Modelling Survival Data in Medical Research

SECOND EDITION

David Collett

CHAPMAN & HALL/CRC

A CRC Press Company

Boca Raton London New York Washington, D.C.

be similar over the patients in each of the two treatment groups. However, it would not be wise to rely on this assumption. For example, it could turn out that patients in the placebo group had larger tumours on average than those in the group treated with DES. If patients with large tumours have a poorer prognosis than those with small tumours, the size of the treatment effect would be overestimated unless proper account was taken of the size of the tumour in the analysis. Consequently, it will first be necessary to determine if any of the covariates are related to survival time. If so, the effect of these variables will need to be allowed for when comparing the survival experiences of the patients in the two treatment groups.

1.3 Survivor function and hazard function

In summarising survival data, there are two functions of central interest, namely the *survivor function* and the *hazard function*. These functions are therefore defined in this first chapter.

The actual survival time of an individual, t, can be regarded as the value of a variable, T, which can take any non-negative value. The different values that T can take have a *probability distribution*, and we call T the *random variable* associated with the survival time. Now suppose that the random variable T has a probability distribution with underlying *probability density function* $f(t)$. The *distribution function* of T is then given by

$$F(t) = \mathrm{P}(T < t) = \int_0^t f(u)\,\mathrm{d}u,$$

and represents the probability that the survival time is less than some value t.

The survivor function, $S(t)$, is defined to be the probability that the survival time is greater than or equal to t, and so

$$S(t) = \mathrm{P}(T \geqslant t) = 1 - F(t). \tag{1.1}$$

The survivor function can therefore be used to represent the probability that an individual survives from the time origin to some time beyond t.

The *hazard function* is widely used to express the risk or hazard of death at some time t, and is obtained from the probability that an individual dies at time t, conditional on he or she having survived to that time. For a formal definition of the hazard function, consider the probability that the random variable associated with an individual's survival time, T, lies between t and $t + \delta t$, conditional on T being greater than or equal to t, written $\mathrm{P}(t \leqslant T < t + \delta t \mid T \geqslant t)$. This conditional probability is then expressed as a probability per unit time by dividing by the time interval, δt, to give a *rate*. The hazard function, $h(t)$, is then the limiting value of this quantity, as δt tends to zero, so that

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{\mathrm{P}(t \leqslant T < t + \delta t \mid T \geqslant t)}{\delta t} \right\}. \tag{1.2}$$

The function $h(t)$ is also referred to as the *hazard rate*, the *instantaneous death rate*, the *intensity rate*, or the *force of mortality*.

From equation (1.2), $h(t)\delta t$ is the approximate probability that an individual dies in the interval $(t, t+\delta t)$, conditional on that person having survived to time t. For example, if the survival time is measured in days, $h(t)$ is the approximate probability that an individual, who is alive on day t, dies in the following day. For this reason, the hazard function is often simply interpreted as the risk of death at time t.

From the definition of the hazard function in equation (1.2), we can obtain some useful relationships between the survivor and hazard functions. According to a standard result from probability theory, the probability of an event A, conditional on the occurrence of an event B, is given by $\mathrm{P}(A \mid B) = \mathrm{P}(AB)/\mathrm{P}(B)$, where $\mathrm{P}(AB)$ is the probability of the joint occurrence of A and B. Using this result, the conditional probability in the definition of the hazard function in equation (1.2) is

$$\frac{\mathrm{P}(t \leqslant T < t+\delta t)}{\mathrm{P}(T \geqslant t)},$$

which is equal to

$$\frac{F(t+\delta t) - F(t)}{S(t)},$$

where $F(t)$ is the distribution function of T. Then,

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{F(t+\delta t) - F(t)}{\delta t} \right\} \frac{1}{S(t)}.$$

Now,

$$\lim_{\delta t \to 0} \left\{ \frac{F(t+\delta t) - F(t)}{\delta t} \right\}$$

is the definition of the derivative of $F(t)$ with respect to t, which is $f(t)$, and so

$$h(t) = \frac{f(t)}{S(t)}. \tag{1.3}$$

It then follows that

$$h(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\{\log S(t)\}, \tag{1.4}$$

and so

$$S(t) = \exp\{-H(t)\}, \tag{1.5}$$

where

$$H(t) = \int_0^t h(u)\,\mathrm{d}u. \tag{1.6}$$

The function $H(t)$ features widely in survival analysis, and is called the *integrated* or *cumulative hazard*. From equation (1.5), the cumulative hazard can be obtained from the survivor function, since

$$H(t) = -\log S(t). \tag{1.7}$$

In the analysis of survival data, the survivor function and hazard function are estimated from the observed survival times. Methods of estimation that

do not require the form of the probability density function of T to be specified are described in Chapters 2 and 3, while methods based on the assumption of a particular survival time distribution are presented in Chapters 5 and 6.

1.4 Further reading

An introduction to the techniques used in the analysis of survival data is included in a number of general books on statistics in medical research, such as those of Altman (1991) and Armitage *et al.* (2001). Parmar and Machin (1995) provide a practical guide to the analysis of survival data from clinical trials, using non-technical language.

There a number of textbooks that provide an introduction to the methods of survival analysis, illustrated with practical examples. Lee and Wang (2003) provides a broad coverage of topics with illustrations drawn from biology and medicine. Kleinbaum (1996) provides a self-learning text in two column format, which, like the texts of Harris and Albert (1991) and Miller (1998), emphasises non-parametric methods. Marubini and Valsecchi (1995) describe the analysis of survival data from clinical trials and observational studies. Hosmer and Lemeshow (1999) give a balanced account of survival analysis, with excellent chapters on model development and the interpretation of the parameter estimates in a fitted model. Klein and Moeschberger (1997) include many example data sets and exercises in their comprehensive textbook. Applications of survival analysis in the analysis of epidemiological data are described by Breslow and Day (1987) and Woodward (1999). Introductory texts that describe the application of survival analysis in other areas include those of Elandt-Johnson and Johnson (1999), who focus on actuarial applications, and Crowder *et al.* (1991) who provide a good introduction to the analysis of reliabilty data.

Comprehensive accounts of the subject are given by Kalbfleisch and Prentice (2002), Le (1987) and Lawless (2002). These books have been written for the postgraduate statistician or research worker, and are usually regarded as reference books rather than introductory texts. A concise review of survival analysis is given in the research monograph of Cox and Oakes (1984), and in the chapter devoted to this subject in Hinkley, Reid and Snell (1991).

The book by Hougaard (2000) on multivariate survival data incorporates more advanced topics, after introductory chapters that covers the basic features of survival analysis. Therneau and Grambsch (2000) base their presentation of survival analysis on the counting process approach, leading to a more mathematical development of the material. Smith (2002) describes how a generalisation of least squares allows linear regression models to be used in modelling censored data. Harrell (2001) gives details on many issues that arise in the development of a statistical model not found in other texts, and includes an extensive discussion of two case studies.