Applied Survival Analysis

Regression Modeling of Time to Event Data

DAVID W. HOSMER, Jr.

Department of Biostatistics and Epidemiology
University of Massachusetts
Amherst, Massachusetts

STANLEY LEMESHOW

Department of Biostatistics and Epidemiology
University of Massachusetts
Amherst, Massachusetts

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

where

$$d_{k+} = \sum_{i=1}^{m} d_{ki},$$

and $\hat{e}_{k+}$ is defined similarly. If we calculate Q_c and reject the hypothesis of equal survival experience, then we would reject using Q.

The estimated survivorship functions for the four age groups are shown in Figure 2.8. The figure confirms our preliminary observations based on estimates of median survival times. We see that the survivorship function for the youngest group lies completely above those of the other three groups. It has a long right tail and does not go to zero since two observations are censored at 60 months. For the first 15 months, the estimated survivorship functions for the youngest three age groups follow the trend observed in the medians. In this interval, the three functions are, for the most part, inversely ordered by age. The functions for the middle two age groups cross four times between 15 and 45 months, suggesting that the survival experience for these two age groups may be similar in this range. The estimated survivorship function for the oldest age group lies completely below that of the other three groups for 34 months. This suggests that we should begin our analysis with a test for the overall equality of the survivorship experience. If we find that the experience of at least one group is different from the others, we should construct single degree-of-freedom contrasts to examine between-group differences, as is typically done in analysis of variance methods.

The values of the four test statistics using their respective weights in (2.27) are given in Table 2.14. Since each statistic is significant at beyond the 1 percent level, we reject the hypothesis that the survivorship functions for the four age groups are the same. We follow the test for overall group differences in survival experience with contrasts to try and describe more precisely the source(s) of the significance of the overall test. The BMDP package, program 1L, offers this option by allowing the user to specify a trend test and to input a set of coefficients to test for trend when the groups are not equally spaced. The SAS package lifetest procedure has a test option that provides a trend test for a numeric covariate. The test does not yield the same numeric value as the trend test in BMDP. We describe the test used in BMDP as it follows directly from the multiple group test in (2.27). The null hypothesis is that the survivorship functions are equal and the alternative is that they are rank-ordered and follow the trend specified by the coefficients denoted by the vector $\mathbf{c}' = (c_1, c_2, \ldots, c_{K-1})$. If the groups are equally

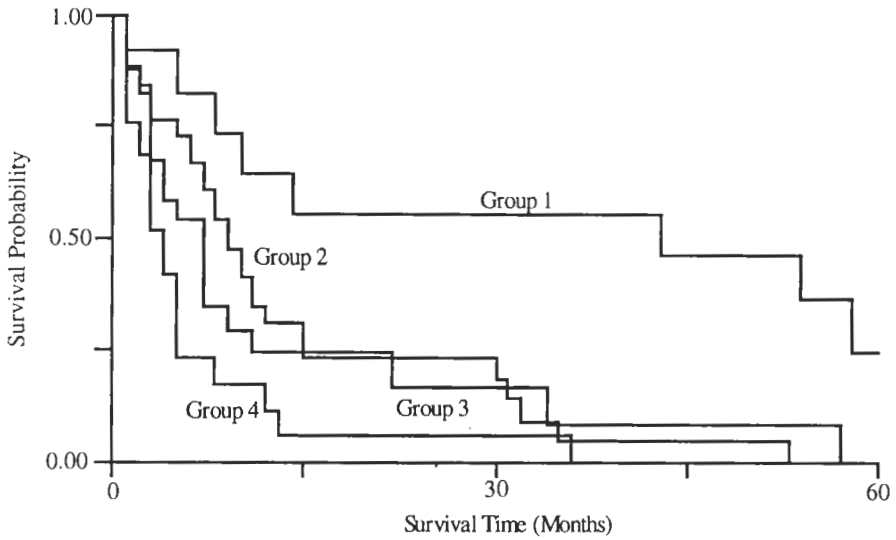

Figure 2.8 Estimated survivorship functions for the four age groups in the HMO-HIV+ study.

spaced, we may use $c_k = k$. The age groups we used in the HMO-HIV+ study are not equally spaced so we will use a vector of coefficients whose values are the midpoints of the four groups, i.e.,

$$\mathbf{c}' = (25, 32.5, 37.5, 47.5).$$

Any linear transformation of these coefficients would yield the same value of the test statistic. The statistic to test for trend, with one degree-of-freedom, is

$$Q_{\text{trend}} = \frac{\left[\mathbf{c}' \sum_{i=1}^{m} \mathbf{W}_i (\mathbf{d}_i - \hat{\mathbf{e}}_i)\right]^2}{\mathbf{c}' \left[\sum_{i=1}^{m} \mathbf{W}_i \hat{\mathbf{V}}_i \mathbf{W}_i\right] \mathbf{c}}. \tag{2.28}$$

The p-value is computed using the chi-square distribution with one degree-of-freedom, i.e., $p = \Pr(\chi^2(1) \geq Q_{\text{trend}})$. Table 2.15 presents the statistics and their p-values for the test of trend among the four age groups in the HMO-HIV+ study. These values are each just slightly

Table 2.14 Test Statistics, Degrees-of-Freedom and *p*-Values for the Equality of the Survivorship Functions for the Four Age Groups in the HMO-HIV+ Study

Statistic	Value	df	*p*-Value
Log-rank	19.906	3	<0.01
Generalized Wilcoxon	14.143	3	<0.01
Tarone–Ware	16.956	3	<0.01
Peto–Prentice (BMDP)	15.665	3	<0.01

smaller than the values in Table 2.14, providing strong evidence for a trend in survival experience that is inversely related to age. We explore this relationship in more detail when we consider regression modeling in the next chapter.

In the examples we have used from the HMO-HIV+ study to illustrate the comparison of the survivorship functions over groups, the magnitude of the test statistics has not varied too dramatically with the choice of weight, and the significance or non-significance of all test statistics has been consistent. However, this is not always the case and to illustrate this we use some data provided to us by our colleagues Drs. Carol Bigelow and Penny Pekow (at the University of Massachusetts) and Dr. Kathy Meyer (at Baystate Medical Center in Springfield, Massachusetts). These data were used as part of Ms. Shiaw-Shyuan Yuan's Masters degree project [Yuan (1993)]. The purpose of the study was to determine factors which predict the length of time low birth weight infants (<1500 grams) with bronchopulmonary dysplasia (BPD) were treated with oxygen. The data were collected retrospectively for the period December 1987 to March 1991. Beginning in August 1989, the treatment of BPD changed to include the use of surfactant replacement therapy. This was done with parental permission since, at the time, this therapy was considered experimental. A total of 78 infants met the study criteria, with 35 receiving surfactant replacement therapy and 43

Table 2.15 Trend Test Statistics, Degrees-of-Freedom and *p*-Values for the Equality of the Survivorship Functions among the Four Age Groups in the HMO-HIV+ Study

Statistic	Value	df	*p*-Value
Log-rank	19.066	1	<0.01
Generalized Wilcoxon	14.080	1	<0.01
Tarone–Ware	16.673	1	<0.01
Peto–Prentice (BMDP)	15.536	1	<0.01

not receiving this therapy. Five babies were still on oxygen at their last follow-up visit and represent censored observations. We refer to this study as the BPD study.

The outcome variable is the total number of days the baby required supplemental oxygen therapy. Figure 2.9 presents the Kaplan–Meier estimates of the survivorship functions for two groups defined by use of surfactant replacement therapy. The estimated median number of days of therapy for those babies who did not have surfactant replacement therapy (group 0) is 107 {95 percent CIE: (55.3, 158.7)}, and the estimated median number of days for those who had the therapy (group 1) is 71 {95 percent CIE: (33.3, 108.7)}. The median number of days of therapy for the babies not on surfactant is about 1.5 times longer than those using the therapy, but there is considerable overlap in the confidence intervals. The plots of the survivorship functions in Figure 2.9 indicate a progressively larger difference in the survivorship experience between the two groups over time. Table 2.16 presents test statistics and associated p-values for the equality of the survivorship functions. The Wilcoxon test is not significant at the 5 percent level, but the log-rank test is significant. The difference in the magnitude of the test statistics is due to the difference in the weights used. The Wilcoxon test uses a weight equal to the size of the risk set and thus is more likely to detect early differences. The log-rank test uses a weight equal to one and is more likely to detect later differences in the survivorship functions.

In any statistical analysis in which more than one test can be used, we need to make a decision about which results we will report. The log-rank test is the most frequently used and reported test for the comparison of survivorship functions. For most analyses, at least when each test has roughly the same level of significance, reporting only the results of the log-rank test is appropriate. When the tests give different results, then more than one result should be reported. This will provide the reader with a clearer picture as to where the survivorship functions are different. The current example demonstrates the importance of computing several of the tests. Most packages have both the log-rank and generalized Wilcoxon tests, and we recommend that both be computed. To our knowledge, only BMDP computes the Tarone–Ware and Peto–Prentice tests. The pattern of censoring can influence the magnitude of the tests, but the values of the Tarone–Ware and Peto–Prentice tests tend to be intermediate between the log-rank and Wilcoxon tests.

We conclude our presentation of the tests for comparison of survivorship functions with a brief discussion of the assumptions underlying the tests and the types of alternative hypotheses the tests have the power

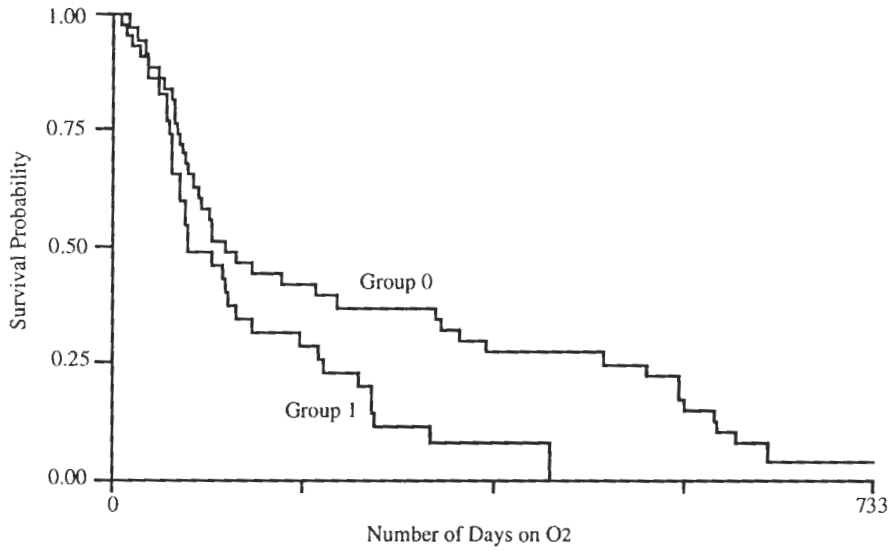

Figure 2.9 Estimated survivorship functions defined by surfactant use in the BPD study (0 = No surfactant, 1 = Surfactant).

to detect. Recall that the Kaplan–Meier estimator assumes that censoring is independent of survival time. In addition, the tests assume that the censoring is independent of the group. Problems in study design and data collection can lead to differential effects due to censoring, and the best protection is a carefully designed study. However, it is good practice to examine the censoring pattern in the data.

In general, we cannot over-emphasize the importance of a careful study of the plot of the Kaplan–Meier estimates of the survivorship functions. Any tests comparing these functions, and within-group point estimates of quantiles, should support what is seen in the plot. The plot is also the basic diagnostic tool to determine whether the tests described

Table 2.16 Test Statistics and *p*-Values for the Equality of the Survivorship Functions for Two Groups Defined by Surfactant Use in the BPD Study

Statistic	Value	df	*p*-Value
Log-rank	5.618	1	0.018
Generalized Wilcoxon	2.490	1	0.115
Tarone–Ware	3.698	1	0.055
Peto–Prentice (BMDP)	2.534	1	0.111

previously should be used or, if used, have any chance of detecting a difference. The alternative hypothesis that the tests are most likely to detect is a monotonic ordering of the survivorship functions (e.g., they lie one above another). The tests have little to no power to detect differences when the survivorship functions cross one another. An example of a worst-case scenario is when the survivorship functions for two groups have the same median and cross each other once at that value. For the early times one group has the more favorable survival experience, but for later times the other group does. None of the tests described in this section are able to detect this kind of difference. This is a situation analogous to the presence of interaction in a Mantel–Haenszel analysis of stratified contingency tables. Unfortunately, tests for interaction used with a Mantel–Haenszel analysis, such as the Breslow–Day test [Breslow and Day (1980)], can't be used, due to small cell frequencies in tables such as Table 2.13. In this case, one approach that can be used is to subdivide the sample on the basis of the stratification variable and then test for group differences within the strata. This approach is limited by the study size, as we can spread the data over only so many strata. Eventually there are too few subjects per stratum to reliably estimate the survivorship function. However, in practice, there may be one or two clinically plausible variables to use for stratification purposes. These types of differences, or interactions, between survivorship functions are much more clearly addressed using the regression modeling approach to be discussed in Chapter 3.

2.5 OTHER FUNCTIONS OF SURVIVAL TIME AND THEIR ESTIMATORS

The Kaplan–Meier estimator of the survivorship function has been, and continues to be, the most frequently used estimator, largely due to the fact that it is routinely calculated by most software packages. To motivate the discussion of another estimator, we begin by presenting a different representation of the survivorship function. If we assume that the underlying time random variable is absolutely continuous, then we may express the survivorship function as

$$S(t) = e^{-H(t)}, \tag{2.29}$$

where $H(t) = -\ln(S(t))$. The expression in (2.29) suggests that estimators of the survivorship function could be based on an estimator of $S(t)$

(e.g., the Kaplan–Meier estimator) or via an estimator of $H(t)$. Aalen (1975, 1978), Nelson (1969, 1972) and Altshuler (1970) have proposed an easily computed estimator of $H(t)$, which we refer to as the Nelson-Aalen estimator.

The work by Aalen is considered to be one of the landmark contributions to the field, as virtually all recent statistical developments for the analysis of survival time have been based on the counting process approach he used to derive his version of the estimator of $H(t)$. The statistical theory and use of this estimator in various applied settings are discussed in detail in Andersen, Borgan, Gill and Keiding (1993) and in Fleming and Harrington (1984, 1991). We will use results derived from the counting process theory to justify various techniques discussed in this text. We will not present the counting process approach in any detail since fully appreciating and understanding it requires having had calculus-based courses in mathematical statistics and probability theory.

Without providing any details as to its derivation (a heuristic argument is given later in this section), the Nelson–Aalen estimator of $H(t)$ is

$$\tilde{H}(t) = \sum_{t_{(i)} \le t} \frac{d_i}{n_i}. \tag{2.30}$$

An estimator of the survivorship function, based on (2.30), is

$$\tilde{S}(t) = e^{-\tilde{H}(t)}. \tag{2.31}$$

One theoretical problem is that the expression in (2.29) is valid for continuous time, but the estimator in (2.31) is discrete. However, the estimator in (2.31) provides the basis for the estimator of the survivorship function used with the proportional hazards regression model discussed in Chapter 3. For this reason, we consider it in some detail.

Even though packages may not provide the Nelson–Aalen estimator of the survivorship function, it is remarkably easy to compute. In the absence of ties, one merely sorts the data into ascending order on the time variable. The size of the risk set at $t_{(i)}$ is $n-i+1$ and the estimator, $\tilde{H}(t)$ in (2.30), is obtained as the cumulative sum of the zero-one censoring indicator variable divided by the size of the risk set. The Nelson-Aalen estimator of the survivorship function is obtained by evaluating the expression in (2.31). When ties are present, one sorts the data into ascending order on time and into descending order on the censoring

variable within values of time. Sorting in this way places the censored observations after the events when ties occur. One then calculates a variable equal to $n-i+1$, and uses a procedure such as STATA's collapse command, or the means procedure in SAS, to provide summary statistics at each value of time observed. One needs to obtain the maximum value of $n-i+1$ among the tied time values and the total number of events and/or censored observations. This reduced data set is used to calculate the Nelson-Aalen estimator using the cumulative sum described for the case where there are no ties.

Peterson (1977) proposed another estimator, which is based on the Kaplan–Meier estimator of the cumulative hazard function, as follows:

$$\hat{H}(t) = -\ln\left(\hat{S}(t)\right) = -\ln\left(\prod_{t_{(i)}\le t}\left(\frac{n_i - d_i}{n_i}\right)\right) = -\ln\left(\sum_{t_{(i)}\le t}\left(\frac{n_i - d_i}{n_i}\right)\right)$$

$$= \sum_{t_{(i)}\le t} -\ln\left(1 - \frac{d_i}{n_i}\right).$$

One may show, by using a Taylor series expansion (see Appendix 1), that $d_i/n_i \le -\ln\left(1-d_i/n_i\right)$ for each survival time. Thus, the Nelson–Aalen estimator of the survivorship function will always be greater than or equal to the Kaplan–Meier estimator. If the size of the risk sets relative to the number of events is large, then $d_i/n_i \cong -\ln\left(1-d_i/n_i\right)$ and there will be little practical difference between the Nelson-Aalen and the Kaplan-Meier estimators of the survivorship function.

The HMO-HIV+ study provides a good illustration of a situation in which there is little practical difference between the two estimators. Table 2.17 presents the results of collapsing the sample of 100 observations to obtain the necessary within-time summary statistics at each observed value of time: the frequency of occurrence (freq), the number of events (d), the size of the risk set (n), the Nelson–Aalen estimator, $\tilde{H}(t)$, the Nelson–Aalen estimator of the survivorship function, $\tilde{S}(t)$ and, for comparison, the Kaplan–Meier estimator, $\hat{S}(t)$. For example, at 3 months the values of the estimators are

$$\tilde{H}(3) = \frac{15}{100} + \frac{5}{83} + \frac{10}{73} = 0.347,$$

$$\tilde{S}(3) = e^{-0.347} = 0.707,$$

and

$$\hat{S}(3) = \left(1 - \frac{15}{100}\right) \times \left(1 - \frac{5}{83}\right) \times \left(1 - \frac{10}{73}\right) = 0.689.$$

Table 2.17 Summary Table Used to Calculate the Nelson-Aalen Estimator of the Survivorship Function for the HMO-HIV+ Study

Time	freq	d	n	$\tilde{H}(t)$	$\tilde{S}(t)$	$\hat{S}(t)$
1	17	15	100	0.150	0.861	0.850
2	10	5	83	0.210	0.810	0.799
3	12	10	73	0.347	0.707	0.689
4	5	4	61	0.413	0.662	0.644
5	7	7	56	0.538	0.584	0.564
6	3	2	49	0.579	0.561	0.541
7	7	6	46	0.709	0.492	0.470
8	4	4	39	0.812	0.444	0.422
9	3	3	35	0.897	0.408	0.386
10	4	3	32	0.991	0.371	0.350
11	3	3	28	1.098	0.333	0.312
12	4	2	25	1.178	0.308	0.287
13	1	1	21	1.226	0.294	0.273
14	1	1	20	1.276	0.279	0.260
15	2	2	19	1.381	0.251	0.232
19	1	0	17	1.381	0.251	0.232
22	1	1	16	1.444	0.236	0.218
24	1	0	15	1.444	0.236	0.218
30	1	1	14	1.515	0.220	0.202
31	1	1	13	1.592	0.204	0.187
32	1	1	12	1.675	0.187	0.171
34	1	1	11	1.766	0.171	0.156
35	1	1	10	1.866	0.155	0.140
36	1	1	9	1.977	0.138	0.125
43	1	1	8	2.102	0.122	0.109
53	1	1	7	2.245	0.106	0.093
54	1	1	6	2.412	0.090	0.078
56	1	0	5	2.412	0.090	0.078
57	1	1	4	2.662	0.070	0.058
58	1	1	3	2.995	0.050	0.039
60	2	0	2	2.995	0.050	0.039

The values at other times are obtained in a similar manner. Figure 2.10 presents graphs of the the Nelson–Aalen and Kaplan–Meier estimators. We see little practical difference between the two estimators, even though $\tilde{S}(t) \geq \hat{S}(t)$ at every observed value of time.

The function $H(t)$ is an important analytic tool for the analysis of survival time data. In much of the survival analysis literature it is called the *cumulative hazard function*, but in the counting process literature it is related to a function called the *cumulative* or *integrated intensity process*. The term "hazard" is used to describe the concept of the risk of "failure" in an interval after time t, conditional on the subject having survived to time t. The word "cumulative" is used to describe the fact that its value is the "sum total" of the hazard up to time t. At this point we focus on the hazard function itself, as it plays a central role in regression modeling of survival data.

Consider a subject in the HMO-HIV+ study who has a survival time of 7 months. For this subject to have died at 7 months, he/she had to be alive at 6 months. The hazard at 7 months is the failure rate "per month," conditional on the fact that the subject has lived 6 months. This is not the same as the unconditional failure rate "per month" at 7

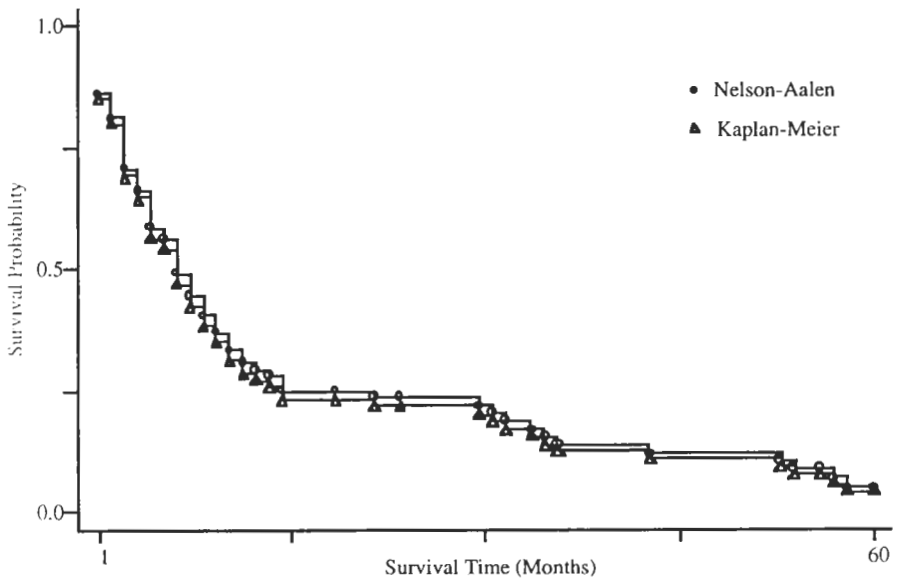

Figure 2.10 Graphs of the Nelson-Aalen and Kaplan-Meier estimators of the survivorship function from the HMO-HIV+ study.

months. The unconditional rate applies to subjects at time zero and, as such, does not use the information available as the study progresses about the survival experience in the sample. This accumulation of knowledge, over time, is generally referred to as *aging*. For example, of 100 subjects who enroll in a study, what fraction is expected to die at 7 months? The conditional failure rate applies only to that subset of the sample that has survived to a particular time, thus it accounts for the aging that has taken place in the sample.

The data from the HMO-HIV+ study can be used to demonstrate the difference between the conditional and the unconditional failure rate. If we assume that there were no censored observations in the study, the "freq" column in Table 2.17 gives the number of deaths. The first two columns of Table 2.17 are a typical presentation of grouped data. A histogram based on these data provides a graphical estimator of the unconditional failure rate.

To construct the histogram, we divide the follow-up time into 10 intervals, each of width 6 months. Each interval is represented graphically by a rectangle with height equaling the frequency drawn over the interval. To construct a relative histogram we divide each frequency by the total sample size. At this point we must decide what we wish to use as the appropriate unit of time. If we do nothing, we implicitly let 6 months denote "one unit" of time. If we wish to have "one unit" equal "one month" then we must further divide by 6. For other intervals of time, we would divide by the correct multiple of interval width and unit. If we divide by 6, the heights of the rectangles give us the relative proportions of the *total* number of subjects beginning at time "zero" who had a survival time in each interval, and the area of each rectangle is the observed unconditional failure rate per month in that interval.

For each time, t, the histogram estimator, $\hat{f}(t)$, is

$$\hat{f}(t) = \frac{(\text{freq})/(\text{width})}{n}, \tag{2.32}$$

where "freq" denotes the number of survival times in the interval, "width" denotes the width of the interval relative to the definition of "one unit" and n is the total sample size. The fact that the numerator of the estimator is expressed relative to the total sample size makes it an unconditional estimator. This is further reflected by the fact that the total area of the histogram rectangles is one, meaning that each subject has been counted once and only once in the presentation of the data.

The interval grouped-data estimator of the hazard function is, for all values of time, t, in an interval,

$$\hat{h}(t) = \frac{(\text{freq})/(\text{width})}{n(t)}, \tag{2.33}$$

where the quantity $n(t)$ is used somewhat imprecisely to denote the number of subjects still alive (at risk) at the beginning of the current interval. The area of the rectangle formed by graphing $\hat{h}(t)$ versus t estimates the conditional, on $n(t)$, per-month failure rate in the interval. The sum of the areas of the rectangles up to and including an interval is an estimate of the cumulative hazard. Since subjects are at risk until they actually die or are censored, they may be counted more than once and the sum of the areas of the rectangles may be greater than one.

Figure 2.11 presents the graphs of the histogram and hazard function estimators of the unconditional and conditional failure rates, computed from the data in Table 2.17, using 6-month intervals (e.g., (0,6], (6,12],...,(54,60]). The shaded rectangles of the histogram, which estimate the overall, unconditional per-month failure rate, are initially high and then drop rapidly, staying consistently low to 60 months. This pattern reflects the many early deaths; relatively few subjects had survival times throughout the period of follow-up. This was described by the Kaplan–Meier estimator in Figure 2.2. On the other hand, the open rectangles of the hazard function estimate the failure rate in the current interval, given that a subject is alive at the beginning of the interval. This pattern is not as consistent as that seen in the shaded histogram due to the fact that each rectangle is based on fewer subjects than the previous one. In other words, the variability is greater in the estimator of the hazard than the histogram. The graph indicates a relatively high initial failure rate which drops and then rises again.

The histogram estimator in (2.32) is useful for providing an estimate of the unconditional rate only when there are no censored observations. It may be modified to handle censored observations by using the difference between the values of the Kaplan–Meier (or Nelson–Aalen) estimator of the survivorship function at the two endpoints of the interval. The hazard function estimator in (2.33) may be modified to accommodate censored observations by having censored values of time contribute to the count in the denominator but not in the numerator. To provide a better approximation of the number at risk over the whole interval in settings in which there are large numbers of subjects and/or the inherently continuous time variable has been recorded at a few dis-

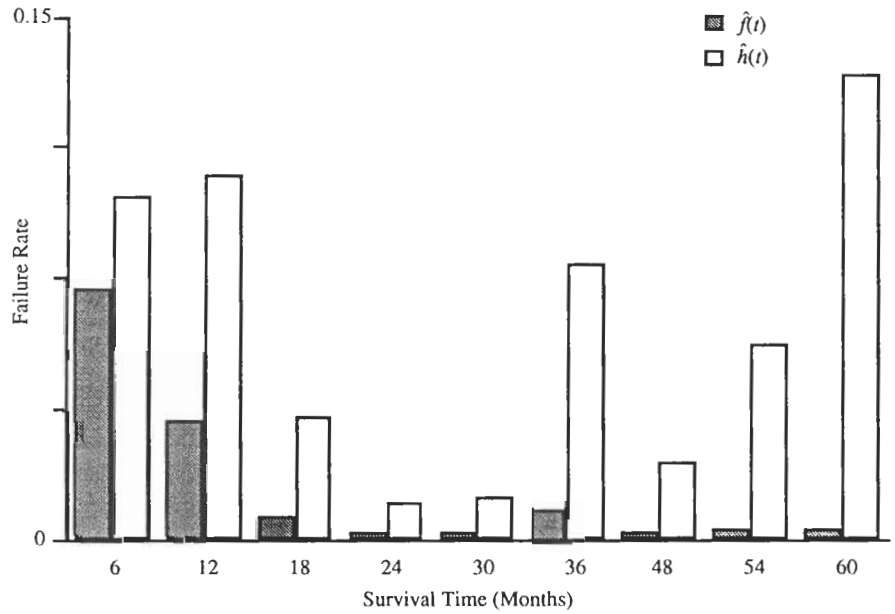

Figure 2.11 Graphs of the histogram estimator (shaded) of the unconditional failure rate and the hazard function estimator (open) of the conditional failure rate from the HMO-HIV+ study.

crete time points, the estimator of the hazard may use a denominator in which the number at risk at the beginning of the interval is reduced by one-half the number of subjects who failed, were censored or were lost for other reasons [see Lee (1992)].

Considering Figure 2.11, it is logical to postulate a function of time that describes, in a concise fashion, the form of either the unconditional or conditional failure rate, which may then be used to express the survivorship function as a function of time. If we can answer this question, then we have taken an important first step toward a more comprehensive analysis that will enable us to study which factors affect survival, namely parametrizing this function with a regression-like model.

As we think about the problem of trying to develop a function to describe survival time in the presence of censored data, we focus attention on the hazard function since it incorporates any aging that might take place. Figure 2.11 may be useful for general descriptive purposes but it is, in a sense, too discrete to be of use in developing a more precise function of time to describe the hazard function. What we would like is a more "continuous" time analysis. If we let the interval width

shrink to the point where it is one measurement unit wide (i.e., one month in the HMO-HIV+ study), then the right-hand side of the estimator of the hazard function in (2.33) is d_i/n_i at observed survival times and is zero elsewhere.

Figure 2.12 presents a scatterplot of the pairs $\left(t_{(i)}, d_i/n_i\right)$, $i = 1,2,\ldots,31$ and a lowess smooth[2] of the plot [see StataCorp (1997), ksm command]. The smoothing done here is for illustrative purposes [see Andersen, Borgan, Gill and Keiding (1993) for a more complete discussion of smoothed estimators of the hazard function]. One difficulty with the plot in Figure 2.12 is that the hazard function should be estimated to be 0 at times when no deaths occurred. The smoothed curve in Figure 2.12 does not incorporate these 0 values. However, the goal in this section is to begin to make the transition from fully nonparametric to regression models discussed in subsequent chapters. Figure 2.12, while not totally correct, does serve to guide the reader in the direction of these regression models.

The smooth of the pointwise estimates of the hazard agrees with our original impression drawn from Figure 2.11 that the conditional risk is relatively high, drops and then rises. On the basis of this observation, we might postulate that the hazard function is a quadratic function of time,

$$h(t) = \theta_0 + \theta_1 t + \theta_2 t^2.$$

Suppose for the moment that we have a parametric form for the hazard function. We need to link the hazard function in a more direct way to the survivorship function. Since we assume the time variable is absolutely continuous, the cumulative hazard is, by methods of calculus,

$$H(t) = \int_0^t h(u)\, du, \tag{2.34}$$

and by (2.29)

$$S(t) = e^{-\int_0^t h(u)\, du}. \tag{2.35}$$

Those readers familiar with calculus will recognize the right-hand side of (2.34) as the integral of the hazard function over the time interval $[0,t]$. For readers not familiar with calculus, the estimator in (2.30) can

[2] For those unfamiliar with scatterplot smoothing methods, the purpose is to remove some of the "noise" in the plot by computing, for each y in the plot, a weighted average of the other y's near it.

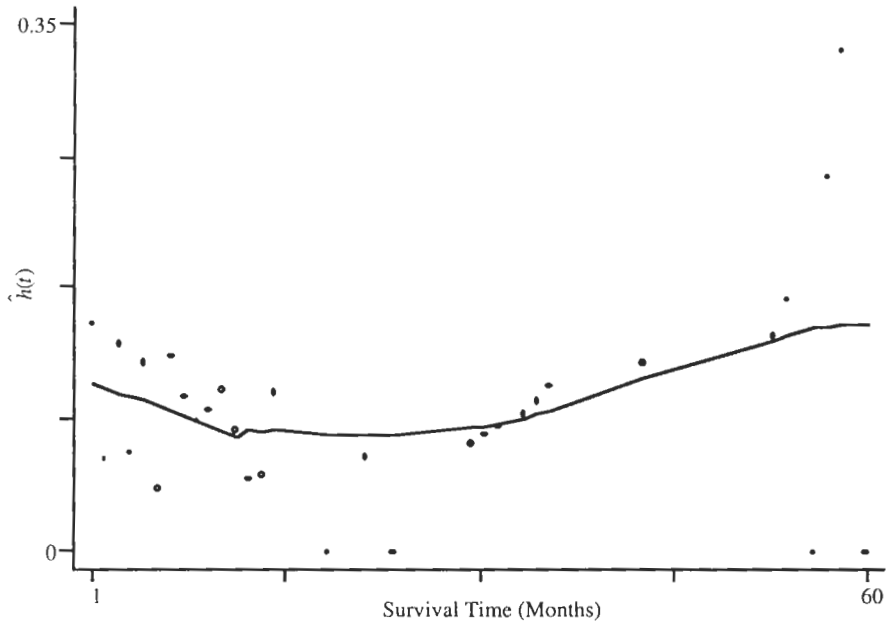

Figure 2.12 Scatterplot of the pointwise estimator of the hazard function, d_i/n, and its lowess smooth from the HMO-HIV+ study.

serve as a convenient mental model of what is being computed in (2.34). Another representation of the hazard function may be obtained by taking the log of (2.35) and then differentiating with respect to t yielding

$$h(t) = \frac{f(t)}{S(t)}, \tag{2.36}$$

where $f(t)$ denotes the probability density function for the time random variable. Those not familiar with methods of calculus may think of the function $f(t)$ as what the histogram estimator in (2.32) becomes if we use larger and larger sample sizes and the width of each interval used in its construction becomes quite small. A similar intuitive argument may be applied to the hazard function estimator in (2.33) to motivate the expression in (2.36).

As noted above, one way to envision the hazard function is to think of it as a limiting, $n \to \infty$, version of the estimator in (2.33). In this argument, we let the width of each interval become quite small and, in the

end, we have a function which describes the failure rate in the next instant following t. The expressions in (2.35)–(2.36) show that if we can specify the hazard function, then it is, in principle, relatively easy to obtain an expression for any of the other functions of survival time. The advantage of using the hazard function is that it characterizes the aging process as a function of time.

To obtain a better understanding of the hazard function and how it specifies the survivorship function, we consider various possible parametric models. A discussion of parametric survival time models is presented in Chapter 8. The goal here is see how this function describes the aging process.

The simplest possible model is for the hazard function to be constant, not depending on time [i.e., $h(t) = \theta$]. This hazard function states that at any particular time the chance that a subject "dies" in the next instant does not depend on how long the subject has survived. For example, in Figure 2.12 the average value of the plotted pointwise estimates of the hazard function is about 0.1. Thus, the constant hazard model is $\hat{h}(t) = 0.1$. The interpretation of this hazard function is that there is about a 10 percent chance that a subject will die in the next month, regardless of how long he/she has already survived. This model for the hazard may be clinically plausible in some studies of human populations when the follow-up time is relatively short. For example, the chance that a "healthy" 35-year-old person dies in the next year is about the same as that of a healthy 36- or 37- or 38- or 39-year-old subject.

The next simplest model is for the hazard to be a linear function of time, $h(t) = \theta_0 + \theta_1 t$. For example, an approximate straight-line fit to the plotted points in Figure 2.12 yields the model $\hat{h}(t) = 0.07 + 0.001t$. The interpretation is that at the beginning of the study subjects had about a 7 percent chance of dying in the next month, and this increases at about 0.1 percent per month. Since the hazard function must be greater than zero, the values of the parameters are constrained. For example, the model $h(t) = 0.12 - 0.004t$ describes the hazard in the first 30 months in Figure 2.12, but yields negative values after 30 months. This leads to the clinically implausible situation of positive probability of infinite physical life. Therefore, we have to use special methods when fitting hazard functions to observed data, since simple least squares regression methods will not be appropriate. We discuss these methods in detail in the next chapter.

On the basis of the lowess smooth in Figure 2.12, we postulated a quadratic function for the hazard function for the HMO-HIV+ study.

This is a more complicated function than the linear or constant model, but a life process of decreasing risk followed by increasing risk is clinically plausible. If one conceptualizes the risk of death in the next "instant" from birth to age 80, the function decreases for the first 5 or so years, remains fairly constant for 40 or so years and then begins to rise rapidly. This is more of a "bathtub" shape and requires a more complex function to describe it than a simple quadratic [see Lawless (1982)].

The major point is that the hazard function itself says a great deal about the fundamental underlying life-length process being studied. Specifying a fully parametric model leads to a specific life-length process. In some settings we may need this level of specificity, but in others it may not be necessary or flexible enough. This point will be dealt with directly in the next chapter.

The univariate descriptive methods discussed in this chapter, computed for the whole study or within a few subgroups, are an important first step in any analysis of survival time; however, these methods cannot be used to address the more sophisticated questions that can typically be addressed through regression modeling techniques. In Chapter 1 we discussed the general similarities and differences between regressions using dependent variables such as weight or disease status and regressions using survival time (with and without censoring) as the dependent variable. At this point, we are in a position to consider the regression methods for survival data in more detail.

Other texts presenting descriptive as well as other methods for survival data include: Collett (1994), Cox and Oakes (1984), Klein and Moeschberger (1997), Kleinbaum (1996), Le (1997), Lee (1992), Miller (1981), Marubini and Valsecchi (1995) and Parmar and Machin (1995).

EXERCISES

1. Listed below are values of survival time (length of follow-up) for 6 males and 6 females from the WHAS. Right-censored times are denoted by a "+" as a superscript.

Males: 1, 3, 4^+, 10, 12, 18
Females: 1, 3^+, 6, 10, 11, 12^+