Collett: Modelling Survival Data in Medical Research Chapter 2

Table 2.4 *Standard error of $\hat{S}(t)$ and confidence intervals for $S(t)$ for the data from Example 1.1.*

Time interval	$\hat{S}(t)$	se $\{\hat{S}(t)\}$	95% confidence interval
0–	1.0000	0.0000	
10–	0.9444	0.0540	(0.839, 1.000)
19–	0.8815	0.0790	(0.727, 1.000)
30–	0.8137	0.0978	(0.622, 1.000)
36–	0.7459	0.1107	(0.529, 0.963)
59–	0.6526	0.1303	(0.397, 0.908)
75–	0.5594	0.1412	(0.283, 0.836)
93–	0.4662	0.1452	(0.182, 0.751)
97–	0.3729	0.1430	(0.093, 0.653)
107	0.2486	0.1392	(0.000, 0.522)

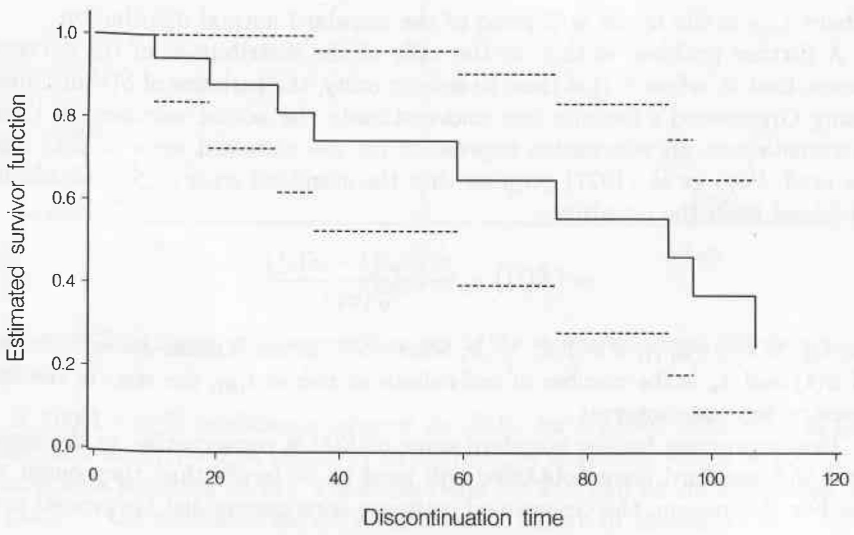

Figure 2.6 *Estimated survivor function and 95% confidence limits for $S(t)$.*

produce confidence bands that are such that there is a given probability, such as 0.95, that the survivor function is contained in the band for all values of t. These bands will tend to be wider than the band formed from the pointwise confidence limits. Details will not be included, but references to these methods are given in the final section of this chapter. Notice also that the width of these intervals is very much greater than the difference between the Kaplan-Meier and Nelson-Aalen estimates of the survivor function, shown in Tables 2.2 and 2.3. Similar calculations lead to confidence limits based on life-table and Nelson-Aalen estimates of the survivor function.

2.3 Estimating the hazard function

A single sample of survival data may also be summarised through the hazard function, which shows the dependence of the instantaneous risk of death on time. There are a number of ways of estimating this function, two of which are described in this section.

2.3.1 *Life-table estimate of the hazard function*

Suppose that the observed survival times have been grouped into a series of m intervals, as in the construction of the life-table estimate of the survivor function. An appropriate estimate of the average hazard of death per unit time over each interval is the observed number of deaths in that interval, divided by the average time survived in that interval. This latter quantity is the average number of persons at risk in the interval, multiplied by the length of the interval. Let the number of deaths in the jth time interval be d_j, $j = 1, 2, \ldots, m$, and suppose that n_j' is the average number of individuals at risk of death in that interval, where n_j' is given by equation (2.2). Assuming that the death rate is constant during the jth interval, the average time survived in that interval is $(n_j' - d_j/2)\tau_j$, where τ_j is the length of the jth time interval. The life-table estimate of the hazard function in the jth time interval is then given by

$$h^*(t) = \frac{d_j}{(n_j' - d_j/2)\tau_j},$$

for $t_j' \leqslant t < t_{j+1}'$, $j = 1, 2, \ldots, m$, so that $h^*(t)$ is a step-function.

The asymptotic standard error of this estimate has been shown by Gehan (1969) to be given by

$$\text{se}\,\{h^*(t)\} = \frac{h^*(t)\surd\{1 - [h^*(t)\tau_j/2]^2\}}{\surd(d_j)},$$

and confidence intervals for the corresponding true hazard over each of the m time intervals can be obtained in the manner described in Section 2.2.3.

Example 2.6 Survival of multiple myeloma patients

The life-table estimate of the survivor function for the data from Example 1.3 on the survival times of 48 multiple myeloma patients was given in Table 2.1. Using the same time intervals as were used in Example 2.2, calculations leading to the life-table estimate of the hazard function are given in Table 2.5.

The estimated hazard function is plotted as a step-function in Figure 2.7. The general pattern is for the hazard to remain roughly constant over the first two years from diagnosis, after which time it declines and then increases gradually. However, some caution is needed in interpreting this estimate, as there are few deaths two years after diagnosis.

Table 2.5 *Life-table estimate of the hazard function for the data from Example 1.3.*

Time period	τ_j	d_j	n'_j	$h^*(t)$
0–	12	16	46.0	0.0351
12–	12	10	26.0	0.0397
24–	12	1	14.0	0.0062
36–	12	3	12.5	0.0227
48–	12	2	8.0	0.0238
60–	36	4	4.5	0.0444

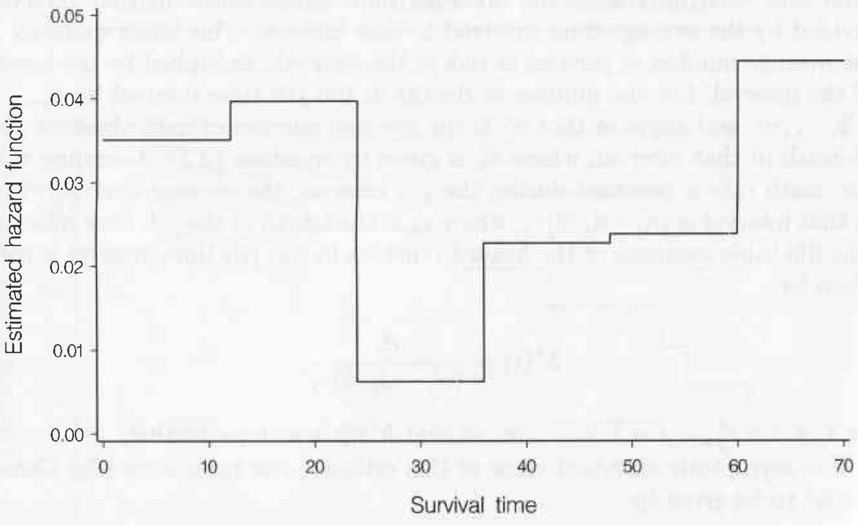

Figure 2.7 *Life-table estimate of the hazard function for the data from Example 1.3.*

2.3.2 Kaplan-Meier type estimate

A natural way of estimating the hazard function for ungrouped survival data is to take the ratio of the number of deaths at a given death time to the number of individuals at risk at that time. If the hazard function is assumed to be constant between successive death times, the hazard per unit time can be found by further dividing by the time interval. Thus if there are d_j deaths at the jth death time, $t_{(j)}$, $j = 1, 2, \ldots, r$, and n_j at risk at time $t_{(j)}$, the hazard function in the interval from $t_{(j)}$ to $t_{(j+1)}$ can be estimated by

$$\hat{h}(t) = \frac{d_j}{n_j \tau_j}, \tag{2.16}$$

for $t_{(j)} \leqslant t < t_{(j+1)}$, where $\tau_j = t_{(j+1)} - t_{(j)}$. Notice that it is not possible to use equation (2.16) to estimate the hazard in the interval that begins at the final death time, since this interval is open-ended.

The estimate in equation (2.16) is referred to as a *Kaplan-Meier type estimate*, because the estimated survivor function derived from it is the Kaplan-Meier estimate. To show this, note that since $\hat{h}(t)$, $t_{(j)} \leqslant t < t_{(j+1)}$, is an estimate of the risk of death per unit time in the jth interval, the probability of death in that interval is $\hat{h}(t)\tau_j$, that is, d_j/n_j. Hence an estimate of the corresponding survival probability in that interval is $1 - (d_j/n_j)$, and the estimated survivor function is as given by equation (2.4).

The approximate standard error of $\hat{h}(t)$ can be found from the variance of d_j, which, following Section 2.2.1, may be assumed to have a binomial distribution with parameters n_j and p_j, where p_j is the probability of death in the interval of length τ. Consequently, var $(d_j) = n_j p_j (1 - p_j)$, and estimating p_j by d_j/n_j gives

$$\text{se}\,\{\hat{h}(t)\} = \hat{h}(t)\sqrt{\left(\frac{n_j - d_j}{n_j d_j}\right)}.$$

However, when d_j is small, confidence intervals constructed using this standard error will be too wide to be of practical use.

Example 2.7 Time to discontinuation of the use of an IUD

Consider again the data on the time to discontinuation of the use of an IUD for 18 women, given in Example 1.1. The Kaplan-Meier estimate of the survivor function for these data was given in Table 2.2, and Table 2.6 gives the corresponding Kaplan-Meier type estimate of the hazard function, computed from equation (2.16). The approximate standard errors of $\hat{h}(t)$ are also given.

Table 2.6 *Kaplan-Meier type estimate of the hazard function for the data from Example 1.1.*

Time interval	τ_j	n_j	d_j	$\hat{h}(t)$	se $\{\hat{h}(t)\}$
0–	10	18	0	0.0000	–
10–	9	18	1	0.0062	0.0060
19–	11	15	1	0.0061	0.0059
30–	6	13	1	0.0128	0.0123
36–	23	12	1	0.0036	0.0035
59–	16	8	1	0.0078	0.0073
75–	18	7	1	0.0079	0.0073
93–	4	6	1	0.0417	0.0380
97–	10	5	1	0.0200	0.0179

Figure 2.8 shows a plot of the estimated hazard function. From this figure, there is some evidence that the longer the IUD is used, the greater is the risk of discontinuation, but the picture is not very clear. The approximate standard

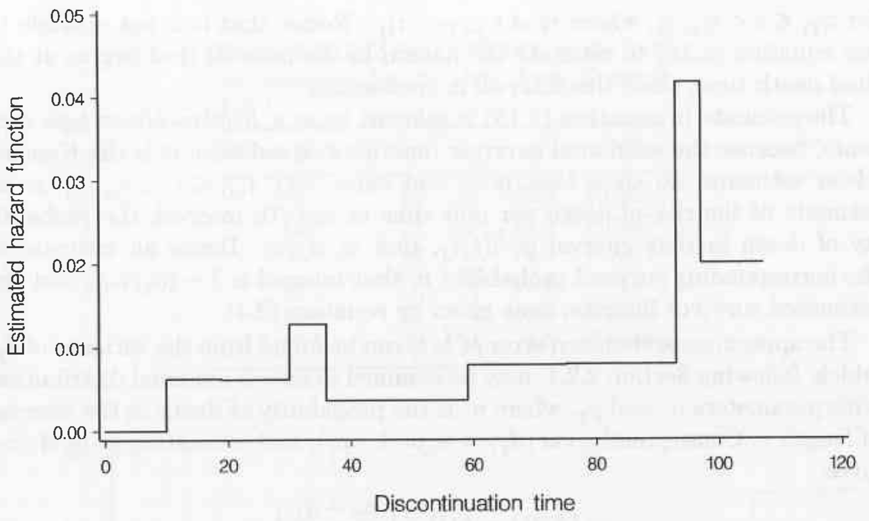

Figure 2.8 *Kaplan-Meier type estimate of the hazard function for the data from Example 1.1.*

errors of the estimated hazard function at different times are of little help in interpreting this plot.

In practice, estimates of the hazard function obtained in this way will often tend to be rather irregular. For this reason, plots of the hazard function may be "smoothed", so that any pattern can be seen more clearly. There are a number of ways of smoothing the hazard function, that lead to a weighted average of values of the estimated hazard $\hat{h}(t)$ at death times in the neighbourhood of t. For example, a *kernel smoothed* estimate of the hazard function, based on the r ordered death times, $t_{(1)}, t_{(2)}, \ldots, t_{(r)}$, with d_j deaths and n_j at risk at time $t_{(j)}$, can be found from

$$h^{\dagger}(t) = b^{-1} \sum_{j=1}^{r} 0.75 \left\{ 1 - \left(\frac{t - t_{(j)}}{b} \right)^2 \right\} \frac{d_j}{n_j},$$

where the value of b needs to be chosen. The function $h^{\dagger}(t)$ is defined for all values of t in the interval from b to $t_{(r)} - b$, where $t_{(r)}$ is the greatest death time. For any value of t in this interval, the death times in the interval $(t-b, t+b)$ will contribute to the weighted average. The parameter b is known as the *bandwidth* and its value controls the shape of the plot; the larger the value of b, the greater the degree of smoothing. There are formulae that lead to "optimal" values of b, but these tend to be rather cumbersome. Fuller details can be found in the references provided in the final section of this chapter. In this book, the use of a modelling approach to the analysis of survival data is advocated, and so model-based estimates of the hazard function will be considered in subsequent chapters.

2.3.3 Estimating the cumulative hazard function

The cumulative hazard function is important in the identification of models for survival data, as will be seen later in Sections 4.4 and 5.2. In addition, since the derivative of the cumulative hazard function is the hazard function itself, the slope of the cumulative hazard function provides information about the shape of the underlying hazard function. In particular, a linear cumulative hazard function over some time interval suggests that the hazard is constant over this interval. Accordingly, methods that can be used to estimate this function will now be described.

The cumulative hazard at time t, $H(t)$, was defined in equation (1.6) to be the integral of the hazard function, but is more conveniently found using equation (1.7). According to this result, $H(t) = -\log S(t)$, and so if $\hat{S}(t)$ is the Kaplan-Meier estimate of the survivor function, $\hat{H}(t) = -\log \hat{S}(t)$ is an appropriate estimate of the cumulative hazard to time t.

Now, using equation (2.4),

$$\hat{H}(t) = -\sum_{j=1}^{k} \log \left(\frac{n_j - d_j}{n_j} \right),$$

for $t_{(k)} \leqslant t < t_{(k+1)}$, $k = 1, 2, \ldots, r$, and $t_{(1)}, t_{(2)}, \ldots, t_{(r)}$ are the r ordered death times, with $t_{(r+1)} = \infty$.

If the Nelson-Aalen estimate of the survivor function is used, the estimated cumulative hazard function, $\tilde{H}(t) = -\log \tilde{S}(t)$, is given by

$$\tilde{H}(t) = \sum_{j=1}^{k} \frac{d_j}{n_j}.$$

This is the cumulative sum of the estimated probabilities of death from the first to the kth time interval, $k = 1, 2, \ldots, r$. This quantity therefore has immediate intuitive appeal as an estimate of the cumulative hazard.

An estimate of the cumulative hazard function also leads to an estimate of the corresponding hazard function, since the differences between adjacent values of the estimated cumulative hazard function provide estimates of the underlying hazard, after dividing by the time interval. In particular, differences in adjacent values of the Nelson-Aalen estimate of the cumulative hazard lead directly to the hazard function estimate in Section 2.3.2.

2.4 Estimating the median and percentiles of survival times

Since the distribution of survival times tends to be positively skew, the median is the preferred summary measure of the location of the distribution. Once the survivor function has been estimated, it is straightforward to obtain an estimate of the *median survival time*. This is the time beyond which 50% of the individuals in the population under study are expected to survive, and is given by that value $t(50)$ which is such that $S\{t(50)\} = 0.5$.

Because the non-parametric estimates of $S(t)$ are step-functions, it will