Chapter 10

Survival Analysis

Aim: Upon completing the chapter, the learner should be able to use the Cox proportional hazards model to estimate the magnitude of the association between the risk of the occurrence of a given clinical event (e.g., disease, death, remission) after a certain period of time and a factor of exposure, controlling for potential confounders.

10.1 Introduction

In this chapter we present the use of a regression model to analyze the occurrence of an event after a certain time. This analysis is regularly identified as a *survival analysis* or a *time to event analysis* (Kleinbaum and Klein, 2005). The objective in survival analysis is to assess the time it takes for an event of interest to occur when there is the possibility that this event will not occur in all subjects under study. Take the following examples:

Event of Interest	*Time of Study*	*At the End of the Study*
Death	Time that a person remains alive after heart surgery until the event of interest occurs.	There are persons who remain alive after heart surgery.
Development of a respiratory disease	Time without developing a respiratory disease after exposure to an environmental contaminant until the event of interest occurs.	There are persons who do not develop a respiratory disease after being exposed to an environmental contaminant.
Hospital discharge after 24 h	Length of hospital stay of a patient who arrives at an emergency room until the event of interest occurs.	There are people who stay in the hospital for longer than 24 h.

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

For the analysis of survival times, it is necessary to identify a start date for participation in the study. Some possible start dates are date of birth, date of diagnosis, therapy start date, and date or time of an exposure to a toxin. In addition, it is also necessary to identify the date on which the event of interest occurs or the date of study completion.

The study time in survival analysis is determined by the difference between the date of the occurrence of the event of interest and the start date of the study:

$$T = \text{date of occurrence of event under study} - \text{start date}$$

where T can be measured in days, months, years, or some other time unit.

The use of survival analysis is justified when there is a possibility that an event of interest will not occur in a high number of subjects during a given study period, meaning that there will be a high number of individuals with incomplete information. The time of occurrence of the event of interest (T) cannot be exactly determined when the event does not happen; only the minimum survival time (t) (in which the event of interest does not occur in the individual) can be determined. Therefore, the formulation of a study problem in survival analysis with the event's date or time of occurrence being unknown is given by the following expression:

$$T \geq t$$

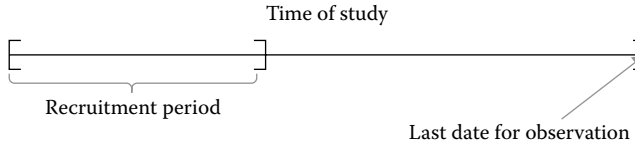
Censoring information may arise in the following situations:

- Termination of the research study
- Loss of follow-up due to the voluntary withdrawal from the study (reasons unrelated to the study)

- Death, assuming that this is not the event of interest
- Development of a disease or health condition that is not associated with the event of interest

When the study time of a subject has not been determined, we use the term censored. If the date of the event's occurrence is unknown, the incomplete data in the survival analysis are called right censoring. The existence of censored observations can be attributed to a selection bias, unless it can be assured that censored individuals are representative of the study population. Therefore, censoring has to be independent of *t*.

A survival analysis involves a longitudinal design in which there is a recruitment period and a maximum date of observation, as illustrated in the following:



Recruitment time comprises a fixed period of time during which the initial measurement of the study subjects for survival analysis is performed. The maximum date of observation indicates the last day or the specific time to observe the occurrence of an event. The possible situations that may occur while observing the event of interest are illustrated in Figure 10.1.

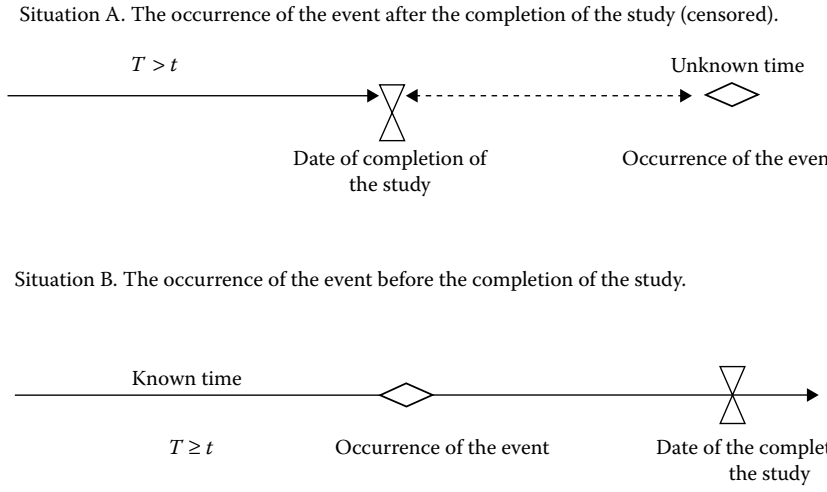


Figure 10.1 Occurrence of the Event in Survival Analysis.

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

10.2 Probability of Survival

In a survival study, the evaluation $T = t$ cannot be established for censored observations. If there are no censored observations, then the variable T can be analyzed using the standard methods for evaluating a continuous variable (e.g., linear regression, ANOVA). In the case of censored observations, we can ensure only that the event occurred after a time greater than t ($T > t$). For this reason, one of the main objectives in a survival analysis is to determine the probability of $T > t$. This probability is expressed as follows:

$$S(t) = \Pr[T > t]$$

$S(t)$ is defined as the *survival function* and indicates the probability of being free of the event of interest at least at t; that is, the probability of the event occurring after t.

10.3 Components of the Study Design

To perform a survival analysis, the minimum information that must be provided by the study design is: (1) recruitment date, (2) date of the occurrence of the study event, and (3) last date of observation (which could be the date of study termination or the occurrence of the study event). For example, let us assume that the time distribution of seven subjects with cancer in a survival analysis design is as illustrated in Figure 10.2.

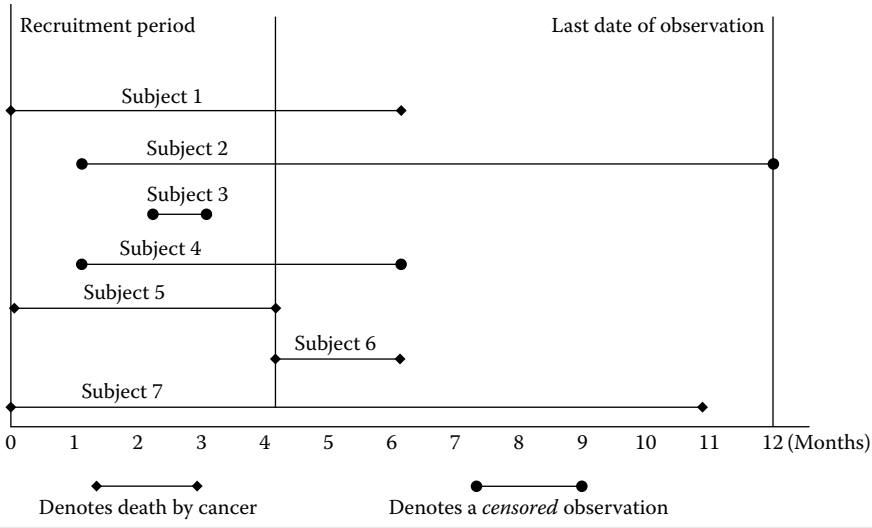


Figure 10.2 Longitudinal design in survival analysis.

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

In the above illustration, the following patterns occur: (1) the recruitment date is different for every subject; (2) the date of the occurrence of the study event is also different in each subject; (3) the last date of observation is the same for all subjects, 12 months; however, not all the subjects get free of the event at this point. Usually, this type of information can be summarized in the following table:

Subject	*Entrance Month*	*Month of Death or Censored*	*Death (D) Censored (C)*	*Survival Time*
1	0	6	D	6
2	1	12	C	11
3	2	3	C	1
4	1	6	C	5
5	0	4	D	4
6	4	6	D	2
7	0	11	D	11

When there are censored observations, there are several nonparametric methods for estimating $S(t)$. The life-table estimate of the survivor function, also known as the *actuarial estimate* of survivor function, assumes that the censoring process is such that the censored survival times occur uniformly within different series of time intervals. Another method for estimating the survival function, $S(t)$, is through the Kaplan–Meier (KM) method. This method determines the probability of surviving at least to time $t_{(j)}$. Time $t_{(j)}$ indicates the times at which one or more events have occurred and is sorted in ascending order:

$$t_{(1)} \leq t_{(2)} \leq t_{(3)} \leq \cdots \leq t_{(j-1)} \leq t_{(j)} \leq \cdots$$

where $t_{(1)}$ is the time at which the event of least time occurred.

10.4 Kaplan–Meier Method

To describe the procedure to estimate the survival function of a person, $S(t_{(j)})$, we use the KM method. This method is based on the product of survival function at time $t_{(j-1)}$ and the conditional probability that the event of interest will occur after $t_{(j)}$, given that the person is alive at least until $t_{(j)}$. This product is expressed as follows:

$$S\left(t_{(j)}\right) = S\left(t_{(j-1)}\right) * \Pr\left(T > t_{(j)} \middle| T \geq t_{(j)}\right)$$

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

where $\Pr[T > t_{(j)} | T \geq t_{(j)}]$ indicates the probability, among those persons who reached $t_{(j)}$ alive, of remaining alive after that specific time,

- $t_{(j)}$ indicates the time in j order in which at least one event occurs after the data are ordered from least to greatest
- $S(t_{(j-1)})$ is the function of survival until time $t_{(j-1)}$

The development of the previous expression with the data from the previous example is presented in the following table:

$T(j)$	r_j[a]	f_j[b]	$Pr[T > t_{(j)} \mid T \geq t_{(j)}]$[c]	$S(t_j)$
0	7	0	1 – (0/7) = 1	1.0
1[d]	–	–	–	–
2	6	1	1 – (1/6) = 0.833	0.833
4	5	1	1 – (1/5) = 0.8	0.666
5[d]	–	–	–	–
6	3	1	1 – (1/3) = 0.667	0.44
11[d]	–	–	–	–
11	1	1	1 – (1/1) = 0	0

[a] r_j indicates the subjects that are at risk an instant before $t_{(j)}$.
[b] f_j indicates the number of deaths in j time.
[c] $\Pr[T > t_{(j)} | T \geq t_{(j)}] = 1 - (f_j / r_j)$.
[d] censored cases.

The $S(t_{(j)})$ usually is graphically represented as a step function; it means that $S(t_{(j)})$ probability remains constant until the time when the next event of interest occurs.

10.5 Programming of *S*(*t*)

Let us assume a study aimed at assessing oropharyngeal cancer mortality by sex (1 = Males, 2 = Females), adjusting for tumor stage (1, 2, and 3). Further, let us suppose that we have the observation time (months) after the diagnosis in 87 subjects, as follows:

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

id	*death*	*sex*	*stage*	*time*	*id*	*death*	*sex*	*stage*	*time*
1	1	2	2	34	46	0	2	1	28
2	1	2	2	61	47	1	1	2	39
3	0	2	1	78	48	1	1	2	8
4	0	1	2	95	49	1	2	1	35
5	1	2	2	49	50	0	1	1	5
6	1	2	2	59	51	0	1	1	45
7	0	2	2	2	52	1	2	2	0
8	0	2	2	1	53	0	2	2	17
9	1	1	3	6	54	0	2	1	0
10	0	1	1	53	55	0	1	2	1
11	1	2	2	8	56	0	2	1	39
12	0	1	2	21	57	1	2	1	9
13	1	2	3	71	58	1	1	2	5
14	1	1	3	47	59	1	2	2	2
15	0	1	2	35	60	0	1	2	38
16	0	1	1	1	61	1	2	1	41
17	1	2	3	10	62	0	2	2	6
18	0	1	2	7	63	1	2	1	28
19	0	2	2	1	64	0	2	1	60
20	1	1	2	27	65	0	2	2	1
21	0	1	2	34	66	1	2	1	81
22	1	2	3	10	67	0	2	1	2
23	1	2	2	43	68	0	1	1	5
24	0	1	2	84	69	0	1	2	2
25	1	2	2	89	70	0	2	2	44
26	1	2	2	6	71	1	2	3	8

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

id	*death*	*sex*	*stage*	*time*	*id*	*death*	*sex*	*stage*	*time*
27	0	2	2	4	72	1	1	3	7
28	0	2	2	0	73	0	1	2	5
29	0	2	1	22	74	0	2	2	3
30	1	2	3	1	75	1	2	1	3
31	1	2	1	23	76	1	2	2	2
32	1	2	3	37	77	1	1	2	55
33	1	2	1	25	78	1	2	1	46
34	0	2	1	0	79	0	2	1	70
35	0	2	3	1	80	1	2	1	39
36	1	2	2	39	81	1	1	1	99
37	1	2	1	20	82	0	1	2	5
38	0	1	2	48	83	1	1	2	52
39	1	1	2	20	84	1	2	3	12
40	1	2	2	6	85	0	2	1	39
41	0	2	1	44	86	1	1	1	40
42	1	2	3	13	87	0	1	2	73
43	0	1	3	1					
44	0	2	2	50					
45	0	2	1	0					

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

To run a survival analysis in Stata, we have to specify the name of the variable that defines the time and the variable that defines the event with the code to be used for the occurrence of the event, as follows:

```
stset time,fa(death=1)
```

Output

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

```
------------------------------------------------------------------------------
      87  total observations
       5  observations end on or before enter()
------------------------------------------------------------------------------
      82  observations remaining, representing
      43  failures in single-record/single-failure data
    2385  total analysis time at risk and under observation
                                     at risk from t =         0
                          earliest observed entry t =         0
                               last observed exit t =        99
```

We defined the time-of-survival variable at the beginning; in this case it was defined with the name *time.* After the comma, the occurrence of the event of interest is indicated with the command **fa** followed by a parenthesis to indicate the variable for the event of interest and the code that indicates when the event occurs.

An estimation of survival probability is obtained in Stata with the *ltable* command, as is demonstrated in the following:

```
ltable time
```

Output

Interval		Beg. Total	Deaths	Lost	Survival	Std. Error	[95% Conf.	Int.]
0	1	87	5	0	0.9425	0.0250	0.8674	0.9757
1	2	82	8	0	0.8506	0.0382	0.7566	0.9104
2	3	74	5	0	0.7931	0.0434	0.6919	0.8642
3	4	69	2	0	0.7701	0.0451	0.6667	0.8451
4	5	67	1	0	0.7586	0.0459	0.6542	0.8354
5	6	66	5	0	0.7011	0.0491	0.5930	0.7856
6	7	61	4	0	0.6552	0.0510	0.5453	0.7446
7	8	57	2	0	0.6322	0.0517	0.5218	0.7238
8	9	55	3	0	0.5977	0.0526	0.4870	0.6920
9	10	52	1	0	0.5862	0.0528	0.4755	0.6813
10	11	51	2	0	0.5632	0.0532	0.4527	0.6597
12	13	49	1	0	0.5517	0.0533	0.4414	0.6489
13	14	48	1	0	0.5402	0.0534	0.4302	0.6380
17	18	47	1	0	0.5287	0.0535	0.4190	0.6270
20	21	46	2	0	0.5057	0.0536	0.3967	0.6049
21	22	44	1	0	0.4943	0.0536	0.3857	0.5938
22	23	43	1	0	0.4828	0.0536	0.3747	0.5826
23	24	42	1	0	0.4713	0.0535	0.3637	0.5714
25	26	41	1	0	0.4598	0.0534	0.3529	0.5601
27	28	40	1	0	0.4483	0.0533	0.3420	0.5488
28	29	39	2	0	0.4253	0.0530	0.3205	0.5260
34	35	37	2	0	0.4023	0.0526	0.2993	0.5029
35	36	35	2	0	0.3793	0.0520	0.2783	0.4797
37	38	33	1	0	0.3678	0.0517	0.2678	0.4680

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

38	39	32	1	0	0.3563	0.0513	0.2575	0.4562
39	40	31	5	0	0.2989	0.0491	0.2067	0.3964
40	41	26	1	0	0.2874	0.0485	0.1967	0.3843
41	42	25	1	0	0.2759	0.0479	0.1868	0.3721
43	44	24	1	0	0.2644	0.0473	0.1770	0.3598
44	45	23	2	0	0.2414	0.0459	0.1577	0.3350
45	46	21	1	0	0.2299	0.0451	0.1481	0.3225
46	47	20	1	0	0.2184	0.0443	0.1387	0.3099
47	48	19	1	0	0.2069	0.0434	0.1293	0.2972
48	49	18	1	0	0.1954	0.0425	0.1200	0.2844
49	50	17	1	0	0.1839	0.0415	0.1109	0.2715
50	51	16	1	0	0.1724	0.0405	0.1019	0.2585
52	53	15	1	0	0.1609	0.0394	0.0930	0.2454
53	54	14	1	0	0.1494	0.0382	0.0842	0.2322
55	56	13	1	0	0.1379	0.0370	0.0756	0.2188
59	60	12	1	0	0.1264	0.0356	0.0671	0.2053
60	61	11	1	0	0.1149	0.0342	0.0589	0.1916
61	62	10	1	0	0.1034	0.0327	0.0508	0.1778
70	71	9	1	0	0.0920	0.0310	0.0430	0.1637
71	72	8	1	0	0.0805	0.0292	0.0354	0.1494
73	74	7	1	0	0.0690	0.0272	0.0282	0.1349
78	79	6	1	0	0.0575	0.0250	0.0213	0.1200
81	82	5	1	0	0.0460	0.0225	0.0150	0.1047
84	85	4	1	0	0.0345	0.0196	0.0092	0.0889
89	90	3	1	0	0.0230	0.0161	0.0044	0.0725
95	96	2	1	0	0.0115	0.0114	0.0010	0.0558
99	100	1	1	0	0.0000	.	.	.

Note: The command *ltable* is followed by the variable that indicates the observation time.

After running the *stset* command, the *sts graph* command can be used in Stata to construct the $S(t)$ graph using the KM method, as illustrated in Figure 10.3.

10.6 Hazard Function

Another way to assess the time of occurrence of an event is by using the *hazard function,* which is denoted by $h(t)$. This function is defined as the instantaneous risk of occurrence of the event of interest after a time, t, given that this event did not occur during at least a single time t. To estimate the hazard function using the KM method, in a given time interval $(t_{(j)}, t_{(j+1)})$, a similar process is used to compute the incidence density:

$$\hat{h}(t) = \frac{f_j}{r_j * \tau_j} = \frac{\text{cases}}{\text{person-time}}$$

where τ_j indicates the size of the interval $(t_{(j)}, t_{(j+1)})$, that is, $\tau_j = t_{(j+1)} - t_{(j)}$. According to the time unit that is used, the product $r_j{}^*\tau_j$ indicates person-time (i.e., person-years,

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

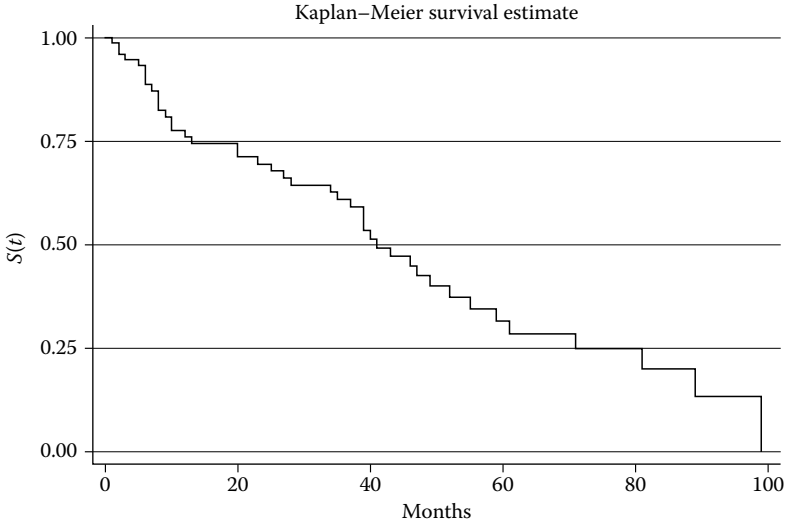


Figure 10.3 Survival function using the Kaplan–Meier method.

person-weeks, person-days, etc.). The estimation of the hazard probabilities can also be obtained with the *ltable* command, but we add the *hazard* option. For example, using the previous data, the command is as follows:

```
ltable time, hazard
```

Output

Interval		Beg. Total	Cum. Failure	Std. Error	Hazard	Std. Error	[95% Conf.	Int.]
0	1	87	0.0575	0.0250	0.0592	0.0265	0.0073	0.1110
1	2	82	0.1494	0.0382	0.1026	0.0362	0.0316	0.1735
2	3	74	0.2069	0.0434	0.0699	0.0313	0.0087	0.1312
3	4	69	0.2299	0.0451	0.0294	0.0208	0.0000	0.0702
4	5	67	0.2414	0.0459	0.0150	0.0150	0.0000	0.0445
5	6	66	0.2989	0.0491	0.0787	0.0352	0.0098	0.1477
6	7	61	0.3448	0.0510	0.0678	0.0339	0.0014	0.1342
7	8	57	0.3678	0.0517	0.0357	0.0252	0.0000	0.0852
8	9	55	0.4023	0.0526	0.0561	0.0324	0.0000	0.1195
9	10	52	0.4138	0.0528	0.0194	0.0194	0.0000	0.0575
10	11	51	0.4368	0.0532	0.0400	0.0283	0.0000	0.0954
12	13	49	0.4483	0.0533	0.0206	0.0206	0.0000	0.0610
13	14	48	0.4598	0.0534	0.0211	0.0211	0.0000	0.0623
17	18	47	0.4713	0.0535	0.0215	0.0215	0.0000	0.0637
20	21	46	0.4943	0.0536	0.0444	0.0314	0.0000	0.1060
21	22	44	0.5057	0.0536	0.0230	0.0230	0.0000	0.0680
22	23	43	0.5172	0.0536	0.0235	0.0235	0.0000	0.0696

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

```
23   24    42   0.5287   0.0535   0.0241   0.0241   0.0000   0.0713
25   26    41   0.5402   0.0534   0.0247   0.0247   0.0000   0.0731
27   28    40   0.5517   0.0533   0.0253   0.0253   0.0000   0.0749
28   29    39   0.5747   0.0530   0.0526   0.0372   0.0000   0.1255
34   35    37   0.5977   0.0526   0.0556   0.0393   0.0000   0.1325
35   36    35   0.6207   0.0520   0.0588   0.0416   0.0000   0.1403
37   38    33   0.6322   0.0517   0.0308   0.0308   0.0000   0.0911
38   39    32   0.6437   0.0513   0.0317   0.0317   0.0000   0.0940
39   40    31   0.7011   0.0491   0.1754   0.0782   0.0223   0.3286
40   41    26   0.7126   0.0485   0.0392   0.0392   0.0000   0.1161
41   42    25   0.7241   0.0479   0.0408   0.0408   0.0000   0.1208
43   44    24   0.7356   0.0473   0.0426   0.0425   0.0000   0.1259
44   45    23   0.7586   0.0459   0.0909   0.0642   0.0000   0.2168
45   46    21   0.7701   0.0451   0.0488   0.0488   0.0000   0.1444
46   47    20   0.7816   0.0443   0.0513   0.0513   0.0000   0.1518
47   48    19   0.7931   0.0434   0.0541   0.0540   0.0000   0.1600
48   49    18   0.8046   0.0425   0.0571   0.0571   0.0000   0.1691
49   50    17   0.8161   0.0415   0.0606   0.0606   0.0000   0.1793
50   51    16   0.8276   0.0405   0.0645   0.0645   0.0000   0.1909
52   53    15   0.8391   0.0394   0.0690   0.0689   0.0000   0.2041
53   54    14   0.8506   0.0382   0.0741   0.0740   0.0000   0.2192
55   56    13   0.8621   0.0370   0.0800   0.0799   0.0000   0.2367
59   60    12   0.8736   0.0356   0.0870   0.0869   0.0000   0.2572
60   61    11   0.8851   0.0342   0.0952   0.0951   0.0000   0.2817
61   62    10   0.8966   0.0327   0.1053   0.1051   0.0000   0.3113
70   71     9   0.9080   0.0310   0.1176   0.1174   0.0000   0.3478
71   72     8   0.9195   0.0292   0.1333   0.1330   0.0000   0.3941
73   74     7   0.9310   0.0272   0.1538   0.1534   0.0000   0.4545
78   79     6   0.9425   0.0250   0.1818   0.1811   0.0000   0.5367
81   82     5   0.9540   0.0225   0.2222   0.2208   0.0000   0.6551
84   85     4   0.9655   0.0196   0.2857   0.2828   0.0000   0.8400
89   90     3   0.9770   0.0161   0.4000   0.3919   0.0000   1.1681
95   96     2   0.9885   0.0114   0.6667   0.6285   0.0000   1.8986
99  100     1   1.0000        .   2.0000   0.0000   2.0000   2.0000
-------------------------------------------------------------------------------
```

To graphically represent $h(t)$ through the KM method, the same command for $S(t)$ is used, but we add the hazard option (***sts graph, hazard***). The output of this command is illustrated in Figure 10.4.

10.7 Relationship between $S(t)$ and $h(t)$

Assuming that the time of observation, T, is a random variable in the survival analysis setting, then, as a consequence, the following functions exist:

$f(t)$: the probability density function of T
$F(t)$: the probability cumulative function of $f(t)$

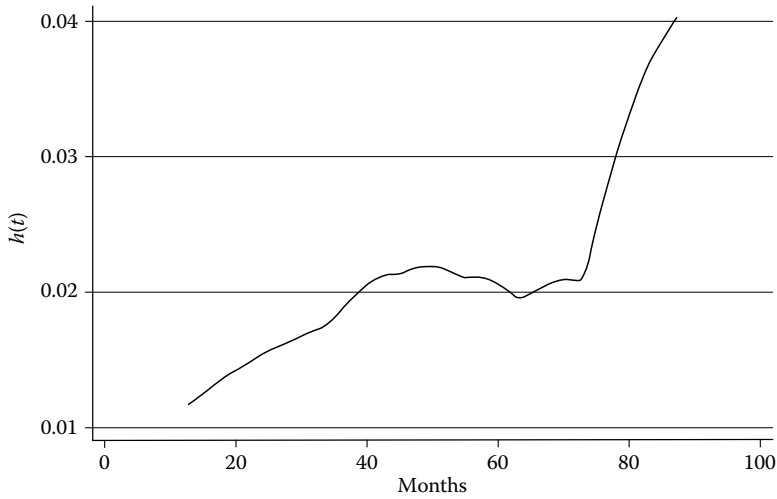


Figure 10.4 Hazard function.

It has been shown that there is a mathematical relationship between these survival and hazard functions, the specifics of which are as follows (Collett, 2003):

a. $$f(t) = \frac{\partial F(t)}{\partial t} = \frac{\partial\left(1 - S(t)\right)}{\partial t} = -S'(t)$$

b. $$h(t) = \lim_{\Delta t \to 0} \frac{\Pr\left(t + \Delta t > T > t \mid T > t\right)}{\Delta t} = \frac{f(t)}{S(t)}$$

c. $$H(t) = \int_{o}^{t} h(u)\partial u = \int_{0}^{t} \frac{f(u)}{S(u)}\partial u = -\int_{o}^{t} \frac{1}{S(u)}\left(\frac{\partial S(u)}{\partial u}\right)\partial u = -\ln\left(S(t)\right)$$

where $S'(t)$ indicates the derivative of $S(t)$ and $H(t)$, the cumulative hazard function.

10.8 Cumulative Hazard Function

The following methods are available for determining the cumulative hazard function:

1. The KM method:

$$\widehat{H}(t) = -\mathrm{Ln}\left(\hat{S}(t)\right) = -\mathrm{Ln}\left(\prod_{t(i)\le t}\left(\frac{r_i - f_i}{r_i}\right)\right) = \sum_{t(i)\le t} -\mathrm{Ln}\left(1 - \frac{f_i}{r_i}\right)$$

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

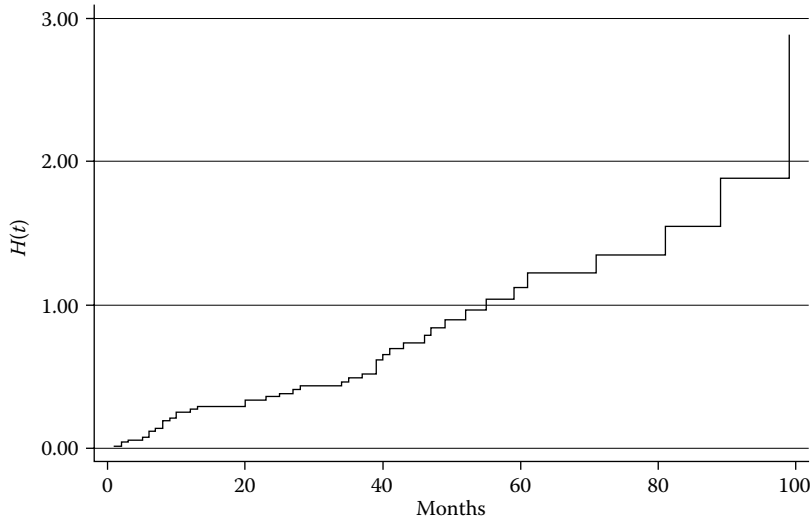


Figure 10.5 Nelson–Aelen cumulative hazard estimate.

2. The Nelson−Aalen method:

$$\widetilde{H}(t) = \sum_{t(i)\le t} \frac{f_i}{r_i}$$

The estimate of $H(t)$ using the Nelson–Aelen method will always be greater than or equal to that which is generated using the KM method. When the number of subjects at risk at any given time is large, the two estimates are basically equal. Based on the Nelson–Aalen method, we can obtain the survival function with the following expression:

$$\widetilde{S(t)} = e^{(-H(t))}$$

The graphic representation of the cumulative hazard $[H(t)]$ using the Nelson–Aalen method is programmed in Stata using the following command to create Figure 10.5:

```
sts graph, cumhaz
```

10.9 Median Survival Time and Percentiles

The minimum survival time in which the probabilities are less than 50% is identified by the median survival time $\{t(50) = \min(\text{time}|S(t) < 50\%)\}$. In Stata this time is obtained as follows:

```
stci, median
```

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

Output

```
      failure _d:  death == 1
  analysis time _t:  time

          |   no. of
          | subjects   50%   Std. Err.      [95% Conf. Interval]
----------+--------------------------------------------------------------
   total  |   82       41     3.705979            34          55
```

Note: The **median** option is added.

If an estimate of other times, based on percentiles [(1–$S(t)$], is required, we use the option **p** with the integer that indicates the percentile. For example, for the percentiles 25 and 75, the Stata commands are as follows:

```
stci, p(25)
```

Output

```
          |    no. of
          |  subjects    25%   Std. Err.  [95% Conf. Interval]
----------+--------------------------------------------------------------

   total  |    82        13    6.687969            8          28
```

```
stci, p(75)
```

Output

```
          |   no. of
          | subjects    75%    Std. Err.  [95% Conf.  Interval]
----------+--------------------------------------------------------------
  total   |       82     71    12.37697            52          .
```

Note: In the previous case, the 25th percentile indicates the minimum time for which the survival probabilities are less than 75%. The 75th percentile indicates the minimum time for which the survival probabilities are less than 25%.

10.10 Comparison of Survival Curves

To compare the overall experience of two or more survival curves, you can use different tests of significance. Initially, it is recommended that a graph of $S_i(t)$ by different subgroups of the stratum be constructed. The Stata programming for visualizing different survival curves in the same plot, using the KM method, uses the option *by* after the *sts graph* command. For example, using the previous database, the survival curves by sex are graphically obtained with the following command to create Figure 10.6:

```
sts graph, by(sex)
```

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

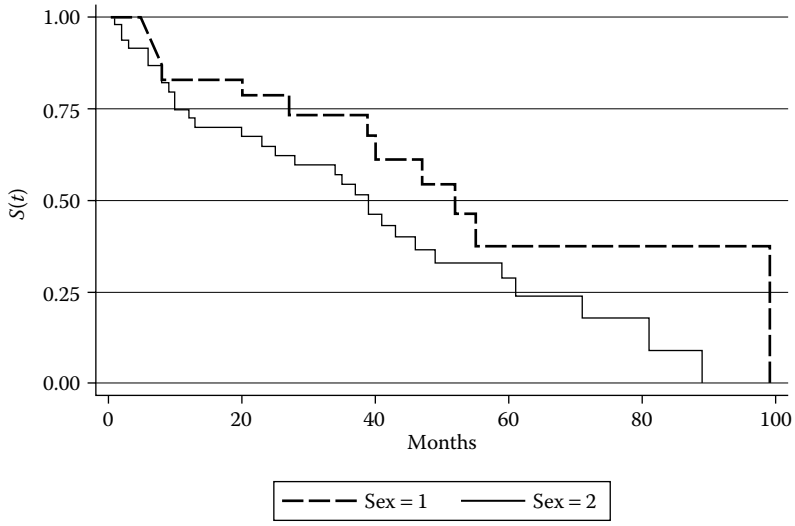


Figure 10.6 Survival function by sex.

To obtain the median survival time in different subgroups, the option *by* can also be used after the *stci* command. For example, the command line for finding the median time by sex would be

```
stci, median by(sex)
```

Output

```
         |    no. of
sex      |  subjects     50%     Std. Err.    [95% Conf. Interval]
---------+--------------------------------------------------------
       1 |        31      52      5.187957             27          .
       2 |        51      39      3.999146             23         49
---------+--------------------------------------------------------
   total |        82      41      3.705979             34         55
```

10.11 Proportional Hazards Assumption

The application of certain significance tests to evaluate hazard functions depends on the behavior of these functions among the study groups over time. If the ratio of hazard functions is constant over time $h_1(t)/h_2(t) =$ constant, it is said that the hazards are proportional (proportional hazards). For example, using the previous database, the hazard function curves by sex are graphically obtained with the following command to create Figure 10.7:

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

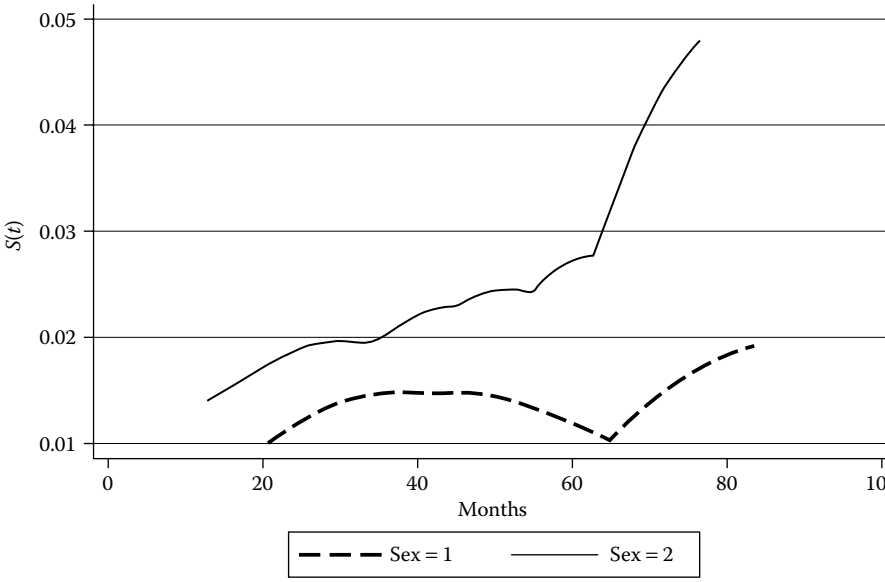


Figure 10.7 Hazard function by sex.

```
sts graph, by(sex) hazard
```

A *proportional hazards (PH)* assessment can be performed through the pattern observed in the survival function curves. It is a necessary but not sufficient condition that the curves not cross over time, suggesting that the PH assumption is met. An alternate way to evaluate this assumption graphically is to determine whether the function ln{–ln[$S(t)$]} is kept parallel through time. For example, for programming this function we use the following command to create Figure 10.8:

```
stphplot, by(sex)
```

10.12 Significance Assessment

Hypothesis testing can be performed to compare the survival curves of different groups. There are several methods that can be used to evaluate these curves, which include the log-rank test, the Wilcoxon–Gehan–Breslow (WGB) test, and the Tarone–Ware test. A log-rank test is recommended when the condition of PH is fulfilled, while a Wilcoxon test is recommended when the condition of PH is not fulfilled.

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

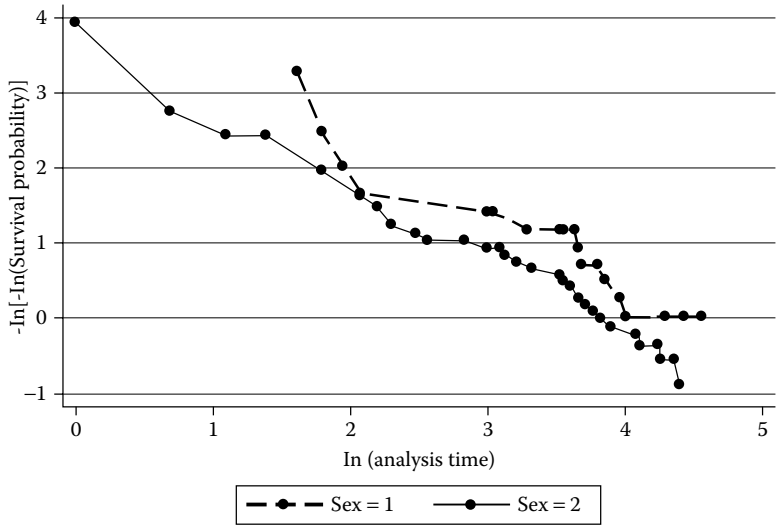


Figure 10.8 ln{–ln(*S*(*t*))} plot by sex.

10.12.1 Log-Rank Test

To compare the survival curves, the log-rank test assumes the following null hypothesis:

$$H_0 : S_1(t) = S_2(t)$$

To evaluate the survival curves with the log-rank test, the following contingency table is constructed at each time $t_{(j)}$:

Group	*Events*	*No Events*	*Number at Risk*
1	f_{1j}	r_{1j-f1j}	r_{1j}
2	f_{2j}	r_{2j-f2j}	r_{2j}
Total	f_j	r_{j-fj}	r_{1j}

Under the null hypothesis of no association between the type of group and the occurrence of the event, you can determine the expected events in each group and compare with the observed event with the following statistics:

$$\frac{U^2{}_L}{V_L} \sim \chi^2{}_{(1)}$$

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

where:

$U_L = \sum w_i(f_{1j} - e_{1j})$ defines the weighted difference (observed events minus expected events $E(f_{1j}) = e_{1j,}$ under H_0) with equal weights ($w_i = 1$) over time

$V_L = \sum v_{ij}$ determines the sum of the variances under the hypergeometric distribution $v_{ij} = \text{Var}(f_{1j}) = r_{1j}{}^* r_{\text{II}j}{}^* f_j{}^* (r_j - f_j)/r^2{}_j{}^* (r_j - 1)$

The Stata command for a log-rank test is:

```
sts test sex, logrank
```

Output

```
Log-rank test for equality of survivor functions
      |   Events          Events
sex   |  observed        expected
------+-------------------------
1     |        12           17.84
2     |        31           25.16
------+-------------------------
Total |        43           43.00
                           chi2(1) =       3.48
                           Pr>chi2 =     0.0620
```

According to the log-rank test performed, there is evidence in favor of H_0: $S_1(t) = S_2(t)$ (P-value = .062). However, some users might even consider this statistical evidence as marginally significant ($0.05 \leq P\text{-value} < .1$).

10.12.2 Wilcoxon–Gehan–Breslow Test

The Wilcoxon–Gehan–Breslow (WGB) test is also based on the null hypothesis

$$H_0 : S_1(t) = S_2(t)$$

To perform this test, the following statistic is computed:

$$\frac{U^2_{\text{WGB}}}{V_{\text{WGB}}} \sim \chi^2_{(1)}$$

where:

$$U_{\text{WGB}} = \sum w_j * \left(f_{1j} - \text{e}_{1j}\right); \quad w_j = r_j$$

$$V_{\text{WGW}} = \sum w_j^2 * v_{1j}$$

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

The difference between the U_W statistic and the statistics of the log-rank test is that $(f_{ij} - e_{ij})$ is being weighted by r_j. As time increases, the r_j decreases; therefore, individuals with very high values in the observation times will have less weight. The Stata command for this test is

```
sts test sex, wilcoxon
```

Output

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

       |   Events          Events         Sum of
sex    |  observed        expected         ranks
-------+-------------------------------------
1      |        12           17.84          -208
2      |        31           25.16           208
-------+-------------------------------------
Total  |        43           43.00             0
           chi2(1) =        2.23
           Pr>chi2 =      0.1350
```

According to the Wilcoxon test, there is evidence in favor of the statement H_0: $S_1(t) = S_2(t)$ (P-value > .1).

10.12.3 Tarone–Ware Test

The Tarone–Ware test is similar to the WGB test, but the weighting factor is r_j^2. The Stata command for this is as follows:

```
sts test sex, tware
```

Output

```
Tarone-Ware test for equality of survivor functions

       |   Events          Events         Sum of
sex    |  observed        expected         ranks
-------+-------------------------------------
1      |        12           17.84     -31.88649
2      |        31           25.16      31.88649
-------+-------------------------------------
Total  |        43           43.00             0
                  chi2(1) =       2.63
                  Pr>chi2 =     0.1049
```

According to the Tarone–Ware test, there is evidence in favor of the statement H_0: $S_1(t) = S_2(t)$ (P-value > .10).

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

10.13 Cox Proportional Hazards Model

To evaluate the effect of an exposure on the hazard function, controlling for the presence of potential confounding variables, we can use the *Cox proportional hazards model* (Royston and Lambert, 2011), as follows:

$$h(t, X) = h_0(t) * e^{\beta_E * E + \sum \beta_i X_i}$$

where:

E defines the exposure variable of interest

X_i defines the group of independent or predictor variables (it includes potential confounding variables and interaction terms)

β_i defines the group of coefficients of the predictor variables

$h_0(t)$ defines the immediate risk in time t. This function depends on the time and indicates the risk at initial conditions ($E = 0$, $X_i = 0$) or in average conditions when the predictor variables are centralized ($E = 0$, $X_i - \overline{X}$)

The predictor variables can be time dependent (e.g., age, blood pressure), but in this book, we are analyzing only those variables that are not time dependent. One of the most important uses of this model in epidemiologic studies is to estimate the hazard ratio (HR) adjusted for potential confounding variables. For example, assuming that the hazard between two persons having different exposure levels is obtained by the Cox model without interaction terms, the HR is estimated as follows:

First person ($E = 1$):

$$h_1(t; x) = h_0(t) * e^{(\beta_E + \beta_1 X_1 + \cdots + \beta_k X_k)}$$

Second person ($E = 0$):

$$h_2(t; x') = h_0(t) * e^{(\beta_1 X_1' + \ldots + \beta_\kappa X_\kappa')}$$

as a consequence,

$$\text{HR} = \frac{h_1(t; x)}{h_2(t; x')}$$

$$\text{HR} = e^{\left[\beta_E + \beta_1 * (X_1 - X_1') + \beta_2 * (X_2 - X_2') + \cdots + \beta_k (X_k - X_k')\right]}$$

If we assume that the difference between both persons is only the exposure, then:

$$(X_1 = X_1'),\ (X_2 = X_2'), \ldots, (X_k = X_k')$$

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

As a consequence, the adjusted HR is obtained with the following:

$$HR_{adjusted} = e^{\beta_E}$$

During the HR estimation, the $h_0(t)$ function is canceled. We assume that the HR remains constant over time; therefore, the HR varies only according to the value of the predictor variables. This process of obtaining the adjusted HR is similar to that used in evaluating the adjusted OR and RR of the logistic and Poisson regression models, respectively.

The Stata commands to evaluate the interaction terms in the Cox regression model is as follows:

```
. quietly xi: stcox i.sex*i.stage
. estimates store model1
. quietly xi: stcox i.sex i.stage
. lrtest model1 .
```

Output

```
Likelihood-ratio test                    LR chi2(2)  =      1.62
(Assumption: . nested in model1)         Prob > chi2 =    0.4440
```

The results show that there is no evidence of any significant interaction terms in the Cox model (P-value > .10). Now, to assess the effect of potential confounding variables, we need to compare the crude and adjusted HRs. To estimate the crude HR, *stcox* is used, as can be seen in the following:

```
stcox b1.sex
```

Output

```
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       2.sex |   1.906899   .6747282     1.82   0.068     .9531086    3.815161
------------------------------------------------------------------------------
```

Note: The use of *b*1 before the predictor is to indicate that the category with a code equal to 1 is the reference category.

The HR between sexes adjusted for stage is estimated with the following Stata command:

```
stcox b1.sex b1.stage
```

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

Output

```
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       2.sex |   1.894716   0.690035     1.75   0.079     0.9279952    3.868499
             |
       stage |
          2  |   1.472089   0.5575111    1.02   0.307     0.7007552    3.092446
          3  |   3.766573   1.596584     3.13   0.002     1.641108     8.644817
------------------------------------------------------------------------------
```

This example indicates that tumor stage is not a confounding variable in the association of *sex* and cancer mortality because the difference between the estimated HR_{Crude} (1.91) and the estimated $HR_{Adjusted}$ (1.89) is very small.

10.14 Assessment of the Proportional Hazards Assumption

If the condition of PH is fulfilled, there should not be an interaction between *time* and the exposure variable (E). To confirm this pattern, you can use the Cox model, as is demonstrated by the following:

$$h(t;x_1) = h_0(t) * e^{\beta_E * E + \beta_X * X}$$

where:

$$X = E * \ln(\text{time})$$

If the PH condition is met, then $\beta_E = 0$. Therefore, the expectation is that the interaction variable X would not be statistically significant to provide evidence that the PH assumption is met.

There are several methods for assessing the PH assumption, which are based on the quantities known as residuals. A description of these methods can be read in Collett (2003). Stata uses Schoenfeld residuals, or partial residuals, as its method for assessing PH. This method can be used with the option *phtest* after running *stcox*, as is demonstrated in the following:

```
quietly: stcox i.sex i.stage
estat phtest, detail
```

Output

```
Test of proportional-hazards assumption
```

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

```
Time:  Time
------------------------------------------------------------------
             |       rho       chi2       df         Prob>chi2
-------------+----------------------------------------------------
1b.sex       |         .          .        1                 .
2.sex        |   0.10298       0.46        1            0.4954
1b.stage     |         .          .        1                 .
2.stage      |  -0.03693       0.06        1            0.8142
3.stage      |  -0.09269       0.33        1            0.5671
-------------+----------------------------------------------------
global test  |                    0.80        3         0.8483
------------------------------------------------------------------
```

The data suggest that the condition of PH is fulfilled for both predictors simultaneously (*P*-value > .10).

10.15 Survival Function Estimation Using the Cox Proportional Hazards Model

Using the Cox model, we can obtain the survival function as follows:

$$S(t,E,x,\beta) = e^{-H(E,t,x,\beta)} = e^{-g(E,x,\beta)H_0(t)} = \left[e^{-H_0(t)}\right]^{g(E,x,\beta)} = \left[S_0(t)\right]^{g(x,\beta)}$$

where:

$$g(E,x,\beta) = e^{\beta_{E*E}+\sum\beta_i X_i}$$

$$S_0(t) = e^{-H_0(t)}$$

Therefore, for visualizing the survival curves by sex at stage 3 (stage = 3) after running the Cox model, the following sequences of commands is used to create Figure 10.9:

```
quietly xi: stcox b1.sex b1.stage
stcurve , survival at1(sex=1 stage=3) at2(sex=2 stage=3)
```

10.16 Stratified Cox Proportional Hazards Model

If the predictors do not satisfy the PH assumption, we recommend using the stratified Cox model, which is expressed as follows:

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

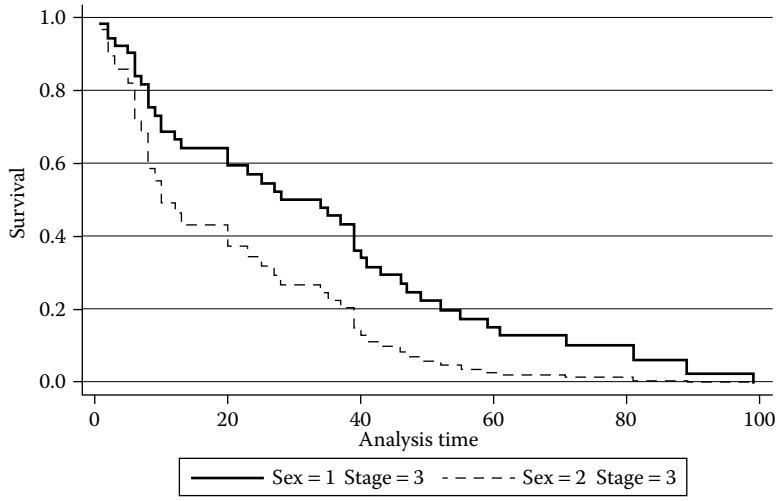


Figure 10.9 Survival curves by sex at stage 3.

$$h_g(t,X) = h_{0g}(t)\mathrm{e}^{\beta_E * E + \sum \beta_i * X_i}$$

where *g* indicates the stratum.

It is necessary to first identify all the predictors that do not satisfy the PH. With these predictor categories, strata are formed. For example, if the *stage* variable does not satisfy the PH assumption, the Cox model can be stratified by *stage* levels. In Stata the command line for the stratified Cox model using the previous data would have the variable *sex* as the sole predictor, as follows:

```
xi: stcox  b1.sex, strata(stage) nolog
```

Output

```
Stratified Cox regr. -- Breslow method for ties
No. of subjects =             82      Number of obs  =        82
No. of failures =             43
Time at risk    =           2385
                                      LR chi2(1)     =      2.84
Log likelihood  =   -98.460443        Prob > chi2    =    0.0919
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       2.sex |   1.818963   0.6677337     1.63   0.103     0.8858321    3.73505
------------------------------------------------------------------------------
                                                        Stratified by stage
```

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017

A slight variation is observed between the HR stratified by stage ($HR_{\text{stratified by stage}}$: 1.82, 95% CI: 0.89, 3.74) and the adjusted HR ($HR_{\text{adjusted by stage}}$: 1.89, 95% CI = 0.93, 3.87).

There are other applications of survival analysis that can be explored in Stata, including time-dependent predictors, competing risks regression, parametric survival models, and multilevel parametric regression. These topics are beyond the scope of this book, but an extensive review of survival analysis can be found in Collett (2003), Peace (2009), Royston and Lambert (2011), and Wienke (2011).

Downloaded by [University of California, San Francisco (CDL)] at 11:43 24 August 2017