CHAPTER 10

Survival Analysis

Survival analysis is based on data where patients are followed over time until the occurrence of a particular event such as death, relapse, recurrence, or some other event that is of interest in to the investigator. Of particular interest is the construction of an estimate of a survival curve which is illustrated in Figure 10.1. Survival probability at t represents the probability that an event does not occur by time t.

In the figure, the x-axis shows time in years and the y-axis survival probability. In this case, the function $S(t)$ is a Weibull curve with $S(t) = \exp(-(\lambda t)^{\beta})$ and $\lambda = 0.4$ and $\beta = 2.0$. In a clinical study, $S(t)$ represents the probability that the time from initiation in the study ($t = 0$) until the occurrence of the event for an arbitrary patient is greater than a specified value t. The curve represents the value for this probability as a function of t. Data on the observed time to the event for each patient is the information to use to estimate the survival curve for all t in $(0, \infty)$. See Section 10.4.2 for more details on the Weibull family of survival curves.

The survival curve or the comparison of two or more survival curves is often important in determining the effectiveness of a new treatment. It can be used for efficacy as in the case of showing that an anticoagulant is effective at reducing stroke for patients with atrial fibrillation. More often, it is used as a safety parameter, such as in the

The Essentials of Biostatistics for Physicians, Nurses, and Clinicians,
First Edition. Michael R. Chernick.

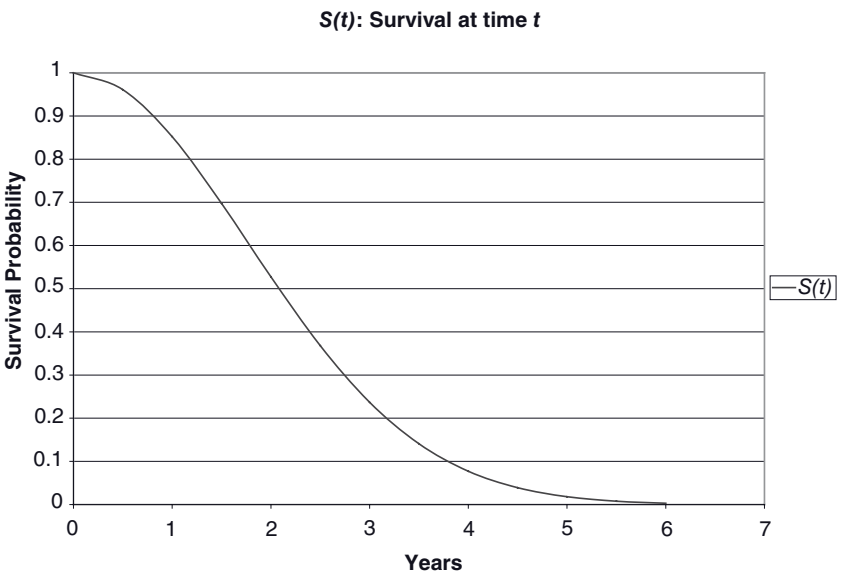

Figure 10.1. A typical survival curve.

determination of a particular adverse event that the treatment is suspected to cause. The term survival analysis came about because it was originally used when mortality was the outcome, but it can be used for time-to-event data for any event. More generally, the curve does not necessarily have to be a function of time. It is even possible for time to be replaced by a variable that increases with time, such as the cost of a worker's compensation claim where the event occurs when the claim is closed.

10.1 TIME-TO-EVENT DATA AND RIGHT CENSORING

What characterizes survival data is that some patients have incomplete results. In a particular study, there is a time at which the study ends and the data must be analyzed. At that point, some of the patients may not have experienced the event (either because they will never have the event or because the event will occur some time later). The data for these patients should not be thrown out because that would (1) ignore valuable information about the time to event, since these patients time

to event must at least be longer than the time from study initiation until the termination of the study (called the censoring time); and (2) leaving them out biases the estimate of parameters, such as median or mean survival time, since the censored observations are more likely to be the longer times than those that were not censored. So from (2), we see that the median time-to-event is underestimated if the censored data are ignored. Other censoring could occur if the patient becomes lost to follow-up prior to the date of completion for the study.

What makes survival analysis different is the existence of incomplete data on some patients whose time to event is right censored (i.e., cut off at the end of the study). The key to the analysis is to find parametric, semi-parametric, or nonparametric ways to estimate the survival curve utilizing both the complete and incomplete observations. This will often allow for a less biased median survival time estimate. The remainder of the chapter will cover various methods.

The first method is the life table. Although the methods we describe here are straightforward, there are many practical difficulties. One of these is the problem of unreported events. This is a very big problem with medical devices. Attempts have been made to address the issue of bias in estimates due to underreporting. But these methods must rely heavily on assumptions about the underreporting. The article by Chernick et al. (2002) covers the issue in detail.

10.2 LIFE TABLES

The survival curve $S(t)$ is defined to be equal to the probability that $T > t$ where T is the random variable representing the time to the event. The data is Table 10.1 is taken from Altman (1991, p. 367). In this example events are restricted to the time $(0, L]$ with events occurring after time L, right censored.

We notice from the table that patients are accrued over time for slightly less than 6 months. The study is terminated at 18 months after the first patient is enrolled in the study. Four patients died during the trial six were either living at the end of the trial or lost to follow-up. Specifically, patients 1, 5, 7, and 10 died, patients 3, 6, 8, and 9 completed the study alive and patients 2 and 4 were lost to follow-up. This table provides us with exactly all we need to construct the various types of survival curves.

Table 10.1
Survival Times for Patients

Censor code*	Patient no.	Time at entry (months)	Time to death or censoring (months)	Survival time (months)
1	1	0.0	11.8	11.8
0	2	0.0	12.5	12.5†
0	3	0.4	18.0	17.6†
0	4	1.2	4.4	3.2†
1	5	1.2	6.6	5.4
0	6	3.0	18.0	15.0†
1	7	3.4	4.9	1.5
0	8	4.7	18.0	13.3†
0	9	5.0	18.0	13.0†
1	10	5.8	10.1	4.3

*Death occurred = 1, censoring = 0, $L = 18.0$.

†Censored observation.

Life tables give survival probability estimates for intervals of time whereas survival curves are continuous over time (although their nonparametric estimates are step functions that only change when events occur). Life tables must be used when the only information that is available is the number of events occurring in the intervals. If we have the exact times when each event occurs, and all the times when censoring occurs, we can estimate the survival curve by parametric or nonparametric methods.

We can also create a life table by choosing time intervals and counting the number of events and censoring times that occur in each specified interval. However, the use of life tables when we have the exact times for the events and censoring is inefficient, since it ignores some of the available information about survival (namely, where in the interval each event occurs). In addition to the interval survival probability, the life table provides an estimate of the cumulative survival probability at the end of the time interval for each interval. Whether we are estimating cumulative survival over time or for life table intervals, there is a key equation that is exploited. It is shown as Equation 10.1.

$$S(t_2) = P(t_2 \mid t_1)S(t_1) \text{ for any } t_2 > t_1 \geq 0, \tag{10.1}$$

Table 10.2
Life Table for Patients From Table 10.1

Time interval I_j	No. of deaths in I_j	No. withdrawn in I_j	No. at risk in I_j	Avg. No. at risk in I_j	Est. prop. of deaths in I_j	Est. prop. Surv. at end of I_j	Est. cum. surv. at end of I_j
[0, 3)	1	0	10	10	0.1	0.9	0.9
[3, 6)	2	1	9	8.5	0.235	0.765	0.688
[6, 9)	0	0	6	6	0.0	1.0	0.688
[9, 12)	1	0	6	6	0.167	0.833	0.573
[12, 15)	0	3	5	5	0	1.0	0.573
[15, 18)	0	2	2	2	0	1.0	0.573
[18, ∞)	0	0	0	0	—	—	—

where $S(t)$ = survival probability at time $t = P(T > t)$, t_1 is the previous time of interest, and t_2 is some later time of interest (for a life table, t_1 is the beginning of the interval, and t_2 is the end of the interval.

For the life table, we must use the data as in Table 10.1 to construct the estimates that we show in Table 10.2. In the first time interval, say $[0, a]$, we know that $S(0) = 1$ and $S(a) = P(a|0)S(0) = P(a|0)$. This is gotten by applying Equation 10.1, with $t_1 = 0$ and $t_2 = a$, and substituting 1 for $S(0)$. The life table estimate was introduced by Cutler and Ederer (1958), and therefore it also is sometimes called the Cutler–Ederer method. We exhibit the life table as Table 10.2, and then will explain the computations.

In constructing Table 10.2 from the data displayed in Table 10.1, we see that including event times and censoring times, the data range from 1.5 to 17.6 months. Note that since time of entry dies not start at the beginning of the study, the time to event is shifted by subtracting the time of entry from the time of the event (death or censoring). We choose to create 3-month intervals out to 18 months. The seven intervals comprising all times greater than 0 are: (0, 3), [3, 6), [6, 9), [9, 12), [12, 15), [15, 18) and [18, ∞). Intervals denoted $[a, b)$ include the number "a" and all real numbers up to but not including "b." Intervals (a, b) include all real numbers greater than "a" and less than "b" but do not include "a" or "b." In each interval, we need to

determine the number of subjects who died during the interval, the number withdrawn during the interval, the total number at risk at the beginning of the interval, and the average number at risk during the interval.

To understand Table 10.2, we need to explain the meaning of the column heading.

Column 1 is labeled "Time interval" and is denoted I_j for the jth interval.

Column 2 contains the number that died in the jth interval and is denoted D_j.

Column 3 contains the number that withdrew during the jth interval and is denoted W_j.

Column 4 contains the number at risk at the start of the jth interval and is denoted N_j.

Column 5 is the average number at risk during the jth interval and is denoted $N_j^{'}$.

Column 6 is the estimated proportion of deaths during the interval and is denoted as q_j.

Column 7 is the estimated proportion of subjects surviving the interval and is denoted by p_j.

Column 8 is the cumulative probability of surviving the interval.

We note that the deaths are determined just by counting the deaths with event time falling in the interval. The withdrawals are simply determined by counting the number of censoring times falling in the interval. The number at risk at the beginning of the interval is just the total at time 0 minus all deaths and withdrawals that occurred from time 0 up to but not including time "a" where "a" is the beginning time for the interval.

Now the average number remaining over the jth interval is $N_j' = N_j - (W_j/2)$. We then get the estimated proportion that are dead, to be $q_j = D_j/N_j'$. Then the estimated proportion surviving the interval is $p_j = 1 - q_j$. Remember the key recursion in Equation 10.1? It gives $S_j = p_j S_{j-1}$. This recursive equation allows S_1 to be determined from the known value S_0 after calculating p_1. Then S_2 is calculated using S_1 and p_2, and this continues up to the time of the last event or censor time.

This is the Cutler–Ederer method, and just about any life table is generated in a very similar fashion.

10.3 KAPLAN–MEIER CURVES

The Kaplan–Meier curve is a nonparametric estimate of the survival function (see Kaplan and Meier 1958). It is computed using the same conditioning principle that we used for the life table. However, here we estimate the survival at every time point, but only do the iterative computations at the event or censoring times. The estimate is taken to be constant between points. It has sometimes been called the product limit estimator, because at each time point, it is calculated as the product of conditional probabilities. Next, we describe in detail how the curve is estimated.

10.3.1 The Kaplan–Meier Curve: A Nonparametric Estimate of Survival

For all time from 0 to t_1, where t_1 is the time of the first event, the Kaplan–Meier survival estimate is $S_{km}(t) = 1$. At time t_1, $S_{km}(t_1) = S_{km}(0)$ $(n_1 - D_1)/n_1$, where n_1 is the total number at risk, and D_1 is the number that die (have an event) at time t_1. Since $S_{km}(0) = 1$, $S_{km}(t_1) = (n_1 - D_1)/n_1$. For the example in Table 10.3, below we see that $S_{km}(t_1) =$

Table 10.3
Kaplan–Meier Survival Estimates for Example in Table 10.1

Time	No. of deaths in D_j	No. withdrawals W_j	No. at risk n_j	Est. prop. of deaths q_j	Est. prop. surviving $p_{j=1} - q_j$	Est. cumulative survival $S_{km}(t_j)$
$t_1 = 1.5$	1	0	10	0.1	0.9	0.9
$t_2 = 4.3$	1	1	9	0.125	0.875	0.788
$t_3 = 5.4$	1	0	6	0.143	0.857	0.675
$t_4 = 11.8$	1	0	6	0.167	0.833	0.562
$18 > t > 11.8$	0	5	5	0	1.0	0.562
$t \geq 18$	0	0	0	0	—	—

$(10 - 1)/10 = 0.9$. At the next death time t_2, $S_{km}(t_2) = S_{km}(t_1)(n_2 - D_2)/n_2$. For n_2, we use the value of N_2 in Table 10.2, and get $S_{km}(t_2) = (0.9)(8 - 1)/8 = 0.9(7/8) = 0.9(0.875) = 0.788$. Note that $n_2 = 8$ because there was one withdrawal between time t_1 and t_2. The usual convention is to assume "deaths before losses." This means that if events occur at the same time as censored observations, the censored observations are left in the patients at risk for each event at that time and removed before the next event occurring at a later time.

We notice a similarity in the computations when comparing Kaplan–Meier with the life table estimates. However the event times do not coincide with the endpoints of the intervals and this leads to quantitative differences. For example, at t = 4.3, the Kaplan–Meier estimate is 0.788, whereas the life table estimate is 0.688. At t = 5.4, the Kaplan–Meier estimate is 0.675 whereas the life table is 0.688. At and after t = 11.8 the Kaplan–Meier estimate is 0.562, and the life table estimate is 0.573. Although there are numerical differences qualitatively, the two methods give similar results.

10.3.2 Confidence Intervals for the Kaplan–Meier Estimate

Approximate confidence intervals at any specific time t can be obtained by using Greenwood's formula for the standard error of the estimate and the asymptotic normality of the estimate. For simplicity, let S_j denote $S_{km}(t_j)$. Greenwood's estimate of variance is $V_j = S_j^2[\sum_{i=1}^{j} q_i/(n_i p_i)]$. Greenwood's approximation for the 95% confidence interval at time t_j is $\left[S_j - 1.96\sqrt{V_j}, S_j + 1.96\sqrt{V_j}\right]$.

Although Greenwood's formula is computationally easy through a recursion equation, the Peto approximation is much simpler. The variance estimate for Peto's approximation is $U_j = S_j^2(1 - S_j)/n_j$. Peto's approximation for the 95% confidence interval at time t_j is $\left[S_j - 1.96\sqrt{U_j}, S_j + 1.96\sqrt{U_j}\right]$.

Dorey and Korn (1987) have shown that Peto's method can give better lower confidence bounds than Greenwood's, especially at long follow-up times where there are very few patients remaining at risk. In the example in Table 10.3, we shall now compare the Peto 95% confidence interval with Greenwood's at time $t = t_3$. For Greenwood, we

need to calculate V_3, which requires recursively calculating V_1 and V_2 first.

$V_1 = (0.9)^2[0.1/\{10[0.9]\})] = (0.9)(0.01) = 0.009$. Then

$$V_2 = S_2^2[q_2/(n_2 p_2) + V_1/S_{j-1}^2] = (0.788)^2[0.125/\{8(0.875) + 0.009/(0.9)^2\}]$$
$$= 0.621(0.0179 + 0.0111) = 0.621(0.029) = 0.0180.$$ Finally,

$V_3 = (0.675)^2[0.143/\{7(0.857)\} + 0.018/(0.788)^2] = 0.4556[0.143/6] = 0.0109$. So the 95% Greenwood confidence interval is

$$\left[0.675 - 1.96\sqrt{0.0109}, 0.675 + 1.96\sqrt{0.0109}\right]$$
$$= [0.675 - 0.2046, 0.675 + 0.2046] = [0.4704, 0.8796].$$

For Peto's estimate of variance, U_3, we simply calculate

$$U_3 = S_3^2(1 - S_3)/n_3 = (0.675)^2(1 - 0.675)/7$$
$$= (0.675)^2(0.325)/7 = 0.4556.(0.0464) = 0.0212\ .$$

So Peto's estimate is

$$\left[0.675 - 1.96\sqrt{0.0212}, 0.675 + 1.96\sqrt{0.0212}\right]$$
$$= [0.675 - 0.285, 0.675 + 0.285] = [0.390, 0.960].$$

In this example, we see that Peto's interval is much wider and hence more conservative than Greenwood's. However, that does not necessarily make it more accurate. Both methods are just approximations, and we cannot say that one is always superior to the other.

10.3.3 The Logrank and Chi-Square Tests: Comparing Two or More Survival Curves

To compare two survival curves in a parametric family of distributions, such as the negative exponential or the Weibull distribution, we only need to test for differences in the parameters. However, for a nonparametric estimate, we look for departures in the two Kaplan–Meier curves. The logrank test is a nonparametric test for testing equality of two survival curves against the alternative of some difference. Details about the test can be found in the original work of Mantel (1966) or in texts such as Lee (1992, pp. 109–112) or Hosmer et al. (2008).

Rather than go into the detail of computing the logrank test for comparing the two survival curves, we can conduct a similar test that

Table 10.4
Computation of Expected Numbers for the Chi-Square Test in the Breast Cancer Example

Remission time T	Number of remissions at T d_T	Number at risk in treatment group n_1	Number at risk in control group n_2	Expected frequency in treatment group E_1	Expected frequency in control group E_2
15	1	5	5	0.5	0.5
18	1	4	4	0.5	0.5
19	2	3	3	1.0	1.0
20	1	3	1	0.75	0.25
23	1	2	0	1.0	0.0
Total	—	—	—	3.75	2.25

has an asymptotic chi-square distribution with $k - 1$ degrees of freedom, where k is the number of survival curves being compared. For comparing two curves, the test statistic is chi-square with 1 degree of freedom under the null hypothesis. The chi-square statistic as usual takes the form $\sum_{i=1}^{k}(O_i - E_i)^2/E_i$, where n is the number of event times.

The expected values E_i are computed by pooling the survival data and computing the expected numbers in each group based on the pooled data (which is the expected number when the null hypothesis is true, and we condition on the total number of events at the event time points and sum up the expected numbers. Our example is from a breast cancer trial.

In the breast cancer study, the remission times for the treatment group, getting cyclophosphamide, methatrexate, and fluorouracil (CMF), are 23 months, and four patients censored at 16, 18, 20, and 24 months. For the control group, remission times were at 15, 18, 19, 19, and 20, and there were no censoring times. Table 10.4 shows the chi-square calculation for expected frequencies in the treatment and control groups in a breast cancer trial.

Based on the table above, we can compute the chi-square statistic, $(1 - 3.75)^2/4.75 + (5 - 2.25)^2/2.25 = 1.592 + 3.361 = 4.953$. From the chi-square table with 1 degree of freedom, we see that a value of 3.841 corresponds to a p-value of 0.05 and 6.635 to a p-value of 0.01. Hence, since $3.841 < 4.953 < 6.635$, we know that the p-value for this test is

between 0.01 and 0.05, and the survival curves differ significantly at the 5% level.

The logrank test is very similar except that instead of E_i in the denominator, we compute $V = \sum_{i=1}^{m} v_i$, where m is the number of time points for events from the pooled data, and $v_i = n_{1i}n_{2i}d_i(n_i - d_i)/[n_i^2(n_i - 1)]$, where n_{1i} = number at risk in group 1 at time t_i, n_{2i} = number at risk at time t_i in group 2, $n_i = n_{1i} + n_{2i}$, and d_i = combined number of deaths (events pooled from all groups) that have occurred by time t_i. For two groups, the logrank test also has an approximate chi-square distribution with 1 degree of freedom under the null hypothesis. A nice illustration of the use of the logrank test with the aid of SAS software can be found in Walker and Shostak (2010). Additional examples of two-sample and k-sample tests can be found in many standard references on survival analysis, including, for example, Hosmer et al. (2008).

10.4 PARAMETRIC SURVIVAL CURVES

When the survival function has a specific parametric form, we can estimate the survival curve by estimating just a few parameters (usually 1 to 4 parameters). We shall describe two of the most common parametric models, the negative exponential and the Weibull distribution models.

10.4.1 Negative Exponential* Survival Distributions

The negative exponential survival distribution is a one-parameter family of probability models determined by a parameter λ, called the rate parameter or failure rate parameter. It has been found to be a good model for simple product failures, such as the electric light bulb. In survival analysis, we have several related functions. For the negative exponential model, the survival function $S(t) = \exp(-\lambda t)$, where $t \geq 0$ and $\lambda > 0$. The distribution function $F(t) = 1 - S(t) = 1 - \exp(-\lambda t)$, $f(t)$ is the density function, which is the derivative of $F(t)$, $f(t) = \lambda\exp(-\lambda t)$. The hazard function $h(t) = f(t)/S(t)$. For the negative exponential model,

* Also simply referred to as the exponential distribution.

Table 10.5
Negative Exponential Survival Estimates for Patients From Table 10.3

Time T	Number of deaths Dj	Number of withdrawals Wj	Number at risk n_j	Est. prop, of deaths q_j	Est. prop. surviving p_j	KM survival estimate	Negative exp. survival estimate
1.5	1	0	10	0.1	0.9	0.9	0.940
4.3	1	1	8	0.125	0.875	0.788	0.838
5.4	1	0	7	0.143	0.857	0.675	0.801
11.8	1	0	6	0.167	0.833	0.562	0.616
18	0	5	5	0	1	0.562	0.478

$h(t) = \lambda\exp(-\lambda t)/\exp(-\lambda t) = \lambda$. In this case, we will fit an exponential model to the data used to fit the Kaplan–Meier curve in Table 10.3. Table 10.5 compares the estimated negative exponential survival curve with the Kaplan–Meier estimate.

The exponential survival curve differs markedly from the Kaplan–Meier curve, indicating that the negative exponential does not adequately fit the data.

10.4.2 Weibull Family of Survival Distributions

The Weibull model is more general and involves two parameters λ and β. The negative exponential is the special case of a Weibull model, when $\beta = 1$. The Weibull is common in reliability primarily because it is the limiting distribution for the minimum of a sequence of independent identically distributed random variables. In some situations, a failure time can be the first of many possible event times, and hence is a minimum. So under common conditions, the Weibull occurs as an extreme value limiting distribution similar to the way the normal distribution is the limiting distribution for sums or averages.

For the Weibull model $S(t) = \exp(-(\lambda t)^{\beta})$, $F(t) = 1 - \exp(-(\lambda t)^{\beta})$, $f(t) = \lambda\beta(\lambda t)^{\beta-1}\exp[-(\lambda t)^{\beta}]$, and $h(t) = \lambda\beta(\lambda t)^{\beta-1}$. For the Weibull model, $\lambda > 0$ and $\beta > 0$.

10.5 COX PROPORTIONAL HAZARD MODELS

The Cox proportional hazards regression model is called semi-parametric because it includes regression parameters for covariates (which may or may not be time dependent), but in terms of the baseline hazard function, it is completely general (hence not parametric). So part of the modeling is parametric, and another part is nonparametric, hence the term semi-parametric. In SAS®, the model can be implemented using the procedure PHREG, or STCOX in STATA. An excellent and detailed treatment with SAS applications can be found in Walker and Shostak (2010, pp. 413–428). A similar treatment using the STATA software package can be found in Cleves et al. (2008).

The purpose of the model is to test for the effects of a specific set of k covariates on the event times. These covariates can be numerical or categorical. In the case of categorical variables, such as treatment groups, the model can estimate relative risks for the occurrence of an event in a fixed interval when the patient gets treatment A versus when the patient gets treatment B.

For example, in the RE-LY trial to compare three treatments, two doses of dabigatran and warfarin as a control, the Cox model was used to estimate the relative risk of the patient getting a stroke during the trial while on one treatment versus another. This ratio was used to test for superiority or noninferiority of the dabigatran doses versus warfarin with respect to stroke or systemic embolism as the event. The model was also used for other types of event, with major bleeding being a primary safety endpoint.

The model is defined by its hazard function $h(t) = \lambda(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m)$, where m is the number of covariates the X_i are the covariates and is the baseline hazard function (t represents time). We only consider $t \geq 0$. It is called a proportional hazards model because $h(t)$ is proportional to $\lambda(t)$, since $h(t)/\lambda(t)$ is a constant (does not depend on t) that is determined by the covariates. The parameters β_i are estimated by maximizing the partial likelihood. The estimation procedure will not be described here, but its computation requires the use of numerical methods and high-speed computers.

There are many books on survival analysis that cover the Cox model, and even some solely dedicate to the method. A recent text providing an up-to-date theoretical treatment is O'Quigley (2008), which includes over 700 references. Other texts worthy of mention are

Cox and Oakes (1984), Kalbfleisch and Prentice (1980, 2002), Therneau and Grambsch (2000), Lachin (2000), Klein and Moeschberger (2003, paperback 2010), Hosmer and Lemeshow (1999), Cleves et al. (2008), Klein and Moeschberger (2003), and Hosmer et al. (2008). There have been a number of extensions of the Cox model, including having the covariates depend on time. See Therneau and Grambsch (2000) if you want a lucid and detailed account of these extensions. Parametric regression models for survival curves can be undertaken using the SAS procedure LIFEREG and the corresponding procedure STREG in STATA.

10.6 CURE RATE MODELS

The methods for analysis of cure rate models are similar to those previously mentioned, and require the same type of survival information. However, the parametric models previously described all have cumulative survival curves tending to zero as time goes to infinity. For cure rate models, a positive probability of a cure is assumed. So the cumulative survival curve for a cure rate model converges to $p > 0$ as time goes to infinity, where p is called the cure probability, cure fraction or cure rate. Often the goal in these models is to estimate p.

For nonparametric methods such as the Kaplan–Meier approach, p is difficult to detect. It would be the asymptotic limit as t gets larger, but the Kaplan–Meier curve gives us no information about the behavior of the survival curve beyond the last event time or censoring time (whichever is last). So to estimate the cure rate requires a parametric mixture model.

The mixture model for cure rates was first introduced by Berkson and Gage (1952). The general model is given by the following equation:

$$S(t) = p + (1 - p)S_1(t)$$

where p is the cure probability, and $S_1(t)$ is the survival curve for those who are not cured. $S_1(t)$ is the conditional survival curve given the patient is not cured. The conditional survival curve can be estimated by parametric or nonparametric methods. For an extensive treatment of cure rate models using the frequentist approach, see Maller and Zhou

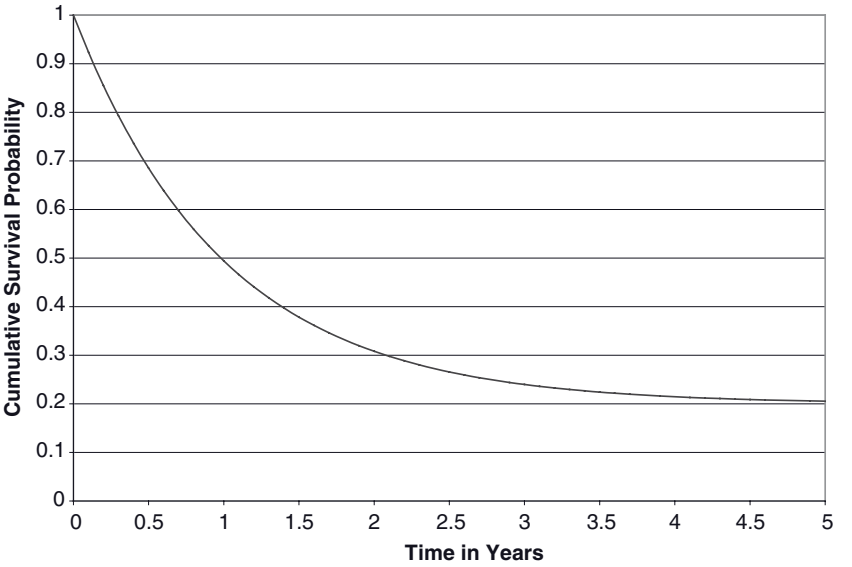

Figure 10.2. Exponential cure rate model with cure rate $p = 0.20$ and exponential rate parameter $\lambda = 1$.Sente videm patum ad inam nonvere timorio rterumunina nihi, catum

(1996). The Bayesian approach to cure rate models can be found in Ibrahim et al. (2001).

We illustrate a parametric mixture survival curve with an exponential survival curve with rate parameter $\lambda = 1$, for the conditional survival curve $S_1(t)$ and with survival probability $p = 0.2$. This curve is shown in Figure 10.2.

Although cure rate modeling began with Berkson and Gage in the 1950s, much of the literature came about in the 1990s when computing became much faster and the EM algorithm for the frequency approach and MCMC methods for Bayesian approaches became easy to implement. Until recently, the free software WinBUGS was the main option for doing MCMC methods for the Bayesian approach to modeling. However, very recently in SAS Version 9.2, MCMC methods have been added as a procedure in SAS/STAT. Users of SAS software may find this more convenient.

10.7 EXERCISES

1. Define the following:
- **(a)** Life table
- **(b)** Kaplan–Meier curve
- **(c)** Negative exponential survival distribution
- **(d)** Cure rate model
- **(e)** Chi-square test to compare two survival curves

2. If the survival function $S(t) = 1 - t/b$ for $0 \le t \le b$, where b is a fixed positive constant, calculate the hazard function. When is the hazard function lowest? Is there a highest rate?

3. Suppose + denotes a censoring event, and that the event times in months for group1 are [8.1, 12, 17 33+, 55, and 61] while for group 2 they are [32, 60, 67, 76+, 80+, and 94]. Test to see if the survival curves are different using the chi-square test.

4. Suppose the survival time since a bone marrow transplant for eight patients who received the transplant is 3, 4.5, 6, 11, 18.5 20, 26, and 35. No observations were censored.
- **(a)** What is the median survival time for these patients?
- **(b)** What is the mean survival time?
- **(c)** Construct a life table where each interval is 5 months.

5. Using the data in example 4:
- **(a)** Calculate a Kaplan–Meier curve for the survival distribution
- **(b)** Fit a negative exponential model.
- **(c)** Compare b with a.
- **(d)** Is the negative exponential survival distribution a good fit in this case?

6. Modify the data in example 4 by making 6, 18.5, and 35 censoring times
- **(a)** Estimate the median survival time.
- **(b)** Why would an average of all the survival times (excluding the censoring times) be inappropriate?
- **(c)** Would an average including the censoring times be appropriate?

7. Now using the data as it has been modified in exercise 6, repeat exercise 5a.

8. Listed below are survival and censoring times (using the + sign for censoring) for six males and six females.

Males: 1, 3, 4+, 9, 11, 15

Females 1, 3+, 6, 9, 10, 11+

(a) Calculate the Kaplan–Meier curve for males

(b) Calculate the Kaplan–Meier curve for females

(c) Test for a difference between the male and female survival curves using the chi-square test.

(d) Compute the logrank statistic and perform the same test as in *c* using this statistic? Do you reach the same conclusion as in *c*? Are the chi-square and logrank test statistics close in value? Are the *p*-values nearly the same?

9. What assumptions are required for the Cox proportional hazard model? Why is it called a semi-parametric method?

10. Suppose a cure model is known to have $S_1(t) = \exp(-0.5t)$.

Recall $S(t) = p + (1 - p)S_1(t)$. Suppose that we know that $S(2) = 0.5259$. Can you calculate the cure rate for this model? If so what is it?