# 5

# Survival Data Analysis

This chapter constitutes an introduction to the modeling of survival data with Stata. After describing the principle of the organization and representation of survival data in Stata, we will focus on the estimation of the survival function with the Kaplan–Meier method and Cox's regression model, which will be then discussed in further detail.

In this chapter, the illustrations are achieved based on a different dataset than the one used so far. The data are available in the R `survival` package and are exported in CSV format. Missing data are coded as a dot (instead of the value `NA`, employed by default in R). They can be imported into Stata as follows:

```
. insheet using "lung.csv", clear
(10 vars, 228 obs)
. label define gender 1 "Male" 2 "Female"
. label values sex gender
```

These consist of data on the survival of patients with lung cancer. The variables of interest are the following: `time` is the survival time in days, `status` is the status at the point date (1 = censored data, 2 = deceased patient), `age` is the patient's age and `sex` is the patient's gender (1 = male, 2 = female). In total, there are 228 patients.

## 5.1. Data representation and descriptive statistics

### 5.1.1. *Survival data representation format*

The representation and the modeling of such data rely on a particular class of Stata commands due to the nature of survival data. These are represented by means of a variable encoding a duration and another variable encoding an event of interest (death, relapse, failure, etc.). This namely consists of the `sts` command that includes several subcommands (`list`, `test`, `graph`, mostly) as well as a few associated commands whose prefix is `st`.

The `stset` command is responsible for creating several auxiliary variables (whose name usually begins with _) that are not directly manipulated but that Stata employs transparently. The same process is used by Stata in the case of data from surveys or from time series data. The `stset` command requires that the variable defining the time durations be indicated (independent of the unit of measure) and with the option `failure()` which represents events encoding. By default, Stata considers that all the non-null and non-missing values signal the event of interest (for instance, death). With $0/1$ values, this does not create any problems, assuming that a value of 1 indicates the event.

It can happen, as is the case here, that the death or the event be represented by another value and that censoring is not coded with 0. In this case, it is necessary to specify the value of the death in the option `failure()` as shown below:

```
. stset time, failure(status=2)

     failure event:  status == 2
obs. time interval:  (0, time]
 exit on or before:  failure

------------------------------------------------------------------------------
      228  total obs.
        0  exclusions
------------------------------------------------------------------------------
      228  obs. remaining, representing
      165  failures in single record/single failure data
    69593  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =      1022
```

## 5.2. Descriptive statistics

Here, follows an overview of the raw data, simply obtained with the `list` command:

```
. list time status age sex in 1/5

     +----------------------------+
     | time   status   age    sex |
     |----------------------------|
```

```
  1. |  306        2    74    Male |
  2. |  455        2    68    Male |
  3. | 1010        1    56    Male |
  4. |  210        2    57    Male |
  5. |  883        2    60    Male |
     +---------------------------+
```

It is naturally possible to make use of the group of Stata commands presented in the previous chapters to achieve the numerical summary of variables. However, care should be taken due to the fact that in this case the data are processed independently of the presence of censoring:

```
. tabulate status, summarize(time)

            |          Summary of time
     status |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          1 |   363.46032   221.13635          63
          2 |         283   202.80508         165
------------+------------------------------------
      Total |   305.23246   210.64554         228
```

## 5.3. Survival function and Kaplan–Meier curve

### 5.3.1. *Mortality table*

In order to build the mortality table and display the probabilities of survival over time, the `sts list` command will be inserted. Without any other option, Stata displays all temporal events:

```
. sts list
```

However, it is possible to limit the display to certain specific values. To display, for example, the survival probability associated with times 200–300, we could write:

```
. sts list, at(200 300) enter

         failure _d:  status == 2
   analysis time _t:  time

              Beg.                      Survivor      Std.
    Time     Total     Fail             Function     Error     [95% Conf. Int.]
```

```
-------------------------------------------------------------------------------
     200        145        72                0.6803     0.0311      0.6149     0.7369
     300         92        29                0.5306     0.0346      0.4605     0.5958
-------------------------------------------------------------------------------
Note:  survivor function is calculated over full data and evaluated at
       indicated times; it is not calculated from aggregates shown at left.
```

The survival median and its 95% confidence interval can be obtained with the `stci` command:

```
. stci, dd(2) noshow

             |    no. of
             |  subjects         50%     Std. Err.     [95% Conf. Interval]
-------------+-------------------------------------------------------------
       total |       228      310.00         21.77           284         361
```

The `noshow` option does not allow for displaying reminders concerning the variables being used (here, `time` and `status`). Regarding the `dd(2)` option, it sets the limit of the display to two decimal places. If it is desirable to obtain the 10th percentile rather than the median, `p(10)` will have to be specified in option:

```
. stci, p(10) dd(2)

         failure _d:  status == 2
   analysis time _t:  time

             |    no. of
             |  subjects         10%     Std. Err.     [95% Conf. Interval]
-------------+-------------------------------------------------------------
       total |       228       79.00         14.94            54         105
```

This command can also be employed when there are several groups and when their survival median have to be compared. In this case, the classification variable will be indicated by using the `by()` option:

```
. stci, by(sex) noshow

             |    no. of
sex          |  subjects         50%     Std. Err.     [95% Conf. Interval]
-------------+-------------------------------------------------------------
        Male |       138         270      26.78831           210         306
      Female |        90         426      44.20601           345         524
-------------+-------------------------------------------------------------
       total |       228         310      21.77251           284         361
```

### 5.3.2. *Kaplan–Meier curve*

The `sts graph` command makes it possible to graphically represent the survival curve of one or more samples. In the case of several samples, the classification criterion is indicated by means of the `by()` option. The basic syntax is therefore (Figure 5.1):

```
. sts graph
```

Note that we can simply type `sts`.

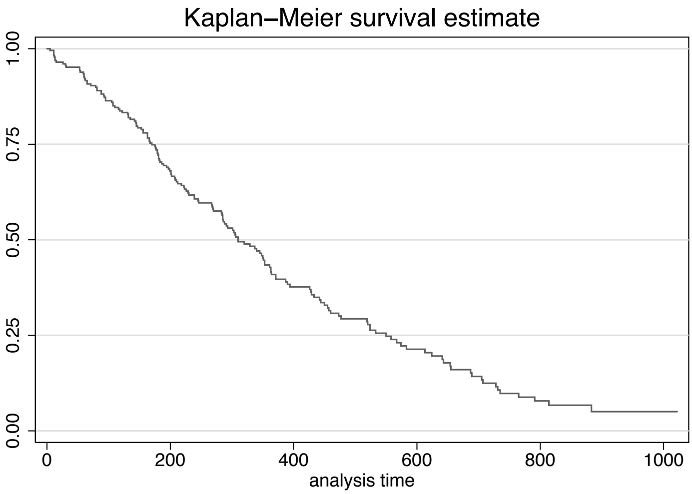


**Figure 5.1.** *Kaplan–Meier survival curve*

The `ci` option will be added to display confidence intervals (for each time value). Here is an example of its use including other options such as the simultaneous display of the number of individuals at risk over time (by default, Stata employs the same time coordinates than those displayed for the x-axis but this can be modified). Another interesting option is `censored()` (it can be abbreviated into `cen()`) that overlays on the survival curve the observed censored data (Figure 5.2):

```
. sts graph, noshow ci risktable censored(single)
```

In the case of two samples, we will include the classification factor via the option `by()`. The rest of the options is applicable. In Figure 5.3, the position of the caption has been modified such that it appear inside the graph and not in the lower margin of the chart:

```
. sts graph, by(sex) legend(ring(0) position(2))

         failure _d:  status == 2
   analysis time _t:  time
```

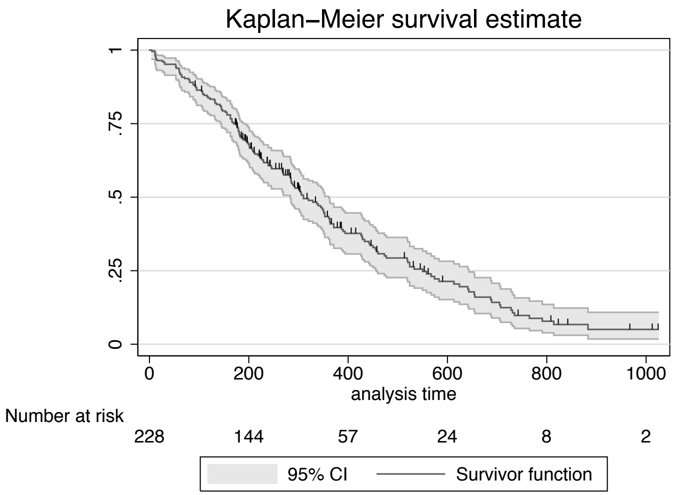


**Figure 5.2.** *Kaplan–Meier survival curve with confidence intervals and number of individuals at risk*

## 5.3.3. *Cumulative hazard function*

If we are willing to work with the cumulative hazard function (most often denoted $H(t)$), it suffices to add the `cumhaz` option when the `sts graph` command is employed:

```
. sts graph, noshow cumhaz ci
```

## 5.3.4. *Survival functions equality test*

The `sts list` command provides the mortality table and the estimated survival values for each time. The `by()` option makes it possible to calculate the survival function for two or more groups of individuals. However, it is also possible to couple this `by()` option to the `compare` option to directly display the survival estimated in each of the groups side by side:

```
. sts list, by(sex) compare noshow

                  Survivor Function
sex                 Male       Female
----------------------------------
time        5     1.0000      0.9889
          132     0.7681      0.8993
          259     0.5192      0.7186
          386     0.3265      0.5089
          513     0.2232      0.4110
          640     0.1228      0.3433
          767     0.0781      0.0832
          894     0.0357      0.0832
         1021     0.0357           .
         1148          .           .
----------------------------------
```

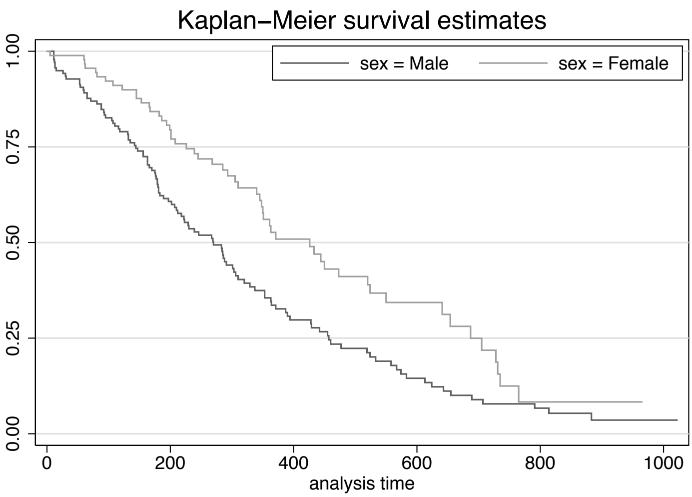


**Figure 5.3.** *Kaplan–Meier survival curve for two samples*

To perform a log-rank test (equality of the survival functions), we will include the command `sts test` simply indicating the variable defining the groups that have to be compared:

```
. sts test sex, noshow
```

```
Log-rank test for equality of survivor functions
------------------------------------------------

       |   Events         Events
sex    |  observed       expected
-------+-------------------------
Male   |       112          91.58
Female |        53          73.42
-------+-------------------------
Total  |       165         165.00

             chi2(1) =      10.33
             Pr>chi2 =     0.0013
```

If we would rather perform a Wilcoxon test, the option `wilcoxon` will be added as follows:

```
. sts test sex, wilcoxon noshow

Wilcoxon (Breslow) test for equality of survivor functions
----------------------------------------------------------

       |   Events         Events        Sum of
sex    |  observed       expected        ranks
-------+--------------------------------------
Male   |       112          91.58         3148
Female |        53          73.42        -3148
-------+--------------------------------------
Total  |       165         165.00            0

             chi2(1) =      12.47
             Pr>chi2 =     0.0004
```

## 5.4. Cox regression

The command to perform a Cox regression is `stcox`. Its use is substantially identical to that of the regression commands seen in previous chapters, except that it is not necessary to specify a response variable: as for the other `sts` commands, Stata transparently manages the time/event representation. It will therefore suffice to indicate the list of explanatory variables after the name of the command. Here is an example of how to use it considering the variable `sex` only:

```
. stcox sex, noshow
```

```
Iteration 0:   log likelihood = -750.12202
Iteration 1:   log likelihood = -744.83027
Iteration 2:   log likelihood = -744.81818
Iteration 3:   log likelihood = -744.81818
Refining estimates:
Iteration 0:   log likelihood = -744.81818

Cox regression -- Breslow method for ties

No. of subjects =          228                     Number of obs   =       228
No. of failures =          165
Time at risk    =        69593
                                                   LR chi2(1)      =     10.61
Log likelihood  =   -744.81818                     Prob > chi2     =    0.0011

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .5883716   .0983645    -3.17   0.002     .4239817    .8165002
------------------------------------------------------------------------------
```

To indicate the presence of a stratification factor, the `strata()` option will be entered as follows:

```
. stcox age, strata(sex) noshow

Iteration 0:   log likelihood = -643.61669
Iteration 1:   log likelihood = -642.03076
Iteration 2:   log likelihood = -642.02946
Refining estimates:
Iteration 0:   log likelihood = -642.02946

Stratified Cox regr. -- Breslow method for ties

No. of subjects =          228                     Number of obs   =       228
No. of failures =          165
Time at risk    =        69593
                                                   LR chi2(1)      =      3.17
Log likelihood  =   -642.02946                     Prob > chi2     =    0.0748

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
         age |   1.016324   .0093351     1.76   0.078      .998191    1.034786
------------------------------------------------------------------------------
                                                             Stratified by sex
```

Note that it is possible to change how Stata addresses tied observations, by specifying the `efron` option for example.

If we want to display the regression coefficients rather than the hazard ratio, the `nohr` option ought to be specified:

```
. stcox sex, noshow nolog nohr

Cox regression -- Breslow method for ties

No. of subjects =          228                     Number of obs   =       228
No. of failures =          165
Time at risk    =        69593
                                                   LR chi2(1)      =     10.61
Log likelihood  =   -744.81818                     Prob > chi2     =    0.0011

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -.5303966   .1671808    -3.17   0.002     -.858065   -.2027282
------------------------------------------------------------------------------
```

The `noshow` and `nolog` options are used to suppress the display of survival variables and the iterations for the convergence of the model.

## 5.5. Key points

– The representation of censored data is carried out via the command `stset` that allows the variables encoding times and events to be defined.

– A set of subcommands is associated with `sts` and other commands are prefixed by `st` (for example `stci`).

– The construction of the Cox regression model follows the same principle as in the case of linear or logistic regression, and postestimation commands enable additional information to be provided (predicted values, goodness of fit of the model, etc.).

## 5.6. Further reading

The book by Cleves *et al.* [CLE 10] definitely remains the reference book for processing survival data with Stata. For a more in-depth coverage of survival analysis, see Royston and Lambert's work [ROY 11].

## 5.7. Applications

1) In a placebo-controlled trial on biliary cirrhosis, D-penicillamine (DPCA) has been introduced in the active arm in a cohort of 312 patients. In total, 154 patients have been randomized in the active arm (variable treatment, `rx`, 1 = Placebo, 2 = DPCA). A data set comprising age, biological data and varied clinical signs including the level of bilirubin serum (`bilirub`) are available in the `pbc.txt` file [VIT 05]. The patient's status is stored in the variable `status` ($0 = $ alive, $1 = $ deceased) and the follow-up duration (`years`) represents the elapsed time in years since the date of the diagnosis.

– How many deceased individuals can be identified? What proportion of these deaths can be found in the active arm?

– display the distribution of the follow-up durations of 312 patients, by distinctively bringing forward the deceased individuals. Calculate the median follow-up time (in years) for each of the two treatment groups. How many positive events are there beyond 10.5 years and what is the gender of these patients?

– the 19 patients, whose number (`number`) appears among the following list have undergone a transplant during the follow-up period:

```
5 105 111 120 125 158 183 241 246 247 254 263 264 265 274 288 291
295 297
```

Indicate their average age, the distribution according to sex and the median duration of the follow-up in days until transplant.

– display a table summarizing the distribution of individuals at risk according to time, with the associated survival value;

– display the Kaplan–Meier curve with a 95% confidence interval, without considering the treatment type;

– calculate the survival median and its 95% confidence interval for each group of subjects and display the corresponding survival curves;

– perform a log-rank test considering as predictor the factor `rx`. Compare with a Wilcoxon test;

– carry out a log-rank test on the factor of interest (rx) by stratifying the age. Three age groups will be considered: 40 years old or less, between 40 and 55 years of age inclusive, more than 55 years old;

– try to find the results of exercise 1(g) with a Cox regression.

The data file is a text file with tabs as field separator. It can be imported in Stata utilizing the insheet command. To display the name of the variables after importing it, it suffices to enter describe with the option simple:

```
. insheet using "pbc.txt", tab
. describe, simple
```

After recoding the labels of the rx and sex variables:

```
. label define trt 1 "Placebo" 2 "DPCA"
. label define sexe 0 "M" 1 "F"
. label values rx trt
. label values sex sexe
```

the proportion of patients who died can be verified (status, $0 =$ alive and $1 =$ deceased) and their distribution according to the treatment group based on simple and crossed tabulation. Regarding the cross-tabulation, the option row will be added to obtain the relative frequencies per status:

```
. tabulate status
. tabulate status rx, row
```

To display the distribution of the follow-up times, we will make use of a simple scatterplot. To distinctly present the observations according to the status (0 or 1), we could very well superimpose two sets of points over the same graph. Here is another way of proceeding:

```
. separate number, by(status)
. twoway scatter number0 number1 years, msymbol(S O)
```

The first command actually allows the numbers of patients to be separated according to the status function in order to display both sets of observation with respect to the follow-up durations in years.

The median of the follow-up duration per treatment group can be obtained with the tabstat command operating on a group basis with the option by:

```
. tabstat years, by(rx) stats(median) nototal
```

The number of deaths beyond 10.5 years of follow-up is obtained with a simple tabulation inserting the tabulate command:

```
. tabulate status if years > 10.49
```

as well as the gender of patients who died after this period:

```
. tabulate sex if years > 10.49 & status == 1
```

Regarding the analysis of transplant patients, it is possible to restrict the data table to these patients only. Since Stata allows working with only one data table at the time, it is however necessary to temporarily save the current data before creating a new table:

```
. preserve
. egen idx = anymatch(number), values(5 105 111 120 125 158 183 241 246 247 254
  263 264 265 274 288 291 295 297)
. keep if idx
. gen days = years*365
. tabstat age sex days, stats(mean median sum)
```

The first command makes it possible to build a list of individuals that we may want to employ to filter the original data table (based on subject IDs contained in the variable `number`). Next, the calculation of the descriptive statistics is carried out utilizing the command `tabstat`. Once the calculations have been completed, the original data can be restored in the following manner:

```
. restore
```

Stata makes use of its own conventions for encoding survival data. The essential commands are thus: `stset` to define the way in which events are logged and the observation time, `sts` to calculate a survival table based on the Kaplan–Meier estimator. Here is how to apply these commands to build the table and the survival curve, regardless of the treatment factor:

```
. stset years, failure(status)
. sts list
```

The second command displays the requested table. For the survival curve, we will use:

```
. sts graph, ci censored(single)
```

The option `ci` enables displaying the 95% confidence interval for the Kaplan–Meier estimator.

The survival median is obtained by employing the command `stci`. Without further information, Stata calculates the median survival and its 95% confidence interval for all of the observations. The `by()` option will be added to obtain the median survival per treatment group:

```
. stci, by(rx)
```

Similarly, it is possible to print the two corresponding survival curves with `sts graph` retaining the `by()` option:

```
. sts graph, by(rx) cen(single)
```

With regard to the log-rank test, the command `sts test` has to be entered, the `logrank` option being the default option. We simply need to indicate the classification variable to test the equality of the survival functions, that is:

```
. sts test rx
```

The `wilcoxon` option will be added to obtain the Wilcoxon test. The `noshow` option allows simplifying the text output by removing the information relative to the variables `sts`:

```
. sts test rx, wilcoxon noshow
```

The same command can be used when we want to include a stratification factor, with the `strata()` option. As a first step, the variable `age` is recoded into a four-class qualitative variable using `egen`:

```
. egen agec = cut(age), at(26,40,55,79)
. sts test rx, strata(agec) noshow
```

Finally, to achieve a Cox regression model, with stratification on the `agec` factor, we employ the command:

```
. stcox rx, strata(agec)
```

By default, the results returned by Stata are expressed in terms of risk (`Haz. Ratio`). In order to obtain the regression coefficients, it is necessary to add the `nohr` option:

```
. stcox rx, strata(agec) nohr
```

2) In a randomized clinical trial, the aim was to compare two treatments for prostate cancer. Patients took orally each day either 1 mg of diethylstilbestrol (DES, active arm) or a placebo, and the survival time is measured in months [COL 94]. The question of interest is knowing whether the survival is different between the two groups of patients, and the other variables present in the `prostate.dat` data file will be ignored.

– Calculate the survival median for the totality of the patients and per treatment group.

– what is the difference between the survival proportions in both groups at 50 months?

– display the survival curves for the two groups of patients;

– perform a log-rank test to verify the hypothesis according to which the DES treatment has a positive effect on the survival of patients.

The text format of the data file `prostate.dat` is identical to that of the file `pbc.txt` of the previous exercise, except that the fields are separated by a single space. The command `insheet` will thus be used:

```
. insheet using "prostate.dat", delimiter(" ")
. list in 1/5
```

The number of living patients at the point time is obtained from `tabulate`:

```
. tabulate status
```

that is six persons still alive at the end of follow-up time.

To tell Stata which variables are utilized to identify the events (`status`) and the time (`time`), we use the command `stset` in the following manner:

```
. stset time, failure(status)
```

The median survival according to the treatment is available with the command `stci` specifying the percentile of interest (here, `p(50)`):

```
. stci, by(treatment) p(50)
```

To plot the survival curves for each treatment arm, we will make use of the command `sts graph` specifying the classification factor by means of the option `by`. The option `censored` displays the censored data:

```
. sts graph, by(treatment) censored(s)
```

The log-rank test is achieved with the command `sts test`. Note that only the treatment factor has to be specified, the notion of response variable being managed from the beginning by `stset`:

```
. sts test treatment, noshow
```

Regarding Cox's model, the `stcox` command will be employed as in the previous exercise, bearing in mind that by default Stata provides the estimated value for the risk, and not the regression coefficient (use `nohr`):

```
. stcox treatment, noshow
```