Comp_advan~e |     -0.034      0.118   -0.289   0.773       -0.265       0.197
       _cons |     -3.089      0.583   -5.293   0.000       -4.233      -1.944
-------------+----------------------------------------------------------------
f14d_new_l~s |
    LnAssets |      0.113      0.029    3.836   0.000        0.055       0.171
  Home_Based |     -0.370      0.092   -4.004   0.000       -0.551      -0.189
       _cons |     -2.210      0.349   -6.341   0.000       -2.894      -1.527
-------------+----------------------------------------------------------------
     /athrho |      1.189      0.793    1.499   0.134       -0.366       2.744
-------------+----------------------------------------------------------------
         rho |      0.830      0.246                       -0.351       0.992
------------------------------------------------------------------------------
Warning: estimation sample varies across imputations; results may be biased.
         Sample sizes vary between 2603 and 2606.
Note: number of primary clusters varies among imputations.
Note: population size varies among imputations.
```

Examples 6.43 Heckman Selection Model

In general, a selected sample is a term used to describe a nonrandom sample. The Heckman selection model assumes an underlying regression relationship:

$$y_{1i} = \alpha x_{1i} + \varepsilon_{1i}$$

where the dependent variable (y_{1i}) for observation *i* is observed only if $y_{2i}^{*} > 0$, with the selection equation:

$$y_{2i}^{*} = \beta x_{2i} + \varepsilon_{2i} \ \{ y_{2i} \ = 1 \ if \ y_{2i}^{*} > 0 \ , 0 \ otherwise$$

The parameter $\rho_{12} = Corr(\varepsilon_1, \varepsilon_2)$ estimates the correlation between the error terms in the two equations when $\rho_{12} \neq 0$, standard probit techniques applied to the first equation yield biased results.

Consider studying the amount of equity injection by owner 01 (f2_owner_amt_eq_invest_01) in the year 2004. This selected sample is not random. Also, we only observe data for f2_owner_amt_eq_invest_01 if the answer to f2_owner_eq_invest_01 is yes.

```
use Cross_Sectional_Long_MI_Long_L2,clear
gen LnAssets=ln( Assets+1)
keep if year==2004
mi svyset  mprid  [pweight=cswgt_final] , strata(sampleinfo_samplestrata)
egen capital_injection=rowtotal(Debt  Equity )
gen Debt_inj= Debt / capital_injection
gen Equity_inj= Equity / capital_injection
recode Debt_inj Equity_inj (.=0) if master!=0 & capital_injection==0
gen Lnf2_owner_amt_eq_invest_01=ln( f2_owner_amt_eq_invest_01+1)

mi   estimate,cmdok   esampvaryok:svy:      heckman   Lnf2_owner_amt_eq_invest_01
LnAssets  Debt_inj Home_Based  OO_D_education_owner    Comp_advantage  , ///
select(f2_owner_eq_invest_01   =LnAssets  OO_work_exp_owner   OO_race_white_owner
OO_gender    )
mi estimate,cformat(%6.3f)  sformat(%6.3f)   nolstretch


Multiple-imputation estimates                   Imputations        =         5
Survey: Heckman selection model                 Number of obs      =      4928

Number of strata  =         6                   Population size    = 73278.441
Number of PSUs    =      4928
                                                Average RVI        =    0.0401
                                                Largest FMI        =    0.1728
                                                Complete DF        =      4922
                                                DF:     min        =    147.06
                                                        avg        =   3023.53
DF adjustment:   Small sample                           max        =   4884.87
                                                F(   5,      .)    =         .
Within VCE type:   Linearized                   Prob > F           =         .
```

```
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Lnf2_owne~01 |
    LnAssets |      0.213      0.013   16.751   0.000        0.188       0.237
    Debt_inj |     -0.642      0.095   -6.786   0.000       -0.828      -0.456
  Home_Based |     -0.778      0.059  -13.083   0.000       -0.895      -0.662
OO_D_educa~r |      0.149      0.057    2.607   0.009        0.037       0.261
Comp_advan~e |      0.069      0.060    1.154   0.248       -0.048       0.187
       _cons |      7.656      0.142   53.740   0.000        7.376       7.935
-------------+----------------------------------------------------------------
f2_owner_e.. |
    LnAssets |      0.046      0.007    6.922   0.000        0.033       0.059
OO_work_ex~r |     -0.004      0.002   -1.644   0.100       -0.008       0.001
OO_race_wh~r |     -0.029      0.063   -0.469   0.639       -0.152       0.094
OO_gender_~r |     -0.074      0.060   -1.229   0.219       -0.192       0.044
       _cons |      0.537      0.087    6.202   0.000        0.367       0.707
-------------+----------------------------------------------------------------
     /athrho |      0.144      0.041    3.542   0.000        0.065       0.224
    /lnsigma |      0.411      0.020   20.073   0.000        0.371       0.452
-------------+----------------------------------------------------------------
         rho |      0.143      0.040                        0.064       0.221
       sigma |      1.509      0.031                        1.449       1.571
      lambda |      0.216      0.060
------------------------------------------------------------------------------
```

```
use Cross_Sectional_Long_MI_Long_L2,clear
gen LnAssets=ln( Assets+1)
keep if year==2004
mi svyset  mprid  [pweight=cswgt_final] , strata(sampleinfo_samplestrata)
egen capital_injection=rowtotal(Debt  Equity )
gen Debt_inj= Debt / capital_injection
gen Equity_inj= Equity / capital_injection
recode Debt_inj Equity_inj (.=0) if master!=0 & capital_injection==0
gen Lnf2_owner_amt_eq_invest_01=ln( f2_owner_amt_eq_invest_01+1)

mi   estimate,cmdok   esampvaryok:svy:      heckman   Lnf2_owner_amt_eq_invest_01
LnAssets  Debt_inj Home_Based  OO_D_education_owner    Comp_advantage  , ///
select(f2_owner_eq_invest_01   =LnAssets  OO_work_exp_owner   OO_race_white_owner
OO_gender    )
mi estimate,cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                   Imputations        =         5
Survey: Heckman selection model                 Number of obs      =      4928

Number of strata  =         6                   Population size    = 73278.441
Number of PSUs    =      4928
                                                Average RVI        =    0.0401
                                                Largest FMI        =    0.1728
                                                Complete DF        =      4922
                                                DF:     min        =    147.06
                                                        avg        =   3023.53
DF adjustment:   Small sample                           max        =   4884.87
                                                F(   5,      .)    =         .
Within VCE type:   Linearized                   Prob > F           =         .
```

```
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Lnf2_owne~01 |
    LnAssets |      0.213      0.013   16.751   0.000        0.188       0.237
    Debt_inj |     -0.642      0.095   -6.786   0.000       -0.828      -0.456
  Home_Based |     -0.778      0.059  -13.083   0.000       -0.895      -0.662
OO_D_educa~r |      0.149      0.057    2.607   0.009        0.037       0.261
Comp_advan~e |      0.069      0.060    1.154   0.248       -0.048       0.187
       _cons |      7.656      0.142   53.740   0.000        7.376       7.935
-------------+----------------------------------------------------------------
f2_owner_e.. |
    LnAssets |      0.046      0.007    6.922   0.000        0.033       0.059
OO_work_ex~r |     -0.004      0.002   -1.644   0.100       -0.008       0.001
OO_race_wh~r |     -0.029      0.063   -0.469   0.639       -0.152       0.094
OO_gender_~r |     -0.074      0.060   -1.229   0.219       -0.192       0.044
       _cons |      0.537      0.087    6.202   0.000        0.367       0.707
-------------+----------------------------------------------------------------
     /athrho |      0.144      0.041    3.542   0.000        0.065       0.224
    /lnsigma |      0.411      0.020   20.073   0.000        0.371       0.452
-------------+----------------------------------------------------------------
         rho |      0.143      0.040                        0.064       0.221
       sigma |      1.509      0.031                        1.449       1.571
      lambda |      0.216      0.060
------------------------------------------------------------------------------
```

Examples 6.44 Interval Regression

Interval regression fits a model of y=[depvar1, depvar2] on indepvars, where y for each observation is point data, interval data, left-censored data, or right-censored data.

Interval regression is used to model outcomes that have interval censoring (*y=[depvar1, depvar2]*) where we know the ordered category into which each observation falls, but we do not know the exact value of the observation. The *depvar1* and *depvar2* should have the following form:

Type of data		*depvar1*	*depvar2*
Point Data	***a=[a,a]***	***a***	***a***
Interval Data	***[a,b]***	***a***	***b***
Left-Censored Data	***(−∞,b]***	**.**	***b***
Right-Censored Data	***[a,∞)***	***a***	**.**

For all the financial variables in the KFS if the respondent did not provide the exact amount of the variable in dollars, the respondent was asked to provide a range of the amount instead. The range interval classes were standard across all the financial variables in the KFS. The interval classes were:

$0.....00
$500 or less,.....01
$501 to $1,000,.....02
$1,001 to $3,000,.....03
$3,001 to $5,000,.....04
$5,001 to $10,000,.....05
$10,001 to $25,000,.....06
$25,001 to $100,000,.....07
$100,001 to $1,000,000,.....08
$1,000,001 or more?.....09
Don't Know....."."
Refused....."."

For businesses that reported the range rather than exact value, replacing missing continuous values by the midpoints of the class interval is a common approach by KFS researchers. Nonetheless, using interval regression could result in less bias in the estimate relative to the use of the midpoints of the class interval.

```
use Cross_Sectional_Long_MI_Long_L2,clear
gen LnAssets=ln( Assets+1)
keep if year==2004

merge m:1 mprid  using Kfs8_enclave_14oct13, keepusing(f16b_rev_2004_amt_ranges_0
)
keep if _merge==3
```

```
mi svyset  mprid  [pweight=cswgt_final] , strata(sampleinfo_samplestrata)
egen capital_injection=rowtotal(Debt  Equity )
gen Debt_inj= Debt / capital_injection
gen Equity_inj= Equity / capital_injection
recode Debt_inj Equity_inj (.=0) if master!=0 & capital_injection==0
* Replacing missing continuous values by the midpoints of the class intervals
* We already did taht for Cross_Sectional_Long_MI_Long_L2.dta see section 3.3.
      Comparing the KFS Imputed to Non-Imputed Data
* OLS with vs. Interval regression
mi xeq 0 :svy: reg  f16a_rev_amt     LnAssets           Debt_inj      Home_Based
OO_D_education_owner         Comp_advantage     ,cformat(%6.3f)    sformat(%6.3f)
nolstretch

Survey: Linear regression

Number of strata   =         6                  Number of obs      =      3749
Number of PSUs     =      3749                  Population size    = 56062.598
                                                 Design df          =      3743
                                                 F(   5,   3739)    =      9.37
                                                 Prob > F           =    0.0000
                                                 R-squared          =    0.0070

------------------------------------------------------------------------------
             |             Linearized
f16a_rev_amt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |  47842.026  12469.010    3.837   0.000    23395.311   72288.742
    Debt_inj |  99550.469  48322.218    2.060   0.039     4810.026    1.94e+05
  Home_Based |  -1.28e+05  21086.004   -6.085   0.000    -1.70e+05   -8.70e+04
OO_D_educa~r |  57333.888  38667.573    1.483   0.138    -1.85e+04    1.33e+05
Comp_advan~e |    254.609  33433.325    0.008   0.994    -6.53e+04   65803.919
       _cons |  -3.00e+05   1.14e+05   -2.624   0.009    -5.25e+05   -7.60e+04
------------------------------------------------------------------------------

*Interval Regression
gen y1_f16a_rev_amt=f16a_rev_amt
gen y2_f16a_rev_amt=f16a_rev_amt

replace y1_f16a_rev_amt =1       if     f16b_rev_2004_amt_ranges_0==1
replace y1_f16a_rev_amt =501     if     f16b_rev_2004_amt_ranges_0==2
replace y1_f16a_rev_amt =1001    if     f16b_rev_2004_amt_ranges_0==3
replace y1_f16a_rev_amt =3001    if     f16b_rev_2004_amt_ranges_0==4
replace y1_f16a_rev_amt =5001    if     f16b_rev_2004_amt_ranges_0==5
replace y1_f16a_rev_amt =10001   if     f16b_rev_2004_amt_ranges_0==6
replace y1_f16a_rev_amt =25001   if     f16b_rev_2004_amt_ranges_0==7
replace y1_f16a_rev_amt =100001  if     f16b_rev_2004_amt_ranges_0==8
replace y1_f16a_rev_amt =1000001 if     f16b_rev_2004_amt_ranges_0==9
replace y2_f16a_rev_amt =500     if     f16b_rev_2004_amt_ranges_0==1
replace y2_f16a_rev_amt =1000    if     f16b_rev_2004_amt_ranges_0==2
replace y2_f16a_rev_amt =3000    if     f16b_rev_2004_amt_ranges_0==3
replace y2_f16a_rev_amt =5000    if     f16b_rev_2004_amt_ranges_0==4
replace y2_f16a_rev_amt =10000   if     f16b_rev_2004_amt_ranges_0==5
replace y2_f16a_rev_amt =25000   if     f16b_rev_2004_amt_ranges_0==6
replace y2_f16a_rev_amt =100000  if     f16b_rev_2004_amt_ranges_0==7
replace y2_f16a_rev_amt =1000000 if     f16b_rev_2004_amt_ranges_0==8
replace y2_f16a_rev_amt =.       if     f16b_rev_2004_amt_ranges_0==9
mi xeq 0 :svy: intreg     y1_f16a_rev_amt  y2_f16a_rev_amt  LnAssets    Debt_inj
Home_Based  OO_D_education_owner   Comp_advantage  ,cformat(%6.3f)  sformat(%6.3f)
nolstretch
```

```
Survey: Interval regression

Number of strata   =         6                  Number of obs      =      3749
Number of PSUs     =      3749                  Population size    = 56062.598
                                                 Design df          =      3743
                                                 F(   5,   3739)    =      9.00
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |  48635.320  12826.786     3.792   0.000    23487.149   73783.491
    Debt_inj |  97456.530  48186.705     2.022   0.043     2981.774    1.92e+05
  Home_Based |  -1.32e+05  21874.013    -6.020   0.000    -1.75e+05   -8.88e+04
OO_D_educa~r |  59976.145  39641.313     1.513   0.130    -1.77e+04    1.38e+05
Comp_advan~e |   2954.801  33736.503     0.088   0.930    -6.32e+04   69098.520
       _cons |  -3.05e+05   1.17e+05    -2.613   0.009    -5.34e+05   -7.63e+04
-------------+----------------------------------------------------------------
    /lnsigma |     14.686      0.436   33.669   0.000       13.831      15.541
-------------+----------------------------------------------------------------
       sigma |   2.39e+06   1.04e+06                      1.02e+06    5.62e+06
------------------------------------------------------------------------------
  Observation summary:         0  left-censored observations
                            3419     uncensored observations
                               9 right-censored observations
                             321       interval observations

* OLS using MI data
mi estimate:svy: reg       f16a_rev_amt    LnAssets        Debt_inj    Home_Based
OO_D_education_owner    Comp_advantage
mi estimate,cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates                    Imputations        =         5
Survey: Linear regression                        Number of obs      =      4928

Number of strata  =         6                  Population size    = 73278.441
Number of PSUs    =      4928
                                                 Average RVI        =    0.0346
                                                 Largest FMI        =    0.0313
                                                 Complete DF        =      4922
DF adjustment:   Small sample                    DF:     min        =   2234.48
                                                         avg        =   3837.94
                                                         max        =   4893.32
Model F test:       Equal FMI                    F(   5, 3561.2)    =     10.36
Within VCE type:   Linearized                    Prob > F           =    0.0000

------------------------------------------------------------------------------
f16a_rev_amt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |  43017.449   9112.627     4.721   0.000    25152.608   60882.289
    Debt_inj |   1.05e+05  53995.946     1.952   0.051     -482.946    2.11e+05
  Home_Based |  -1.49e+05  23035.461    -6.484   0.000    -1.95e+05   -1.04e+05
OO_D_educa~r |  55215.326  36713.163     1.504   0.133    -1.68e+04    1.27e+05
Comp_advan~e |  22458.542  29366.824     0.765   0.444    -3.51e+04   80035.202
       _cons |  -2.23e+05  84196.668    -2.652   0.008    -3.88e+05   -5.83e+04
------------------------------------------------------------------------------
```

Examples 6.45 Two-Limit Tobit Regression

Consider the regression model:

$$y_i = \alpha x_i + \varepsilon_i$$

where y_i is the dependent variable for the i[th] observation and y_i is bounded by 0 form below and 1 from above (double censoring). Using OLS regression could results in fitted value below zero or greater than 1. The tobit regression can limit the fitted values of to the lower and upper limits [0,1].

```
 use Cross_Sectional_Long_MI_Long_L2,clear
gen LnAssets=ln( Assets+1)
egen capital=rowtotal(Equity_AllYrs Debt_Owed )
gen tdca= Debt_Owed / capital
recode tdca (.=.a) if  Debt_Owed==.a & Equity_AllYrs==.a
recode tdca (.=0) if master!=0

keep if year==2004

mi svyset  mprid  [pweight=cswgt_final] , strata(sampleinfo_samplestrata)
egen capital_injection=rowtotal(Debt  Equity )
gen Debt_inj= Debt / capital_injection
gen Equity_inj= Equity / capital_injection

mi xeq 0:svy: tobit tdca   LnAssets  i.Home_Based  i.Have_IP OO_D_education_owner
OO_work_exp_owner  OO_race_white_owner PO_gender   ,  ll(0) ul(1) cformat(%6.3f)
sformat(%6.3f)   nolstretch

Survey: Tobit regression

Number of strata   =         6                  Number of obs      =      3773
Number of PSUs     =      3773                  Population size    =  56504.48
                                                 Design df          =      3767
                                                 F(   7,   3761)    =     18.13
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |      0.034      0.005     7.119   0.000        0.024       0.043
1.Home_Based |     -0.105      0.027    -3.886   0.000       -0.158      -0.052
   1.Have_IP |     -0.082      0.032    -2.567   0.010       -0.145      -0.019
OO_D_educa~r |     -0.091      0.026    -3.504   0.000       -0.142      -0.040
OO_work_ex~r |     -0.006      0.001    -4.589   0.000       -0.009      -0.004
OO_race_wh~r |      0.035      0.035     0.999   0.318       -0.033       0.102
   PO_gender |     -0.025      0.030    -0.831   0.406       -0.084       0.034
       _cons |      0.104      0.062     1.692   0.091       -0.017       0.225
-------------+----------------------------------------------------------------
      /sigma |      0.641      0.013    48.775   0.000        0.615       0.666
------------------------------------------------------------------------------
  Obs. summary:       1401  left-censored observations at tdca<=0
                      1963     uncensored observations
                       409 right-censored observations at tdca>=1


mi  estimate,cmdok:svy:  tobit  tdca      LnAssets    i.Home_Based    i.Have_IP
OO_D_education_owner  OO_work_exp_owner  OO_race_white_owner PO_gender   ,  ll(0)
ul(1)
mi estimate,cformat(%6.3f)  sformat(%6.3f)   nolstretch
```

```
Multiple-imputation estimates                   Imputations       =          5
Survey: Tobit regression                        Number of obs     =       4825

Number of strata  =         6                   Population size   = 71737.374
Number of PSUs    =      4825
                                                Average RVI       =     0.0311
                                                Largest FMI       =     0.1138
                                                Complete DF       =       4819
DF adjustment:   Small sample                   DF:     min       =     314.63
                                                        avg       =    2713.07
                                                        max       =    4748.20
Model F test:       Equal FMI                   F(   7, 3839.4)   =      31.16
Within VCE type:   Linearized                   Prob > F          =     0.0000

------------------------------------------------------------------------------
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LnAssets |       0.046      0.004   10.686   0.000        0.037       0.054
1.Home_Based |      -0.109      0.025   -4.315   0.000       -0.159      -0.060
   1.Have_IP |      -0.061      0.030   -2.039   0.042       -0.119      -0.002
OO_D_educa~r |      -0.086      0.024   -3.552   0.000       -0.134      -0.039
OO_work_ex~r |      -0.008      0.001   -6.111   0.000       -0.010      -0.005
OO_race_wh~r |       0.024      0.034    0.704   0.482       -0.043       0.091
   PO_gender |      -0.016      0.028   -0.584   0.559       -0.070       0.038
       _cons |      -0.062      0.059   -1.048   0.295       -0.178       0.054
-------------+----------------------------------------------------------------
      /sigma |       0.661      0.012   53.150   0.000        0.637       0.686
------------------------------------------------------------------------------
```

Examples 6.46 Instrumental Variables Regression

Consider the regression model:

$$y_{1i} = \alpha x_{1i} + \beta y_{2i} + \varepsilon_i$$

where y_i is the dependent variable for the ith observation. y_2 would be correlated with ε if:

1. Omitted variables are correlated with y_2 and y_1.
2. y_2 is measured with errors.
3. y_1 and y_2 are simultaneously determined.

Assume that y_2 is an endogenous variable correlated with ε ($Cov(y_2, \varepsilon) \neq 0$). Because we are treating y_2 as an endogenous regressor, we must have one or more additional variables available that are correlated with y_2 but uncorrelated with ε. Assume that we have an instrumental variable, z_i that is uncorrelated with ε , but is correlated with y_2. That is $Cov(z, \varepsilon) = 0$ and $Cov(y_2, \mathrm{z}) \neq 0$.

By estimating y_2 as

$$y_{2i} = \alpha x_{1i} + \delta z_i + \nu_i$$

$$y_{2i} = \hat{y}_{2i} + \nu_i$$

Because $\hat{y}_{2i}$ is estimated using all exogenous variables that are not correlated with the error term, ε, $\hat{y}_{2i}$ is not correlated with ε, while ν is correlated with ε.

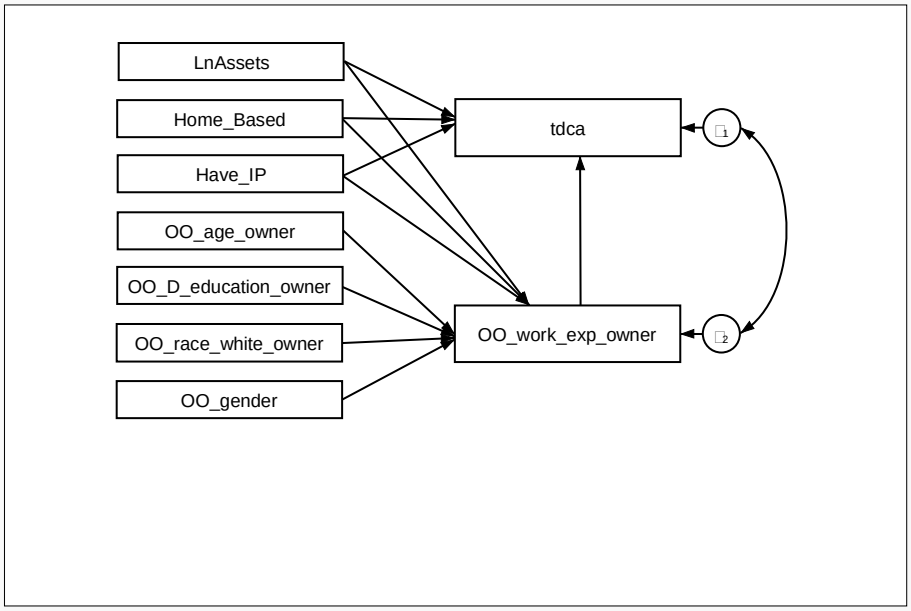
Now we can estimate y_{1i} :

$$y_{1i} = \alpha x_{1i} + \beta \hat{y}_{2i} + \varepsilon_i$$

By isolating the part of y_{2i} that is uncorrelated with ε, we solve the problem of endogeneity.

```
* Assume that OO_work_exp_owner is a choice variable ( an endogenous variable)
* Aumme that OO_age_owner OO_D_education_owner     OO_race_white_owner OO_gender
are instrumental variables


*  Fit a regression using the LIML estimator
mi estimate,cmdok:svy: ivregress liml tdca    LnAssets   i.Home_Based   i.Have_IP
(OO_work_exp_owner=OO_age_owner  OO_D_education_owner        OO_race_white_owner
OO_gender)
mi estimate,cformat(%6.3f)  sformat(%6.3f)   nolstretch

Multiple-imputation estimates
Survey: Instrumental variables (LIML) regression

                                                Imputations        =         5
                                                Number of obs      =      4928
Number of strata  =         6                   Population size    = 73278.441
Number of PSUs    =      4928
                                                Average RVI        =    0.0454
                                                Largest FMI        =    0.0761
                                                Complete DF        =      4922
DF adjustment:   Small sample                   DF:     min        =    636.46
                                                        avg        =   1468.13
                                                        max        =   2893.68
Model F test:       Equal FMI                   F(   4, 2543.2)    =     47.12
```

```
Within VCE type:   Linearized                    Prob > F          =    0.0000

------------------------------------------------------------------------------
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
OO_work_ex~r |     -0.005      0.002   -2.957   0.003       -0.008      -0.002
    LnAssets |      0.021      0.002   11.364   0.000        0.017       0.024
1.Home_Based |     -0.047      0.013   -3.658   0.000       -0.073      -0.022
   1.Have_IP |     -0.043      0.015   -2.801   0.005       -0.073      -0.013
       _cons |      0.239      0.028    8.554   0.000        0.184       0.294
------------------------------------------------------------------------------


*      Fit a regression via 2SLS
mi estimate,cmdok:svy: ivregress 2sls  tdca   LnAssets  i.Home_Based  i.Have_IP
(OO_work_exp_owner=OO_age_owner  OO_D_education_owner        OO_race_white_owner
OO_gender)
mi estimate,cformat(%6.3f)  sformat(%6.3f)   nolstretch



Multiple-imputation estimates
Survey: Instrumental variables (2SLS) regression

                                                 Imputations        =         5
                                                 Number of obs      =      4928
Number of strata  =         6                    Population size    = 73278.441
Number of PSUs    =      4928
                                                 Average RVI        =    0.0454
                                                 Largest FMI        =    0.0768
                                                 Complete DF        =      4922
DF adjustment:   Small sample                    DF:     min        =    626.08
                                                         avg        =   1464.13
                                                         max        =   2881.64
Model F test:       Equal FMI                    F(   4, 2542.1)    =     47.17
Within VCE type:   Linearized                    Prob > F           =    0.0000

------------------------------------------------------------------------------
        tdca |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
OO_work_ex~r |     -0.005      0.002   -2.990   0.003       -0.008      -0.002
    LnAssets |      0.021      0.002   11.364   0.000        0.017       0.024
1.Home_Based |     -0.047      0.013   -3.658   0.000       -0.073      -0.022
   1.Have_IP |     -0.043      0.015   -2.800   0.005       -0.073      -0.013
       _cons |      0.239      0.028    8.599   0.000        0.184       0.294
------------------------------------------------------------------------------

*      Fit a regression via SEM
mi  estimate,cmdok:svy:sem  (OO_work_exp_owner<-OO_age_owner  OO_D_education_owner
OO_race_white_owner  OO_gender  LnAssets      Home_Based      Have_IP)  (tdca<-
OO_work_exp_owner LnAssets  ///
  Home_Based   Have_IP), cov(e.OO_work_exp_owner e.tdca)  nocapslatent
mi estimate,cformat(%6.3f)  sformat(%6.3f)   nolstretch
/*you obtain slightly different results from those you would obtain with ivregress
liml. This is because sem with default method(ml) produces full-information maximum
likelihood rather than limited-information maximum likelihood results. */
```




```
Multiple-imputation estimates                   Imputations       =          5
Survey: Structural equation model               Number of obs     =       4928

Number of strata  =         6                   Population size   = 73278.441
Number of PSUs    =      4928
                                                Average RVI       =     0.0192
                                                Largest FMI       =     0.0843
                                                Complete DF       =       4922
DF adjustment:   Small sample                   DF:     min       =     534.70
                                                        avg       =    3471.92
Within VCE type:   Linearized                           max       =    4909.73

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  OO_wo~r <- |
  OO_age_o~r |      0.323      0.017   19.495   0.000        0.291       0.355
  OO_D_edu~r |     -0.785      0.305   -2.576   0.010       -1.382      -0.188
  O~te_owner |      0.509      0.392    1.299   0.194       -0.259       1.278
  OO_gende~r |      4.496      0.368   12.228   0.000        3.775       5.216
    LnAssets |     -0.006      0.047   -0.138   0.890       -0.098       0.085
  Home_Based |     -0.450      0.308   -1.458   0.145       -1.054       0.155
     Have_IP |     -0.307      0.362   -0.849   0.396       -1.017       0.402
       _cons |     -5.680      0.851   -6.675   0.000       -7.349      -4.012
  -----------+----------------------------------------------------------------
  tdca <-    |
  OO_work_~r |     -0.004      0.001   -6.074   0.000       -0.005      -0.002
    LnAssets |      0.021      0.002   11.326   0.000        0.017       0.024
  Home_Based |     -0.047      0.013   -3.642   0.000       -0.072      -0.022
     Have_IP |     -0.043      0.015   -2.774   0.006       -0.073      -0.012
       _cons |      0.228      0.022   10.485   0.000        0.185       0.270
-------------+----------------------------------------------------------------
Variance     |
  e.OO_wor~r |     82.470      1.881                       78.864      86.241
      e.tdca |      0.136      0.002                        0.131       0.140
------------------------------------------------------------------------------
```

Aday, L.A. and Llewellyn, J.C. 2006. Designing and Conducting HealthSurveys: A Comprehensive Guide, 3rd Ed. San Francisco, CA: Jossey Bass.

Cochran, W.G. 1977. Sampling Techniques, 3rd Ed. New York, NY: John Wiley and Sons.

Haviland, Amelia and Savych, Bogdan, A Description and Analysis of Evolving Data Resources on Small Business (September 2007). RAND Corporation Working Paper No. WR-293-1-ICJ.

Kish, L. 1965. Survey Sampling. New York: John Wiley and Sons.

Kish, L. 1987. Statistical Design for Research. New York: John Wiley & Sons, Inc.

Korn, E. L., and Graubard, B. I. 1995, "Examples of differing weighted and unweighted estimates from a sample survey," The American Statistician, 49, 291-295.

Lee, E. S., and R. N. Forthofer. 2005. Analyzing Complex Survey Data. 2nd ed. Thousand Oaks, CA: Sage.

Lohr, S. L. 2010. Sampling: Design and Analysis, Second Edition, Boston: Brooks/Cole.

Marsden, Peter V., and James D. Wright (eds). 2010. Handbook of Survey Research (second edition). Bingley, UK: Emerald Publishing Group.

Pfeffermann, D. 1993. The Role of Sampling Weights When Modeling Survey Data. International Statistical Review, 61, 317-337.

Pfeffermann, D. and Holmes, D. 1985. Robustness Considerations in the Choice of a Method of Inference for Regression Analysis of Survey Data. Journal of the Royal Statistical Society, Series A, 198, 268-278.

West, B.T., Berglund, P., and Heeringa, S.G. 2008. A Closer Examination of Subpopulation Analysis of Complex Sample Survey Data. The Stata Journal, 8(3), 1-12.