theory of counting processes, Fleming and Harrington (1991) and Therneau and Grambsch (2000) show how different types of residual can be used, and give detailed practical examples. Two other types of residual, introduced by Nardi and Schemper (1999), are particularly suitable for the detection of outlying survival times.

Influence diagnostics for the Cox regression model have been considered by many authors, but the major papers are those of Cain and Lange (1984), Reid and Crépeau (1985), Storer and Crowley (1985), Pettitt and Bin Daud (1989) and Weissfeld (1990). Pettitt and Bin Daud (1990) show how time-dependence in the Cox proportional hazards model can be detected by smoothing the Schoenfeld residuals. The LOWESS smoother was introduced by Cleveland (1979), and the algorithm is also presented in Collett (2003).

Some other graphical methods for evaluating survival models, not mentioned in this chapter, have been proposed by Cox (1979) and Arjas (1988). Gray (1990) describes the use of smoothed estimates of cumulative hazard functions in evaluating the fit of a Cox model.

Most of the diagnostic procedures presented in this chapter rely on an informal evaluation of tabular or graphical presentations of particular statistics. In addition to these procedures, a variety of significance tests have been proposed that can be used to asses the goodness of fit of the model. Examples include the methods of Schoenfeld (1980), Andersen (1982), Nagelkerke *et al.* (1984), Ciampi and Etezadi-Amoli (1985), Moreau *et al.* (1985), Gill and Schumacher (1987), O'Quigley and Pessione (1989), Quantin *et al.* (1996), Grønnesby and Borgan (1996), and Verweij *et al.* (1998). Reviews of some of these goodness of fit tests for the Cox regression model are included in Lin and Wei (1991) and Quantin *et al.* (1996). Many of these test involve statistics that are quite complicated, and the procedures are not widely in computer software for survival analysis. A more simple procedure for evaluating the overall fit of a model has been proposed by May and Hosmer (1998).

Modelling Survival Data in Medical Research
David Collett. 2nd Ed. 2003.

CHAPTER 5

Parametric proportional hazards models

When the Cox regression model is used in the analysis of survival data, there is no need to assume a particular form of probability distribution for the survival times. As a result, the hazard function is not restricted to a specific functional form, and the model has flexibility and widespread applicability. On the other hand, if the assumption of a particular probability distribution for the data is valid, inferences based on such an assumption will be more precise. In particular, estimates of quantities such as relative hazards and median survival times will tend to have smaller standard errors than they would in the absence of a distributional assumption. Models in which a specific probability distribution is assumed for the survival times are known as *parametric models*, and parametric versions of the proportional hazards model, described in Chapter 3, are the subject of this chapter.

A probability distribution that plays a central role in the analysis of survival data is the Weibull distribution, introduced by W. Weibull in 1951 in the context of industrial reliability testing. Indeed, this distribution is as central to the parametric analysis of survival data as the normal distribution is in linear modelling. Proportional hazards models based on the Weibull distribution are therefore considered in some detail.

5.1 Models for the hazard function

Once a distributional model for survival times has been specified in terms of a probability density function, the corresponding survivor and hazard functions can be obtained from the relations

$$S(t) = 1 - \int_0^t f(u)\,\mathrm{d}u, \tag{5.1}$$

and

$$h(t) = \frac{f(t)}{S(t)} = -\frac{\mathrm{d}}{\mathrm{d}t}\{\log S(t)\}, \tag{5.2}$$

where $f(t)$ is the probability density function of the survival times. These relationships were derived in Section 1.3. An alternative approach is to specify a functional form for the hazard function, from which the survivor function and probability density functions can be determined from the equations

$$S(t) = \exp\{-H(t)\}, \tag{5.3}$$

and

$$f(t) = h(t)S(t) = -\frac{\mathrm{d}S(t)}{\mathrm{d}t}, \tag{5.4}$$

where

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

is the integrated hazard function.

5.1.1 *The exponential distribution*

The simplest model for the hazard function is to assume that it is constant over time. The hazard of death at any time after the time origin of the study is then the same, irrespective of the time elapsed. Under this model, the hazard function may be written as

$$h(t) = \lambda,$$

for $0 \leqslant t < \infty$. The parameter λ is a positive constant that would be estimated by fitting the model to the observed data. From equation (5.3), the corresponding survivor function is

$$\begin{aligned} S(t) &= \exp\left\{-\int_0^t \lambda\,\mathrm{d}u\right\}, \\ &= e^{-\lambda t}, \end{aligned} \tag{5.5}$$

and so the implied probability density function of the survival times is

$$f(t) = \lambda e^{-\lambda t}, \tag{5.6}$$

for $0 \leqslant t < \infty$. This is the probability density function of a random variable T that has an *exponential distribution* with a mean of λ^{-1}. It is sometimes convenient to write $\mu = \lambda^{-1}$, so that the hazard function is μ^{-1}, and the survival time distribution has a mean of μ. However, the former specification of the hazard function will generally be used in this book.

The median of the exponential distribution, $t(50)$, is such that $S\{t(50)\} = 0.5$, that is,

$$\exp\{-\lambda t(50)\} = 0.5,$$

so that

$$t(50) = \frac{1}{\lambda}\log 2.$$

More generally, the pth percentile of the survival time distribution is the value $t(p)$ such that $S\{t(p)\} = 1 - (p/100)$, and using equation (5.5), this is

$$t(p) = \frac{1}{\lambda}\log\left(\frac{100}{100-p}\right).$$

A plot of the hazard function for three values of λ, namely 1.0, 0.1 and 0.01, is given in Figure 5.1, and the corresponding probability density functions are shown in Figure 5.2. For these values of λ, the means of the corresponding exponential distributions are 1, 10 and 100, and the median survival times are 0.69, 6.93 and 69.31, respectively.

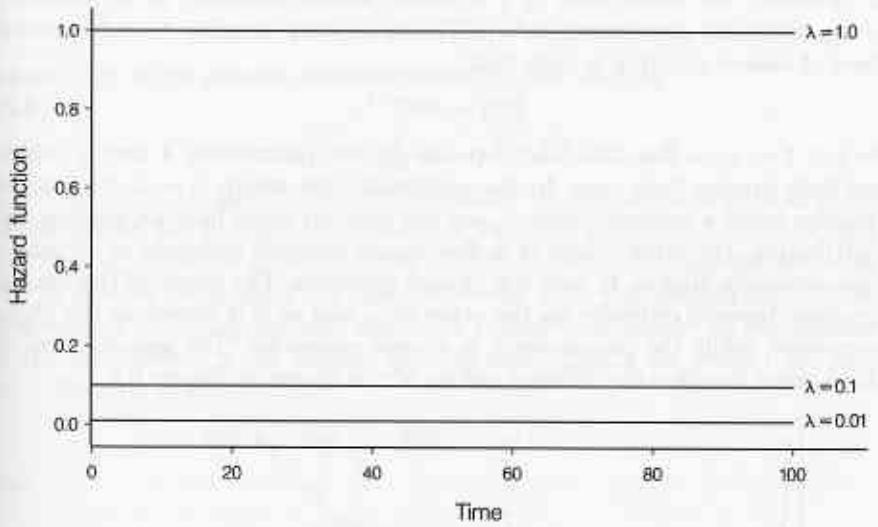

Figure 5.1 *Hazard functions for exponential distributions with $\lambda = 1.0$, 0.1 and 0.01.*

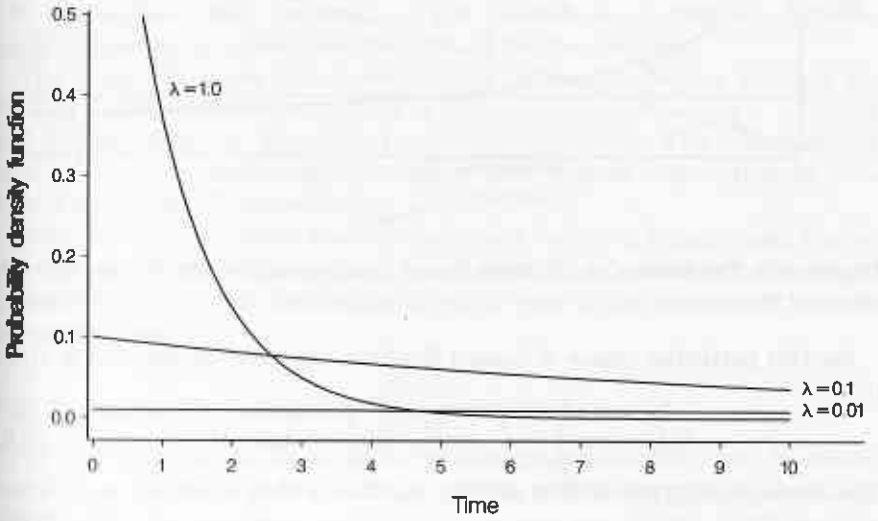

Figure 5.2 *Probability density functions for exponential distributions with $\lambda = 1.0$, 0.1 and 0.01.*

5.1.2 The Weibull distribution

In practice, the assumption of a constant hazard function, or equivalently of exponentially distributed survival times, is rarely tenable. A more general form of hazard function is such that

$$h(t) = \lambda\gamma t^{\gamma-1}, \tag{5.7}$$

for $0 \leqslant t < \infty$, a function that depends on two parameters λ and γ, which are both greater than zero. In the particular case where $\gamma = 1$, the hazard function takes a constant value λ, and the survival times have an exponential distribution. For other values of γ, the hazard function increases or decreases monotonically, that is, it does not change direction. The shape of the hazard function depends critically on the value of γ, and so γ is known as the *shape parameter*, while the parameter λ is a *scale parameter*. The general form of this hazard function for different values of γ is shown in Figure 5.3.

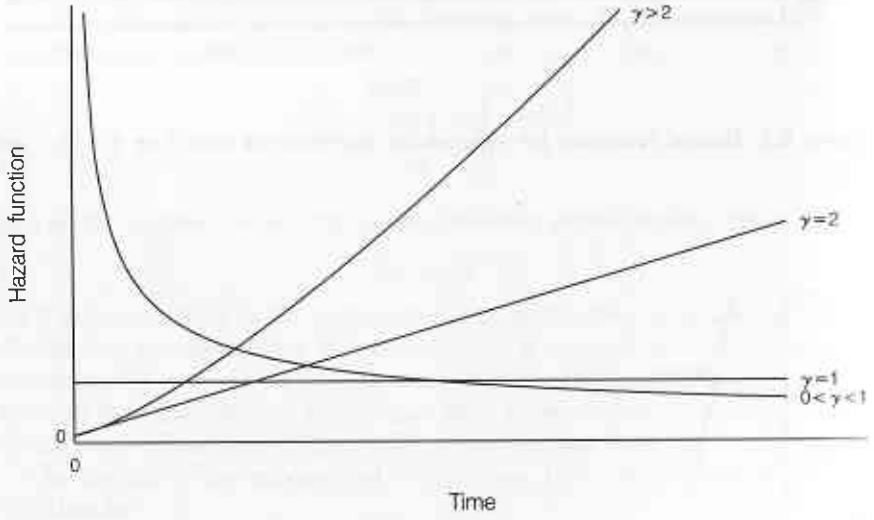

Figure 5.3 *The form of the Weibull hazard function, $h(t) = \lambda\gamma t^{\gamma-1}$, for different values of γ.*

For this particular choice of hazard function, the survivor function is given by

$$S(t) = \exp\left\{-\int_0^t \lambda\gamma u^{\gamma-1}\,\mathrm{d}u\right\} = \exp(-\lambda t^{\gamma}). \tag{5.8}$$

The corresponding probability density function is then

$$f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^{\gamma}),$$

for $0 \leqslant t < \infty$, which is the density of a random variable that has a *Weibull distribution* with scale parameter λ and shape parameter γ. This distribution will be denoted $W(\lambda, \gamma)$. The right-hand tail of this distribution is longer than the left-hand one, and so the distribution is positively skewed.

The mean, or expected value, of a random variable T that has a $W(\lambda, \gamma)$ distribution can be shown to be given by

$$\mathrm{E}(T) = \lambda^{-1/\gamma}\Gamma(\gamma^{-1} + 1),$$

where $\Gamma(x)$ is the gamma function defined by the integral

$$\Gamma(x) = \int_0^{\infty} u^{x-1}e^{-u}\,\mathrm{d}u.$$

The value of this integral is $(x-1)!$, and so for integer values of x it can easily be calculated. For non-integer values of x, tables of the gamma function, such as those in Abramowitz and Stegun (1972), or suitable computer software, will be needed to compute the mean. However, since the Weibull distribution is skewed, a more appropriate, and more tractable, summary of the location of the distribution is the median survival time. This is the value $t(50)$ such that $S\{t(50)\} = 0.5$, so that

$$\exp\{-\lambda[t(50)]^{\gamma}\} = 0.5,$$

and

$$t(50) = \left\{\frac{1}{\lambda}\log 2\right\}^{1/\gamma}.$$

More generally, the pth percentile of the Weibull distribution, $t(p)$, is such that

$$t(p) = \left\{\frac{1}{\lambda}\log\left(\frac{100}{100-p}\right)\right\}^{1/\gamma}. \tag{5.9}$$

The median and other percentiles of the Weibull distribution are therefore much simpler to compute than the mean of the distribution.

The hazard function and corresponding probability density function for Weibull distributions with a median of 20, and shape parameters $\gamma = 0.5$, 1.5 and 3.0, are shown in Figures 5.4 and 5.5, respectively. The corresponding value of the scale parameter, λ, for these three Weibull distributions is 0.15, 0.0078 and 0.000087, respectively.

Since the Weibull hazard function can take a variety of forms, depending on the value of the shape parameter, γ, and appropriate summary statistics can be easily obtained, this distribution is widely used in the parametric analysis of survival data.

5.2 Assessing the suitability of a parametric model

Prior to fitting a model based on an assumed parametric form for the hazard function, a preliminary study of the validity of this assumption should be carried out. One approach would be to estimate the hazard function using the methods outlined in Section 2.3. If the hazard function were reasonably constant over time, this would indicate that the exponential distribution might be a suitable model for the data. On the other hand, if the hazard function increased or decreased monotonically with increasing survival time, a model based on the Weibull distribution would be indicated.

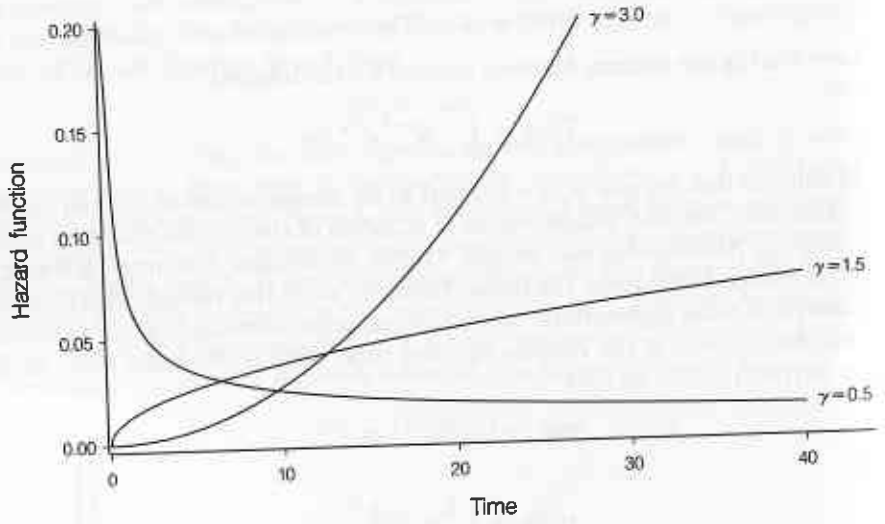

Figure 5.4 *Hazard functions for a Weibull distribution with a median of 20 and $\gamma = 0.5$, 1.5 and 3.0.*

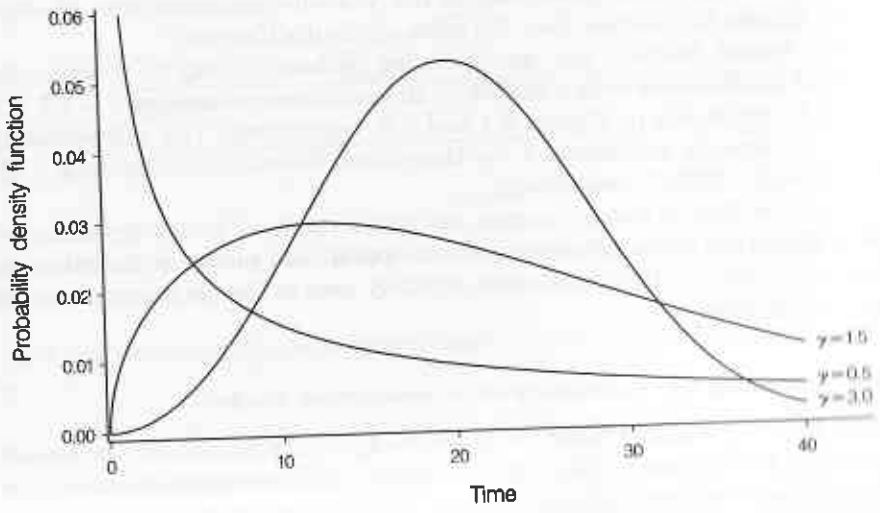

Figure 5.5 *Probability density functions for a Weibull distribution with a median of 20 and $\gamma = 0.5$, 1.5 and 3.0.*

A more informative way of assessing whether a particular distribution for the survival times is plausible is to compare the survivor function for the data with that of a chosen model. This is greatly helped by transforming the survivor function to produce a plot that should give a straight line if the assumed model is appropriate.

Suppose that a single sample of survival data is available, and that a Weibull distribution for the survival times is contemplated. Since the survivor function for a Weibull distribution, with scale parameter λ and shape parameter γ, is given by

$$S(t) = \exp\{-\lambda t^{\gamma}\},$$

taking the logarithm of $S(t)$, multiplying by -1, and taking logarithms a second time, gives

$$\log\{-\log S(t)\} = \log\lambda + \gamma \log t. \tag{5.10}$$

We now substitute the Kaplan-Meier estimate of the survivor function, $\hat{S}(t)$, for $S(t)$ in equation (5.10). If the Weibull assumption is tenable, $\hat{S}(t)$ will be "close" to $S(t)$, and a plot of $\log\{-\log \hat{S}(t)\}$ against $\log t$ would then give an approximately straight line. From equation (1.7), the cumulative hazard function, $H(t)$, is $-\log S(t)$ and so $\log\{-\log S(t)\}$ is the log-cumulative hazard. A plot of the values of $\log\{-\log \hat{S}(t)\}$ against $\log t$ is a log-cumulative hazard plot, introduced in Section 4.4.1 of Chapter 4.

If the log-cumulative hazard plot gives a straight line, the plot can be used to provide a rough estimate of the two parameters of the Weibull distribution. Specifically, from equation (5.10), the intercept and slope of the straight line will be $\log\lambda$ and γ, respectively. Thus, the slope of the line in a log-cumulative hazard plot gives an estimate of the shape parameter, and the exponent of the intercept provides an estimate of the scale parameter. Note that if the slope of the log-cumulative hazard plot is close to unity, the survival times could have an exponential distribution.

Example 5.1 Time to discontinuation of the use of an IUD

In Example 2.3, the Kaplan-Meier estimate of the survivor function, $\hat{S}(t)$, for the data on the time to discontinuation of an IUD, was obtained. A log-cumulative hazard plot for these data, that is, a plot of $\log\{-\log \hat{S}(t)\}$ against $\log t$, is shown in Figure 5.6.

The plot indicates that there is a straight line relationship between the log-cumulative hazard and $\log t$, confirming that the Weibull distribution is an appropriate model for the discontinuation times. From the graph, the intercept of the line is approximately -6.0 and the slope is approximately 1.25. Approximate estimates of the parameters of the Weibull distribution are therefore $\lambda^* = \exp(-6.0) = 0.002$ and $\gamma^* = 1.25$. The estimated value of γ, the shape parameter of the Weibull distribution, is quite close to unity, suggesting that the discontinuation times might be adequately modelled by an exponential distribution.

These informal estimates of λ and γ can be used to estimate the parameters

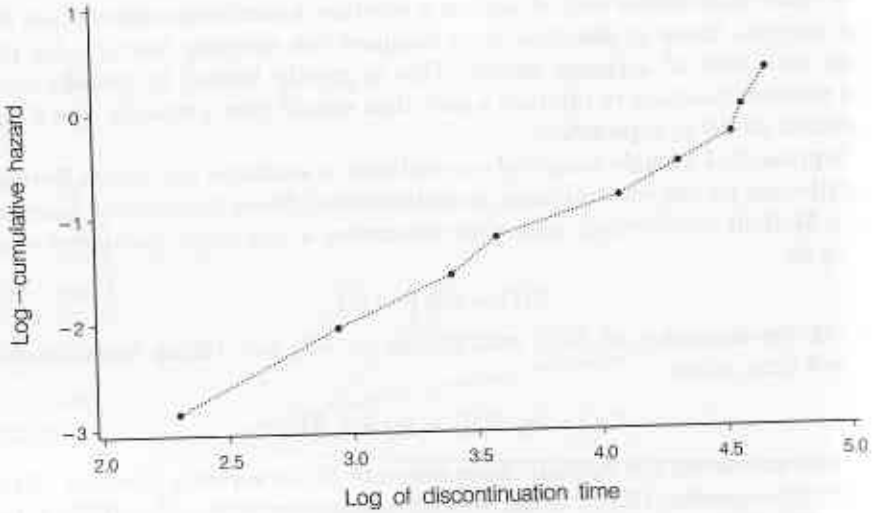

Figure 5.6 *Log-cumulative hazard plot for the data from Example 1.1.*

of the distribution, and hence functions of these estimates, such as the median of the survival time distribution. However, this graphical approach does not lead to a measure of the precision with which the quantities have been estimated. In view of this limitation, a more formal way of fitting parametric models to survival data is developed in the next section.

5.3 Fitting a parametric model to a single sample

Parametric models can be fitted to an observed set of survival data using the method of maximum likelihood, outlined in Section 3.3. Consider first the situation where actual survival times have been observed for n individuals, so that there are no censored observations. If the probability density function of the random variable associated with survival time is $f(t)$, the likelihood of the n observations $t_1, t_2, \ldots, t_n$ is simply the product

$$\prod_{i=1}^{n} f(t_i).$$

This likelihood will be a function of the unknown parameters in the probability density function, and the maximum likelihood estimates of these parameters are those values for which the likelihood function is a maximum. In practice, it is generally more convenient to work with the logarithm of the likelihood function. Those values of the unknown parameters in the density function that maximise the log-likelihood are of course the same values that maximise the likelihood function itself.

We now consider the more usual situation where the survival data include one or more censored survival times. Specifically, suppose that r of the n individuals die at times $t_1, t_2, \ldots, t_r$, and that the survival times of the remaining $n - r$ individuals, $t_1^*, t_2^*, \ldots, t_{n-r}^*$, are right-censored. The r death times contribute a term of the form

$$\prod_{j=1}^{r} f(t_j)$$

to the overall likelihood function. Naturally, we cannot ignore information about the survival experience of the $n - r$ individuals for whom a censored survival time has been recorded. If a survival time is censored at time t^*, say, we know that the lifetime of the individual is at least t^*, and the probability of this event is $\mathrm{P}(T \geqslant t^*)$, which is $S(t^*)$. Thus each censored observation contributes a term of this form to the likelihood of the n observations. The total likelihood function is therefore

$$\prod_{j=1}^{r} f(t_j) \prod_{l=1}^{n-r} S(t_l^*), \tag{5.11}$$

in which the first product is taken over the r death times and the second over the $n - r$ censored survival times.

More compactly, suppose that the data are regarded as n pairs of observations, where the pair for the ith individual is (t_i, δ_i), $i = 1, 2, \ldots, n$. In this notation, δ_i is an indicator variable that takes the value zero when the survival time t_i is censored and unity when t_i is an uncensored survival time. The likelihood function can then be written as

$$\prod_{i=1}^{n} \{f(t_i)\}^{\delta_i} \{S(t_i)\}^{1-\delta_i}. \tag{5.12}$$

This function, which is equivalent to that in expression (5.11), can then be maximised with respect to the unknown parameters in the density and survivor functions. A more careful derivation of this likelihood function is given in Appendix B, which shows the relevance of the assumption of non-informative censoring, mentioned in Section 1.1 of Chapter 1.

An alternative expression for the likelihood function can be obtained by writing expression (5.12) in the form

$$\prod_{i=1}^{n} \left\{\frac{f(t_i)}{S(t_i)}\right\}^{\delta_i} S(t_i),$$

so that, from equation (1.3) of Chapter 1, this becomes

$$\prod_{i=1}^{n} \{h(t_i)\}^{\delta_i} S(t_i). \tag{5.13}$$

This version of the likelihood function is particularly useful when the probability density function has a complicated form, as it often does. Estimates of the unknown parameters in this likelihood function are then found by maximising the logarithm of the likelihood function.

We now consider fitting exponential and Weibull distributions to a single sample of survival data.

5.3.1* *Fitting the exponential distribution*

Suppose that the survival times of n individuals, $t_1, t_2, \ldots, t_n$, are assumed to have an exponential distribution with mean λ^{-1}. Further suppose that the data give the actual death times of r individuals, and that the remaining $n-r$ survival times are right-censored.

For the exponential distribution,

$$f(t) = \lambda e^{-\lambda t}, \qquad S(t) = e^{-\lambda t},$$

and on substituting into expression (5.12), the likelihood function for the n observations is given by

$$L(\lambda) = \prod_{i=1}^{n} \left(\lambda e^{-\lambda t_i}\right)^{\delta_i} \left(e^{-\lambda t_i}\right)^{1-\delta_i},$$

where δ_i is zero if the survival time of the ith individual is censored and unity otherwise. After some simplification,

$$L(\lambda) = \prod_{i=1}^{n} \lambda^{\delta_i} e^{-\lambda t_i},$$

and the corresponding log-likelihood function is

$$\log L(\lambda) = \sum_{i=1}^{n} \delta_i \log \lambda - \lambda \sum_{i=1}^{n} t_i.$$

Since the data contain r deaths, $\sum_{i=1}^{n} \delta_i = r$ and the log-likelihood function becomes

$$\log L(\lambda) = r \log \lambda - \lambda \sum_{i=1}^{n} t_i.$$

We now need to identify the value $\hat{\lambda}$, for which the log-likelihood function is a maximum. Differentiation with respect to λ gives

$$\frac{\mathrm{d} \log L(\lambda)}{\mathrm{d}\lambda} = \frac{r}{\lambda} - \sum_{i=1}^{n} t_i,$$

and equating the derivative to zero and evaluating it at $\hat{\lambda}$ gives

$$\hat{\lambda} = r / \sum_{i=1}^{n} t_i \tag{5.14}$$

for the maximum likelihood estimator of λ.

The mean of an exponential distribution is $\mu = \lambda^{-1}$, and so the maximum

likelihood estimator of μ is

$$\hat{\mu} = \hat{\lambda}^{-1} = \frac{1}{r} \sum_{i=1}^{n} t_i.$$

This estimator of μ is the total time survived by the n individuals in the data set divided by the number of deaths observed. The estimator therefore has intuitive appeal as an estimate of the mean lifetime from censored survival data.

The standard error of either $\hat{\lambda}$ or $\hat{\mu}$ can be obtained from the second derivative of the log-likelihood function, using a result from the theory of maximum likelihood estimation given in Appendix A. Differentiating $\log L(\lambda)$ a second time gives

$$\frac{\mathrm{d}^2 \log L(\lambda)}{\mathrm{d}\lambda^2} = -\frac{r}{\lambda^2},$$

and so the asymptotic variance of $\hat{\lambda}$ is

$$\mathrm{var}\,(\hat{\lambda}) = \left\{ -\mathrm{E}\left(\frac{\mathrm{d}^2 \log L(\lambda)}{\mathrm{d}\lambda^2} \right) \right\}^{-1} = \frac{\lambda^2}{r}.$$

Consequently, the standard error of $\hat{\lambda}$ is given by

$$\mathrm{se}\,(\hat{\lambda}) = \hat{\lambda}/\sqrt{r}. \tag{5.15}$$

This result could be used to obtain a confidence interval for the mean survival time. In particular, the limits of a $100(1-\alpha)\%$ confidence interval for λ are $\hat{\lambda} \pm z_{\alpha/2}\,\mathrm{se}\,(\hat{\lambda})$, where $z_{\alpha/2}$ is the upper $\alpha/2$-point of the standard normal distribution.

In presenting the results of a survival analysis, the estimated survivor and hazard functions, and the median and other percentiles of the distribution of survival times, are useful. Once an estimate of λ has been found, all these functions can be estimated using the results given in Section 5.1.1. In particular, under the assumed exponential distribution, the estimated hazard function is $\hat{h}(t) = \hat{\lambda}$ and the estimated survivor function is $\hat{S}(t) = \exp(-\hat{\lambda}t)$. In addition, the estimated pth percentile is given by

$$\hat{t}(p) = \frac{1}{\hat{\lambda}} \log\left(\frac{100}{100-p} \right), \tag{5.16}$$

and the estimated median survival time is

$$\hat{t}(50) = \hat{\lambda}^{-1} \log 2. \tag{5.17}$$

The standard error of an estimate of the pth percentile of the distribution of survival times can be found using the result for the approximate variance of a function of a random variable given in equation (2.9) of Chapter 2. According to this result, an approximation to the variance of a function $g(\hat{\lambda})$ of $\hat{\lambda}$ is such that

$$\mathrm{var}\,\{g(\hat{\lambda})\} \approx \left\{ \frac{\mathrm{d}g(\hat{\lambda})}{\mathrm{d}\hat{\lambda}} \right\}^2 \mathrm{var}\,(\hat{\lambda}). \tag{5.18}$$

Using this result, the approximate variance of the estimated pth percentile is given by

$$\text{var}\{\hat{t}(p)\} \approx \left\{-\frac{1}{\hat{\lambda}^2}\log\left(\frac{100}{100-p}\right)\right\}^2 \text{var}(\hat{\lambda}).$$

On simplifying this and taking the square root, we get

$$\text{se}\{\hat{t}(p)\} = \frac{1}{\hat{\lambda}^2}\log\left(\frac{100}{100-p}\right)\text{se}(\hat{\lambda}),$$

and on further substituting for $\text{se}(\hat{\lambda})$ from equation (5.15) and $\hat{t}(p)$ from equation (5.16), we find

$$\text{se}\{\hat{t}(p)\} = \hat{t}(p)/\sqrt{r}. \tag{5.19}$$

In particular, the standard error of the estimated median survival time is

$$\text{se}\{\hat{t}(50)\} = \hat{t}(50)/\sqrt{r}. \tag{5.20}$$

Confidence intervals for a true percentile are best obtained from exponentiating the confidence limits for the logarithm of the percentile. This procedure ensures that confidence limits for the percentile will be non-negative. Again making use of the result in equation (5.18), the standard error of $\log\hat{t}(p)$ is given by

$$\text{se}\{\log\hat{t}(p)\} = \hat{t}(p)^{-1}\,\text{se}\{\hat{t}(p)\},$$

and after substituting for $\text{se}\{\hat{t}(p)\}$ from equation (5.19), this standard error becomes

$$\text{se}\{\log\hat{t}(p)\} = 1/\sqrt{r}.$$

Using this result, $100(1-\alpha)\%$ confidence limits for the $100p$th percentile are $\exp\{\log\hat{t}(p) \pm z_{\alpha/2}/\sqrt{r}\}$, that is, $\hat{t}(p)\exp\{\pm z_{\alpha/2}/\sqrt{r}\}$, where $z_{\alpha/2}$ is the upper $\alpha/2$-point of the standard normal distribution.

Example 5.2 Time to discontinuation of the use of an IUD

In this example, the data of Example 1.1 on the times to discontinuation of an IUD for 18 women are analysed under the assumption of a constant hazard of discontinuation. An exponential distribution is therefore fitted to the discontinuation times. For these data, the total of the observed and right-censored discontinuation times is 1046 days, and the number of uncensored times is 9. Therefore, using equation (5.14), $\hat{\lambda} = 9/1046 = 0.0086$, and the standard error of $\hat{\lambda}$ from equation (5.15) is $\text{se}(\hat{\lambda}) = 0.0086/\sqrt{9} = 0.0029$. The estimated hazard function is therefore $\hat{h}(t) = 0.0086$, $t > 0$, and the estimated survivor function is $\hat{S}(t) = \exp(-0.0086\,t)$. The estimated hazard and survivor functions are shown in Figures 5.7 and 5.8, respectively.

Estimates of the median and other percentiles of the distribution of discontinuation times can be found from Figure 5.8, but more accurate estimates are obtained from equation (5.16). In particular, using equation (5.17), the median discontinuation time is 81 days, and an estimate of the 90th percentile of the distribution of discontinuation times is, from equation (5.16), $\hat{t}(90) = \log 10/0.0086 = 267.61$. This means that on the assumption that the

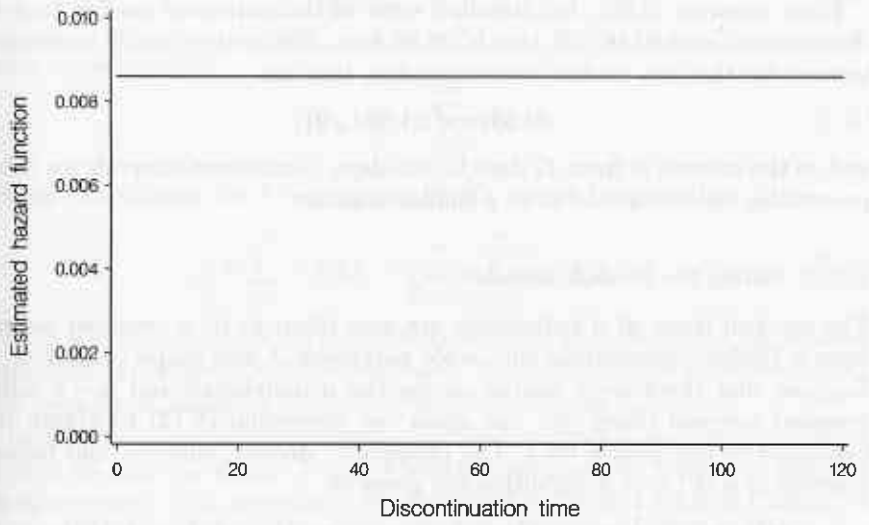

Figure 5.7 *Estimated hazard function on fitting the exponential distribution.*

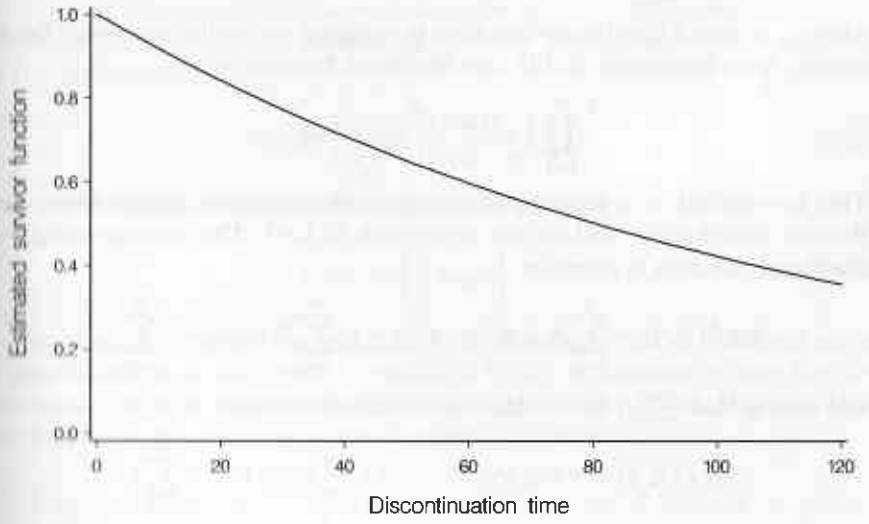

Figure 5.8 *Estimated survivor function on fitting the exponential distribution.*

risk of discontinuing the use of an IUD is independent of time, 90% of women will have a discontinuation time of less than 268 days.

From equation (5.20), the standard error of the estimated median time to discontinuation is $80.56/\sqrt{9}$, that is, 26.85 days. The limits of a 95% confidence interval for the true median discontinuation time are

$$80.56 \exp\{\pm 1.96/\sqrt{9}\},$$

and so the interval is from 42 days to 155 days. Confidence intervals for other percentiles can be calculated in a similar manner.

5.3.2 Fitting the Weibull distribution*

The survival times of n individuals are now taken to be a censored sample from a Weibull distribution with scale parameter λ and shape parameter γ. Suppose that there are r deaths among the n individuals and $n-r$ right-censored survival times. We can again use expression (5.12) to obtain the likelihood of the sample data. The probability density, survivor and hazard function of a $W(\lambda, \gamma)$ distribution are given by

$$f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^{\gamma}), \quad S(t) = \exp(-\lambda t^{\gamma}), \quad h(t) = \lambda\gamma t^{\gamma-1},$$

and so, from expression (5.12), the likelihood of the n survival times is

$$\prod_{i=1}^{n}\left\{\lambda\gamma t_i^{\gamma-1}\exp(-\lambda t_i^{\gamma})\right\}^{\delta_i}\{\exp(-\lambda t_i^{\gamma})\}^{1-\delta_i},$$

where δ_i is zero if the ith survival time is censored and unity otherwise. Equivalently, from expression (5.13), the likelihood function is

$$\prod_{i=1}^{n}\left\{\lambda\gamma t_i^{\gamma-1}\right\}^{\delta_i}\exp(-\lambda t_i^{\gamma}).$$

This is regarded as a function of λ and γ, the unknown parameters in the Weibull distribution, and so can be written $L(\lambda, \gamma)$. The corresponding log-likelihood function is given by

$$\log L(\lambda,\gamma) = \sum_{i=1}^{n}\delta_i\log(\lambda\gamma) + (\gamma-1)\sum_{i=1}^{n}\delta_i\log t_i - \lambda\sum_{i=1}^{n}t_i^{\gamma},$$

and noting that $\sum_{i=1}^{n}\delta_i = r$, the log-likelihood becomes

$$\log L(\lambda,\gamma) = r\log(\lambda\gamma) + (\gamma-1)\sum_{i=1}^{n}\delta_i\log t_i - \lambda\sum_{i=1}^{n}t_i^{\gamma}.$$

The maximum likelihood estimates of λ and γ are found by differentiating this function with respect to λ and γ, equating the derivatives to zero, and evaluating them at $\hat{\lambda}$ and $\hat{\gamma}$. The resulting equations are

$$\frac{r}{\hat{\lambda}} - \sum_{i=1}^{n}t_i^{\hat{\gamma}} = 0, \tag{5.21}$$

and

$$\frac{r}{\hat{\gamma}} + \sum_{i=1}^{n}\delta_i\log t_i - \hat{\lambda}\sum_{i=1}^{n}t_i^{\hat{\gamma}}\log t_i = 0. \tag{5.22}$$

From equation (5.21),

$$\hat{\lambda} = r/\sum_{i=1}^{n}t_i^{\hat{\gamma}}, \tag{5.23}$$

and on substituting for $\hat{\lambda}$ in equation (5.22), we get the equation

$$\frac{r}{\hat{\gamma}} + \sum_{i=1}^{n}\delta_i\log t_i - \frac{r}{\sum_i t_i^{\hat{\gamma}}}\sum_{i=1}^{n}t_i^{\hat{\gamma}}\log t_i = 0. \tag{5.24}$$

This is a non-linear equation in $\hat{\gamma}$, which can only be solved using an iterative numerical procedure. Once the estimate, $\hat{\gamma}$, which satisfies equation (5.24), has been found, equation (5.23) can be used to obtain $\hat{\lambda}$.

In practice, a numerical procedure, such as the Newton-Raphson algorithm, is used to find the values $\hat{\lambda}$ and $\hat{\gamma}$ which maximise the likelihood function simultaneously. This procedure was described in Section 3.3.3 of Chapter 3, in connection with fitting the Cox regression model. In that section it was noted that an important by-product of the Newton-Raphson procedure is an approximation to the variance-covariance matrix of the parameter estimates, from which their standard errors can be obtained.

Once estimates of the parameters λ and γ have been found from fitting the Weibull distribution to the observed data, percentiles of the survival time distribution can be estimated using equation (5.9). The estimated pth percentile of the distribution is

$$\hat{t}(p) = \left\{\frac{1}{\hat{\lambda}}\log\left(\frac{100}{100-p}\right)\right\}^{1/\hat{\gamma}}, \tag{5.25}$$

and so the estimated median survival time is given by

$$\hat{t}(50) = \left\{\frac{1}{\hat{\lambda}}\log 2\right\}^{1/\hat{\gamma}}. \tag{5.26}$$

The standard error of the estimated pth percentile can be obtained using a generalisation of the result in equation (5.18) to the case where the approximate variance of a function of two estimates is required. Details of the derivation are given in Appendix C, where it is shown that

$$\text{se}\left\{\hat{t}(p)\right\} = \frac{\hat{t}(p)}{\hat{\lambda}\hat{\gamma}^2}\left\{\hat{\gamma}^2\,\text{var}\,(\hat{\lambda}) + \hat{\lambda}^2\left(c_p - \log\hat{\lambda}\right)^2\text{var}\,(\hat{\gamma}) + 2\hat{\lambda}\hat{\gamma}\left(c_p - \log\hat{\lambda}\right)\text{cov}\,(\hat{\lambda},\hat{\gamma})\right\}^{\frac{1}{2}}, \tag{5.27}$$

where

$$c_p = \log\log\left(\frac{100}{100-p}\right).$$

The variances of $\hat{\lambda}$ and $\hat{\gamma}$, and their covariance, are found from the variance-covariance matrix of the estimates.

As before, a confidence interval for the true value of the pth percentile, $t(p)$, is best obtained from the corresponding interval for $\log t(p)$. The standard error of $\log \hat{t}(p)$ is

$$\text{se}\{\log \hat{t}(p)\} = \frac{1}{\hat{t}(p)} \text{se}\{\hat{t}(p)\}, \tag{5.28}$$

and $100(1-\alpha)\%$ confidence limits for $\log t(p)$ are

$$\log \hat{t}(p) \pm z_{\alpha/2} \text{se}\{\log \hat{t}(p)\}.$$

Corresponding interval estimates for $t(p)$ are found by exponentiating these limits. For example, the limits of a $100(1-\alpha)\%$ confidence interval for the median survival time, $t(50)$, are $\hat{t}(50) \exp\left[\pm z_{\alpha/2} \text{se}\{\log \hat{t}(50)\}\right]$.

There is a substantial amount of arithmetic involved in these calculations, and care must be taken to ensure that significant figures are not lost during the course of the calculation. For this reason, it is better to perform the calculations using a suitable computer program.

Example 5.3 Time to discontinuation of the use of an IUD

In Example 5.1, it was found that an exponential distribution provides a satisfactory model for the data on the discontinuation times of 18 IUD users. For comparison, a Weibull distribution will be fitted to the same data set. The distribution can be fitted using computer software, and from the resulting output, the estimated scale parameter of the distribution is found to be $\hat{\lambda} = 0.000454$, while the estimated shape parameter is $\hat{\gamma} = 1.676$. The standard errors of these estimates are given by $\text{se}(\hat{\lambda}) = 0.000965$ and $\text{se}(\hat{\gamma}) = 0.460$, respectively. Note that approximate confidence limits for the shape parameter, γ, found using $\hat{\gamma} \pm 1.96\,\text{se}(\hat{\gamma})$, include unity, suggesting that the exponential distribution would provide a satisfactory model for the discontinuation times.

The estimated hazard and survivor functions are obtained by substituting these estimates into equations (5.7) and (5.8), whence

$$\hat{h}(t) = \hat{\lambda}\hat{\gamma}t^{\hat{\gamma}-1},$$

and

$$\hat{S}(t) = \exp\left(-\hat{\lambda}t^{\hat{\gamma}}\right).$$

These two functions are shown in Figures 5.9 and 5.10.

Although percentiles of the discontinuation time can be read from the estimated survivor function in Figure 5.10, they are better estimated using equation (5.25). Hence, under the Weibull distribution, the median discontinuation time can be estimated using equation (5.26), and is given by

$$\hat{t}(50) = \left\{\frac{1}{0.000454}\log 2\right\}^{1/1.676} = 79.27.$$

As a check, notice that this is perfectly consistent with the value of the discontinuation time corresponding to $\hat{S}(t) = 0.5$ in Figure 5.10. The standard

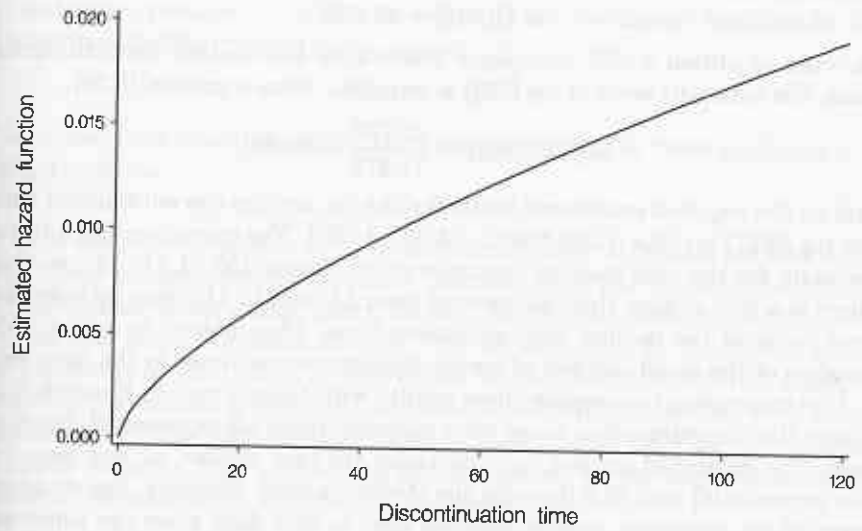

Figure 5.9 *Estimated hazard function on fitting the Weibull distribution.*

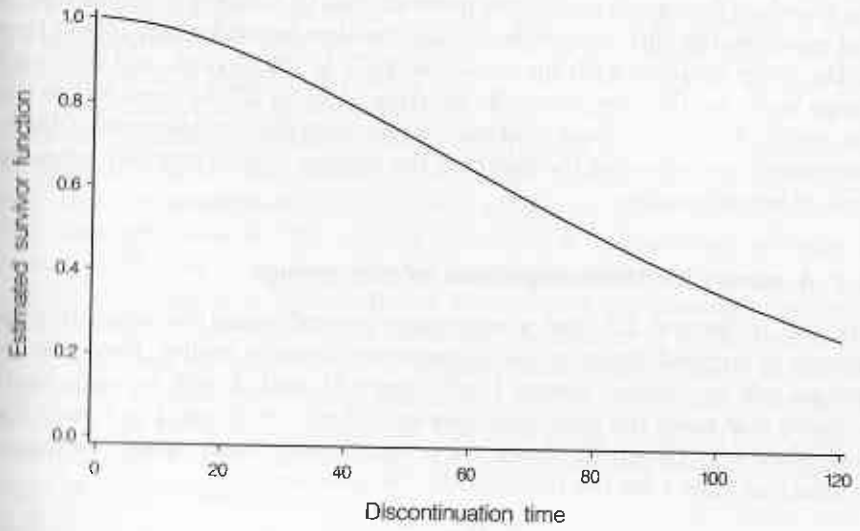

Figure 5.10 *Estimated survivor function on fitting the Weibull distribution.*

error of this estimate, from equation (5.27) is, after much arithmetic, found to be

$$\text{se}\{\hat{t}(50)\} = 15.795.$$

In order to obtain a 95% confidence interval for the median discontinuation time, the standard error of $\log \hat{t}(50)$ is required. From equation (5.28),

$$\text{se}\{\log \hat{t}(50)\} = \frac{15.795}{79.272} = 0.199,$$

and so the required confidence limits for the log median discontinuation time are $\log 79.272 \pm 1.96 \times 0.199$, that is, $(3.982, 4.763)$. The corresponding interval estimate for the true median discontinuation time is $(53.64, 117.15)$, so that there is a 95% chance that the interval from 54 days to 117 days includes the true value of the median discontinuation time. This interval is rather wide because of the small number of actual discontinuation times in the data set.

It is interesting to compare these results with those found in Example 5.2, where the discontinuation times were modelled using an exponential distribution. The estimated median survival times are very similar, at 80.6 days for the exponential and 79.3 days for the Weibull model. However, the standard error of the estimated median survival time is 26.8 days when the times are assumed to have an exponential distribution, and only 15.8 days under the Weibull model. The median is therefore estimated more precisely when the discontinuation times are assumed to have a Weibull distribution.

Other percentiles of the discontinuation time distribution, and accompanying standard errors and confidence intervals, can be found in a similar fashion. For example, the 90th percentile, that is, the time beyond which 10% of those in the study continue with the use of the IUD, is 162.23 days, and 95% confidence limits for the true percentile are from 95.41 to 275.84 days. Notice that the width of this confidence interval is larger than that for the median discontinuation time, reflecting the fact that the median is more precisely estimated than other percentiles.

5.4 A model for the comparison of two groups

We saw in Section 3.1 that a convenient general model for comparing two groups of survival times is the proportional hazards model. Here, the two groups will be labelled Group I and Group II, and X will be an indicator variable that takes the value zero if an individual is in Group I and unity if an individual is in Group II. Under the proportional hazards model, the hazard of death at time t for the ith individual is given by

$$h_i(t) = e^{\beta x_i} h_0(t), \tag{5.29}$$

where x_i is the value of X for the ith individual. Consequently, the hazard at time t for an individual in Group I is $h_0(t)$, and that for an individual in Group II is $\psi h_0(t)$, where $\psi = \exp(\beta)$. The quantity β is then the logarithm of the ratio of the hazard for an individual in Group II, to that of an individual in Group I.

We will now make the additional assumption that the survival times for the individuals in Group I have a Weibull distribution with scale parameter λ and shape parameter γ. Using equation (5.29), the hazard function for the individuals in this group is $h_0(t)$, where

$$h_0(t) = \lambda\gamma t^{\gamma-1}.$$

Now, also from equation (5.29), the hazard function for those in Group II is $\psi h_0(t)$, that is,

$$\psi\lambda\gamma t^{\gamma-1}.$$

This is the hazard function for a Weibull distribution with scale parameter $\psi\lambda$ and shape parameter γ. We therefore have the result that if the survival times of individuals in one group have a Weibull distribution with shape parameter γ, and the hazard of death at time t for an individual in the second group is proportional to that of an individual in the first, the survival times of those in the second group will also have a Weibull distribution with shape parameter γ. The Weibull distribution is then said to have the *proportional hazards property*. This property is another reason for the importance of the Weibull distribution in the analysis of survival data.

5.4.1 The log-cumulative hazard plot

When a single sample of survival times has a Weibull distribution $W(\lambda, \gamma)$, the log-cumulative hazard plot described in Section 5.2 will give a straight line with intercept $\log \lambda$ and slope γ. It then follows that if the survival times in a second group have a $W(\psi\lambda, \gamma)$ distribution, as they would under the proportional hazards model in equation (5.29), the log-cumulative hazard plot will give a straight line, also of slope γ, but with intercept $\log \psi + \log \lambda$. If the estimated log-cumulative hazard function is plotted against the logarithm of the survival time for individuals in two groups, parallel straight lines would mean that the assumptions of a proportional hazards model and Weibull survival times were tenable. The vertical separation of the two lines provides an estimate of $\beta = \log \psi$, the logarithm of the relative hazard.

If the two lines in a log-cumulative hazard plot are essentially straight, but not parallel, this means that the shape parameter, γ, is different in the two groups, and the hazards are no longer proportional. If the lines are not particularly straight, the Weibull model may not be appropriate. However, if the curves can be taken to be parallel, this would mean that the proportional hazards model is valid, and the Cox regression model discussed in Chapter 3 might be more satisfactory.

Example 5.4 Prognosis for women with breast cancer

In this example, we investigate whether the Weibull proportional hazards model is likely to be appropriate for the data of Example 1.2 on the survival times of breast cancer patients. These data relate to women classified according to whether their tumours were positively or negatively stained. The Kaplan-Meier estimate of the survivor functions for the women in each group