10

Survival and Event-Count Models

This chapter presents methods for analyzing event data. *Survival analysis* involves several related techniques that focus on times until the event of interest occurs. Although the event could be good or bad, by convention we refer to the event as a "failure." The time until failure is "survival time." Survival analysis is important in biomedical research, but it can be applied equally well to other fields from engineering to social science — for example, in modeling the time until an unemployed person gets a job, or a single person gets married. Stata offers a full range of survival analysis procedures, a few of which are illustrated in this chapter.

We also look briefly at Poisson regression and its relatives. These methods focus not on survival times but, rather, on the rates or counts of events over a specified interval of time. Event-count methods include Poisson regression and negative binomial regression. Such models can be fit either through specialized commands or through the broader approach of generalized linear modeling (GLM).

Consult the *Survival Analysis and Epidemiological Tables Reference Manual* for more information about Stata's capabilities. Type **help st** to see an online overview. Selvin (2004, 2008) provides well-illustrated introductions to survival analysis and Poisson regression. I have borrowed (with permission) several of his examples. Other good introductions to survival analysis include the Stata-oriented volume by Cleves et al. (2010), a chapter in Rosner (1995), and comprehensive treatments by Hosmer, Lemeshow and May (2008) and Lee (1992). McCullagh and Nelder (1989) describe generalized linear models. Long (1997) has a chapter on regression models for count data (including Poisson and negative binomial), and also some material on generalized linear models. An extensive and current treatment of generalized linear models is found in Hardin and Hilbe (2012).

Stata menu groups most relevant to this chapter include:

Statistics > Survival analysis

Graphics > Survival analysis graphs

Statistics > Count outcomes

Statistics > Generalized linear models

Regarding epidemiological tables, not covered in this chapter, further information can be found by typing **help epitab** or exploring the menus for

Statistics > Epidemiology and related

Example Commands

Most of Stata's survival-analysis (**st***) commands require that the data have previously been identified as survival-time by an **stset** command. **stset** need only be run once, and the data subsequently saved.

. **stset** *timevar*, **failure(***failvar***)**
Identifies single-record survival-time data. Variable *timevar* indicates the time elapsed before either a particular event (called a "failure") occurred, or the period of observation ended ("censoring"). Variable *failvar* indicates whether a failure (*failvar* = 1) or censoring (*failvar* = 0) occurred at *timevar*. The dataset contains only one record per individual. The dataset must be **stset** before any further **st*** commands will work. If we subsequently **save** the dataset, however, the **stset** definitions are saved as well. **stset** creates new variables named *_st*, *_d*, *_t* and *_t0* that encode information necessary for subsequent **st*** commands.

. **stset** *timevar*, **failure(***failvar***) id(***patient***) enter(time** *start***)**
Identifies multiple-record survival-time data. In this example, the variable *timevar* indicates elapsed time before failure or censoring; *failvar* indicates whether failure (1) or censoring (0) occurred at this time. *patient* is an identification number. The same individual might contribute more than one record to the data, but always has the same identification number. *start* records the time when each individual came under observation.

. **stdescribe**
Describes survival-time data, listing definitions set by **stset** and other characteristics of the data.

. **stsum**
Obtains summary statistics: the total time at risk, incidence rate, number of subjects, and percentiles of survival time.

. **ctset** *time nfail ncensor nenter*, **by(***ethnic sex***)**
Identifies count-time data. In this example, the variable *time* is a measure of time; *nfail* is the number of failures occurring at *time*. We also specified *ncensor* (number of censored observations at *time*) and *nenter* (number entering at *time*), although these can be optional. *ethnic* and *sex* are other categorical variables defining observations in these data.

. **cttost**
Converts count-time data, previously identified by the **ctset** command, into survival-time form that can be analyzed by **st*** commands.

. **sts graph**
Graphs the Kaplan–Meier survivor function. To visually compare two or more survivor functions, such as one for each value of the categorical variable *sex*, use a **by( )** option such as **sts graph, by(***sex***)**. To adjust, through Cox regression, for the effects of a continuous independent variable such as *age*, use an **adjustfor( )** option such as **sts graph, by(***sex***) adjustfor(***age***)**. The **by( )** and **adjustfor( )** options work similarly with the **sts list** and **sts generate** commands.

. **sts list**
Lists the estimated Kaplan–Meier survivor (or failure) function.

. **sts test** *sex*
Tests the equality of the Kaplan–Meier survivor function across categories of *sex*.

. **sts generate** *survfunc* **= S**
Creates a new variable arbitrarily named *survfunc*, containing the estimated Kaplan–Meier survivor function.

. **stcox** *x1 x2 x3*
Fits a Cox proportional hazard model, regressing time to failure on continuous or dummy variable predictors *x1*, *x2* and *x3*.

. **stcox** *x1 x2 x3*, **strata(***x4***) vce(robust)**
. **predict** *hazard*, **basechazard**
Fits a Cox proportional hazard model, stratified by *x4*. The **vce(robust)** option requests robust standard error estimates. See Chapter 8, or for a more complete explanation of robust standard errors, consult the *User's Guide*. The **predict** command stores the group-specific baseline cumulative hazard function as a new variable named *hazard*; type **help stcox postestimation** for more options.

. **stphplot, by(***sex***)**
Plots –ln(–ln(survival)) versus ln(analysis time) for each level of the categorical variable *sex*, from the previous **stcox** model. Roughly parallel curves support the Cox model assumption that the hazard ratio does not change with time. Other checks on the Cox assumptions are performed by the commands **stcoxkm** (compares Cox predicted curves with Kaplan–Meier observed survival curves) and **estat phtest** (performs test based on Schoenfeld residuals). See **help stcox diagnostics** for syntax and options.

. **streg** *x1 x2*, **dist(weibull)**
Fits Weibull-distribution model regression of time-to-failure on continuous or dummy variable predictors *x1* and *x2*.

. **streg** *x1 x2 x3 x4*, **dist(exponential) vce(robust)**
Fits exponential-distribution model regression of time-to-failure on continuous or dummy predictors *x1*–*x4*. Obtains heteroskedasticity-robust standard error estimates. In addition to Weibull and exponential, other **dist( )** specifications for **streg** include lognormal, log-logistic, Gompertz or generalized gamma distributions. Type **help streg** for more information.

```
. stcurve, survival
```
After **streg**, plots the survival function from this model at mean values of all the *x* variables.

```
. stcurve, cumhaz at(x3=50, x4=0)
```
After **streg**, plots the cumulative hazard function from this model at mean values of *x1* and *x2*, *x3* set at 50, and *x4* set at 0.

```
. poisson count x1 x2 x3, irr exposure(x4)
```
Performs Poisson regression of event-count variable *count* (assumed to follow a Poisson distribution) on continuous or dummy independent variables *x1–x3*. Independent-variable effects will be reported as incidence rate ratios (**irr**). The **exposure( )** option identifies a variable indicating the amount of exposure, if this is not the same for all observations.
Note: A Poisson model assumes that the event probability remains constant, regardless of how many times an event occurs for each observation. If the probability does not remain constant, we should consider using **nbreg** (negative binomial regression) or **gnbreg** (generalized negative binomial regression) instead.

```
. glm count x1 x2 x3, link(log) family(poisson) exposure(x4) eform
```
Performs the same regression specified in the **poisson** example above, but as a generalized linear model (GLM). **glm** can fit Poisson, negative binomial, logit and many other types of models, depending on what **link( )** (link function) and **family( )** (distribution family) options we employ.

Survival-Time Data

Survival-time data contain, at a minimum, one variable measuring how much time elapsed before a certain event occurred for each observation. The literature often terms this event of interest a "failure," regardless of its real-world meaning. When failure has not occurred to an observation by the time data collection ends, that observation is said to be "censored." The **stset** command sets up a dataset for survival-time analysis by identifying which variable measures time and (if necessary) which variable is a {0, 1} indicator for whether the observation failed or was censored. The dataset can also contain any number of other measurement or categorical variables. Individuals (for example, medical patients) can be represented by more than one observation.

To illustrate the use of **stset**, we will begin with an example from Selvin (1995:453) concerning 51 individuals diagnosed with HIV. The data initially reside in a raw-data file (*aids.raw*) that looks like this:

```
 1   1   1   34
 2   17  1   42
 3   37  0   47
 (rows 4–50 omitted)
51   81  0   29
```

The first column values are case numbers (1, 2, 3, . . . , 51). The second column tells how many months elapsed after the diagnosis, before that person either developed symptoms of AIDS or the study ended (1, 17, 37, . . .). The third column holds a 1 if the individual developed AIDS

symptoms (failure), or a 0 if no symptoms had appeared by the end of the study (censoring). The last column reports the individual's age at the time of diagnosis.

We can read the raw data into memory using **infile**, then label the variables and data:

```
. infile case time aids age using C:\data\aids.raw, clear
. label variable case "Case ID number"
. label variable time "Months since HIV diagnosis"
. label variable aids "Developed AIDS symptoms"
. label variable age "Age in years"
. label data "AIDS (Selvin 1995:453)"
. compress
```

The next step is to identify which variable measures time and which indicates failure/censoring. Although not necessary with these single-record data, we can also note which variable holds individual case identification numbers. In an **stset** command, the first-named variable measures time. Subsequently, we identify with **failure( )** the dummy representing whether an observation failed (1) or was censored (0). After using **stset**, we save the dataset in Stata format to preserve this information.

```
. stset time, failure(aids) id(case)

                id:  case
     failure event:  aids != 0 & aids < .
obs. time interval:  (time[_n-1], time]
 exit on or before:  failure

------------------------------------------------------------------------------
        51  total obs.
         0  exclusions
------------------------------------------------------------------------------
        51  obs. remaining, representing
        51  subjects
        25  failures in single failure-per-subject data
      3164  total analysis time at risk, at risk from t =         0
                            earliest observed entry t =         0
                                 last observed exit t =        97

. save aids.dta, replace
```

stdescribe yields a brief description of how our survival-time data are structured. In this simple example we have only one record per subject, so some of this information is unneeded.

```
. stdescribe
```

```
        failure _d:  aids
  analysis time _t:  time
                id:  case
```

Category	total	per subject: mean	min	median	max
no. of subjects	51				
no. of records	51	1	1	1	1
(first) entry time		0	0	0	0
(final) exit time		62.03922	1	67	97
subjects with gap	0				
time on gap if gap	0	.	.	.	.
time at risk	3164	62.03922	1	67	97
failures	25	.4901961	0	0	1

The **stsum** command obtains summary statistics. We have 25 failures out of 3,164 person-months, giving an incidence rate of 25/3164 = .0079014. The percentiles of survival time derive from a Kaplan–Meier survivor function (next section). This function estimates about a 25% chance of developing AIDS within 41 months after diagnosis, and 50% within 81 months. Over the observed range of the data (up to 97 months) the probability of AIDS does not reach 75%, so there is no 75th percentile given.

```
. stsum
```

```
        failure _d:  aids
  analysis time _t:  time
                id:  case
```

	time at risk	incidence rate	no. of subjects	Survival time: 25%	50%	75%
total	3164	.0079014	51	41	81	.

If the data happen to include a grouping or categorical variable such as *sex*, we could obtain summary statistics on survival time separately for each group by a command of the following form:

```
. stsum, by(sex)
```

Later sections describe more formal methods for comparing survival times from two or more groups.

Count-Time Data

Survival-time (**st**) datasets like *aids.dta* contain information on individual people or things, with variables indicating the time at which failure or censoring occurred for each individual. A different type of dataset called count-time (**ct**) contains aggregate data, with variables counting the number of individuals that failed or were censored at time *t*. For example, *diskdriv.dta* contains hypothetical test information on 25 disk drives. All but 5 drives failed before testing ended at 1,200 hours.


```
. use C:\data\diskdriv.dta, clear
. describe

Contains data from C:\data\diskdriv.dta
  obs:             6                          Count-time data on disk drives
 vars:             3                          30 Jun 2012 10:19
 size:            24
```

variable name	storage type	display format	value label	variable label
hours	int	%8.0g		Hours of continuous operation
failures	byte	%8.0g		Number of failures observed
censored	byte	%9.0g		Number still working

```
Sorted by:

. list
```

	hours	failures	censored
1.	200	2	0
2.	400	3	0
3.	600	4	0
4.	800	8	0
5.	1000	3	0
6.	1200	0	5

To set up a count-time dataset, we specify the time variable, the number-of-failures variable, and the number-censored variable, in that order. After **ctset**, the **cttost** command automatically converts our count-time data to survival-time format.

```
. ctset hours failures censored

    dataset name:  C:\data\diskdriv.dta
            time:  hours
       no. fail:  failures
       no. lost:  censored
      no. enter:  --                    (meaning all enter at time 0)

. cttost

     failure event:  failures != 0 & failures < .
obs. time interval:  (0, hours]
 exit on or before:  failure
           weight:  [fweight=w]

------------------------------------------------------------------------------
        6  total obs.
        0  exclusions
------------------------------------------------------------------------------
        6  physical obs. remaining, equal to
       25  weighted obs., representing
       20  failures in single record/single failure data
    19400  total analysis time at risk, at risk from t =         0
                                earliest observed entry t =         0
                                     last observed exit t =      1200

. list
```

	hours	failures	censored	w	_st	_d	_t	_t0
1.	200	1	0	2	1	1	200	0
2.	400	1	0	3	1	1	400	0
3.	600	1	0	4	1	1	600	0
4.	800	1	0	8	1	1	800	0
5.	1000	1	0	3	1	1	1000	0
6.	1200	0	5	5	1	0	1200	0

```
. stdescribe

        failure _d:  failures
  analysis time _t:  hours
            weight:  [fweight=w]
```

Category	unweighted total	per subject: unweighted mean	min	unweighted median	max
no. of subjects	6				
no. of records	6	1	1	1	1
(first) entry time		0	0	0	0
(final) exit time		700	200	700	1200
subjects with gap	0				
time on gap if gap	0				
time at risk	4200	700	200	700	1200
failures	5	.8333333	0	1	1

The **cttost** command defines a set of frequency weights, *w*, in the resulting **st**-format dataset. **st*** commands automatically recognize and use these weights in any survival-time analysis, so the data now are viewed as containing 25 observations (25 disk drives) instead of the previous 6 (six time periods).

```
. stsum

        failure _d:  failures
  analysis time _t:  hours
            weight:  [fweight=w]
```

	time at risk	incidence rate	no. of subjects	Survival time: 25%	50%	75%
total	19400	.0010309	25	600	800	1000

Kaplan–Meier Survivor Functions

Let n_j represent the number of observations that have not failed, and are not censored, at the beginning of time period t. d_j represents the number of failures that occur to these observations during time period t. The Kaplan–Meier estimator of surviving beyond time t is the product of survival probabilities in t and the preceding periods:

$$S(t) = \prod_{j=0}^{t} \{ (n_j - d_j) / n_j \} \qquad [10.1]$$

For example, in the AIDS data seen earlier, one of the 51 individuals developed symptoms only one month after diagnosis. No observations were censored this early, so the probability of "surviving" (meaning, not developing AIDS) beyond *time* = 1 is

$$S(1) = (51 - 1) / 51 = .9804$$

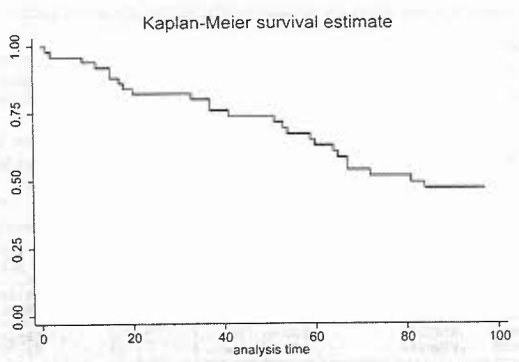
A second patient developed symptoms at *time* = 2, and a third at *time* = 9:

$$S(2) = .9804 \times (50 - 1) / 50 = .9608$$
$$S(9) = .9608 \times (49 - 1) / 49 = .9412$$

Graphing $S(t)$ against t produces a Kaplan–Meier survivor curve, like the one seen in Figure 10.1. Stata draws such graphs automatically with the **sts graph** command. For example,

```
. use C:\data\aids, clear
. sts graph
```

Figure 10.1



For a second example of survivor functions, we turn to data in *smoking1.dta*, adapted from Rosner (1995). The observations are data on 234 former smokers, attempting to quit. Most did not succeed. Variable *days* records how many days elapsed between quitting and starting up again. The study lasted one year, and variable *smoking* indicates whether an individual resumed smoking before the end of this study (*smoking* = 1, "failure") or not (*smoking* = 0, "censored"). With new data, we should begin by using **stset** to set the data up for survival-time analysis.

```
. use C:\data\smoking1.dta, clear
. describe
```

```
Contains data from C:\data\smoking1.dta
  obs:           234                          Smoking (Rosner 1995:607)
 vars:             8                          30 Jun 2012 10:19
 size:         2,808
-------------------------------------------------------------------------
              storage  display    value
variable name   type   format     label      variable label
-------------------------------------------------------------------------
id              int    %9.0g                 Case ID number
days            int    %9.0g                 Days abstinent
smoking         byte   %9.0g                 Resumed smoking
age             byte   %9.0g                 Age in years
sex             byte   %9.0g      sex        Sex (female)
cigs            byte   %9.0g                 Cigarettes per day
co              int    %9.0g                 Carbon monoxide x 10
minutes         int    %9.0g                 Minutes elapsed since last cig
-------------------------------------------------------------------------
Sorted by:
```

```
. stset days, failure(smoking)

     failure event:  smoking != 0 & smoking < .
obs. time interval:  (0, days]
 exit on or before:  failure

-------------------------------------------------------------------------
      234  total obs.
        0  exclusions
-------------------------------------------------------------------------
      234  obs. remaining, representing
      201  failures in single record/single failure data
    18946  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =       366
```

The study involved 110 men and 124 women. Incidence rates for both sexes appear to be similar:

```
. stsum, by(sex)

         failure _d:  smoking
   analysis time _t:  days

         |               incidence       no. of    |------ Survival time -----|
sex      | time at risk     rate        subjects        25%       50%       75%
---------+---------------------------------------------------------------------
    Male |         8813   .0105526           110          4        15        68
  Female |        10133   .0106582           124          4        15        83
---------+---------------------------------------------------------------------
   total |        18946   .0106091           234          4        15        73
```

Figure 10.2 confirms this similarity. There appears to be little difference between the survivor functions of men and women. That is, both sexes returned to smoking at about the same rate. The survival probabilities of nonsmokers decline very steeply during the first 30 days after quitting. For either sex, there is less than a 15% chance of surviving as a nonsmoker beyond a full year.

```
. sts graph, by(sex) plot1opt(lwidth(medium)) plot2opt(lwidth(thick))
```

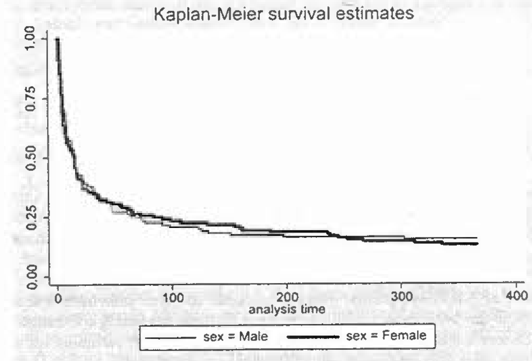



Figure 10.2

We can also formally test for the equality of survivor functions using a log-rank test. Unsurprisingly, this test finds no significant difference (p = .6772) between the smoking recidivism of men and women.

```
. sts test sex

         failure _d:  smoking
   analysis time _t:  days

Log-rank test for equality of survivor functions

       |   Events         Events
sex    |   observed       expected
-------+-------------------------
Male   |         93          95.88
Female |        108         105.12
-------+-------------------------
Total  |        201         201.00

             chi2(1) =       0.17
             Pr>chi2 =     0.6772
```

Cox Proportional Hazard Models

Regression methods allow us to take survival analysis further and examine the effects of multiple continuous or categorical predictors. One widely-used method known as Cox regression employs a proportional hazard model. The hazard rate for failure at time t is defined as the rate of failures at time t among those who have survived to time t:

$$h(t) = \frac{\text{probability of failing between times } t \text{ and } t + \Delta t}{(\Delta t)\,(\text{probability of failing after time } t)} \qquad [10.2]$$