denominator cancels out, and we are left with

$$\frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_{(j)})}{\sum_{l\in R(t_{(j)})}\exp(\boldsymbol{\beta}'\boldsymbol{x}_l)}.$$

Finally, taking the product of these conditional probabilities over the r death times gives the likelihood function in equation (3.4).

The likelihood function that has been obtained is not a true likelihood, since it does not make direct use of the actual censored and uncensored survival times. For this reason it is referred to as a *partial likelihood function.*

In order to throw more light on the structure of the partial likelihood, consider a sample of survival data from five individuals, numbered from 1 to 5. The survival data are illustrated in Figure 3.1.

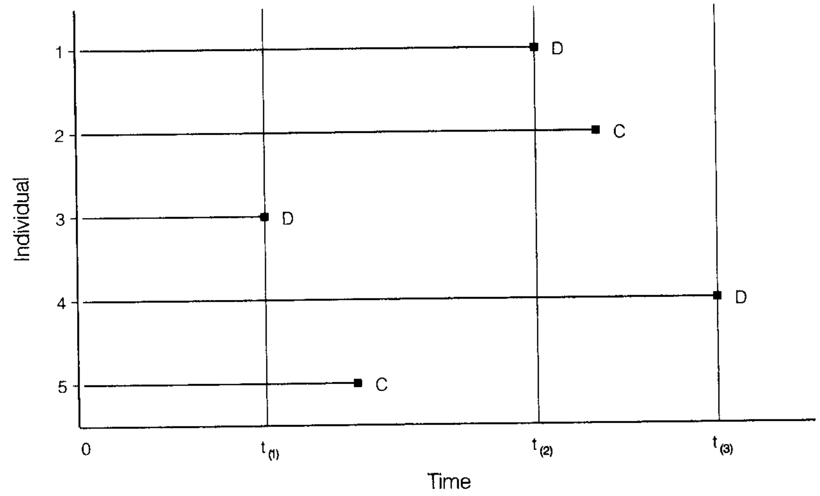

Figure 3.1 *Survival times of five individuals.*

The observed survival times of individuals 2 and 5 will be taken to be right-censored, and the three ordered death times are denoted $t_{(1)} < t_{(2)} < t_{(3)}$. Then, $t_{(1)}$ is the death time of individual 3, $t_{(2)}$ is that of individual 1, and $t_{(3)}$ that of individual 4.

The risk set at each of the three ordered death times consists of the individuals who are alive and uncensored just prior to each death time. Hence, the risk set $R(t_{(1)})$ consists of all five individuals, risk set $R(t_{(2)})$ consists of individuals 1, 2 and 4, while risk set $R(t_{(3)})$ only includes individual 4. Now write $\psi(i) = \exp(\boldsymbol{\beta}'\boldsymbol{x}_i)$, $i = 1, 2, \ldots, 5$, for the risk score for the ith individual, where $\boldsymbol{x}_i$ is the vector of explanatory variables for that individual. The numerators of the partial likelihood function for times $t_{(1)}$, $t_{(2)}$ and $t_{(3)}$, respectively, are $\psi(3)$, $\psi(1)$ and $\psi(4)$, since individuals 3, 1 and 4, respectively, die at the three ordered death times. The partial likelihood function over the three death times is then

$$\frac{\psi(3)}{\psi(1)+\psi(2)+\psi(3)+\psi(4)+\psi(5)} \times \frac{\psi(1)}{\psi(1)+\psi(2)+\psi(4)} \times \frac{\psi(4)}{\psi(4)}.$$

It turns out that standard results used in maximum likelihood estimation carry over without modification to maximum partial likelihood estimation. In particular, the results given in Appendix A for the variance-covariance matrix of the estimates of the β's can be used, as can distributional results associated with likelihood ratio testing, to be discussed in Section 3.4.

3.3.2 Treatment of ties*

The proportional hazards model for survival data assumes that the hazard function is continuous, and under this assumption, tied survival times are not possible. Of course, survival times are usually recorded to the nearest day, month or year, and so tied survival times can arise as a result of this rounding process. Indeed, Examples 1.2, 1.3 and 1.4 in Chapter 1 all contain tied observations.

In addition to the possibility of more than one death at a given time, there might also be one or more censored observations at a death time. When there are both censored survival times and deaths at a given time, the censoring is assumed to occur after all the deaths. Potential ambiguity concerning which individuals should be included in the risk set at that death time is then resolved and tied censored observations present no further difficulties in the computation of the likelihood function using equation (3.4). Accordingly, we only need consider how tied survival times can be handled in fitting the proportional hazards model.

In order to accommodate tied observations, the likelihood function in equation (3.4) has to be modified in some way. The appropriate likelihood function in the presence of tied observations has been given by Kalbfleisch and Prentice (2002). However, this likelihood has a very complicated form, and will not be reproduced here. In addition, the computation of this likelihood function can be very time consuming, particularly when there are a relatively large number of ties at one or more death times. Fortunately, there are a number of approximations to the likelihood function that have computational advantages over the exact method. But before these are given, some additional notation needs to be developed.

Let $\boldsymbol{s}_j$ be the vector of sums of each of the p covariates for those individuals who die at the jth death time, $t_{(j)}$, $j = 1, 2, \ldots, r$. If there are d_j deaths at $t_{(j)}$, the hth element of $\boldsymbol{s}_j$ is $s_{hj} = \sum_{k=1}^{d_j} x_{hjk}$, where x_{hjk} is the value of the hth explanatory variable, $h = 1, 2, \ldots, p$, for the kth of d_j individuals, $k = 1, 2, \ldots, d_j$, who die at the jth death time, $j = 1, 2, \ldots, r$.

The simplest approximation to the likelihood function is that due to Breslow

(1974), who proposed the approximate likelihood

$$\prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{s}_j)}{\left\{\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}'\boldsymbol{x}_l)\right\}^{d_j}}. \tag{3.9}$$

In this approximation, the d_j deaths at time $t_{(j)}$ are considered to be distinct and to occur sequentially. The probabilities of all possible sequences of deaths are then summed to give the likelihood in equation (3.9). Apart from a constant of proportionality, this is also the approximation suggested by Peto (1972). This likelihood is quite straightforward to compute, and is an adequate approximation when the number of tied observations at any one death time is not too large. For these reasons, this method is usually the default procedure for handling ties in statistical software for survival analysis, and will be used in the examples given in this book.

Efron (1977) proposed

$$\prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{s}_j)}{\prod_{k=1}^{d_j}\left[\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}'\boldsymbol{s}_l) - (k-1)d_k^{-1}\sum_{l \in D(t_{(j)})} \exp(\boldsymbol{\beta}'\boldsymbol{x}_l)\right]} \tag{3.10}$$

as an approximate likelihood function for the proportional hazards model, where $D(t_{(j)})$ is the set of all individuals who die at time $t_{(j)}$. This is a closer approximation to the appropriate likelihood function than that due to Breslow, although in practice, both approximations often give similar results.

Cox (1972) suggested the approximation

$$\prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{s}_j)}{\sum_{l \in R(t_{(j)};d_j)} \exp(\boldsymbol{\beta}'\boldsymbol{s}_l)}, \tag{3.11}$$

where the notation $R(t_{(j)}; d_j)$ denotes a set of d_j individuals drawn from $R(t_{(j)})$, the risk set at $t_{(j)}$. The summation in the denominator is the sum over all possible sets of d_j individuals sampled from the risk set without replacement. The approximation in expression (3.11) is based on a model for the situation where the time-scale is discrete, so that under this model, tied observations are permissible. Now, the hazard function for an individual with vector of explanatory variables $\boldsymbol{x}_i$, $h_i(t)$, is the probability of death in the unit time interval $(t, t+1)$, conditional on survival to time t. A discrete version of the proportional hazards model of equation (3.3) is the model

$$\frac{h_i(t)}{1-h_i(t)} = \exp(\boldsymbol{\beta}'\boldsymbol{x}_i)\frac{h_0(t)}{1-h_0(t)},$$

for which the likelihood function is that given in equation (3.11). In fact, in the limit as the width of the discrete time intervals becomes zero, this model tends to the proportional hazards model of equation (3.3).

When there are no ties, that is, when $d_j = 1$ for each death time, the approximations in equations (3.9), (3.10), and (3.11) all reduce to the likelihood function in equation (3.4).

3.3.3* *The Newton-Raphson procedure*

Models for censored survival data are usually fitted by using the Newton-Raphson procedure to maximise the partial likelihood function, and so the procedure is outlined in this section.

Let $\boldsymbol{u}(\boldsymbol{\beta})$ be the $p \times 1$ vector of first derivatives of the log-likelihood function in equation (3.5) with respect to the β-parameters. This quantity is known as the *vector of efficient scores.* Also, let $\boldsymbol{I}(\boldsymbol{\beta})$ be the $p \times p$ matrix of negative second derivatives of the log-likelihood, so that the (j,k)th element of $\boldsymbol{I}(\boldsymbol{\beta})$ is

$$-\frac{\partial^2 \log L(\boldsymbol{\beta})}{\partial\beta_j \partial\beta_k}.$$

The matrix $\boldsymbol{I}(\boldsymbol{\beta})$is known as the *observed information matrix.*

According to the Newton-Raphson procedure, an estimate of the vector of β-parameters at the $(s+1)$th cycle of the iterative procedure, $\hat{\boldsymbol{\beta}}_{s+1}$, is

$$\hat{\boldsymbol{\beta}}_{s+1} = \hat{\boldsymbol{\beta}}_s + \boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}}_s)\boldsymbol{u}(\hat{\boldsymbol{\beta}}_s),$$

for $s = 0, 1, 2, \ldots$, where $\boldsymbol{u}(\hat{\boldsymbol{\beta}}_s)$ is the vector of efficient scores and $\boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}}_s)$ is the inverse of the information matrix, both evaluated at $\hat{\boldsymbol{\beta}}_s$. The procedure can be started by taking $\hat{\boldsymbol{\beta}}_0 = \mathbf{0}$. The process is terminated when the change in the log-likelihood function is sufficiently small, or when the largest of the relative changes in the values of the parameter estimates is sufficiently small.

When the iterative procedure has converged, the variance-covariance matrix of the parameter estimates can be approximated by the inverse of the information matrix, evaluated at $\hat{\boldsymbol{\beta}}$, that is, $\boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}})$. The square root of the diagonal elements of this matrix are then the standard errors of the estimated values of $\beta_1, \beta_2, \ldots, \beta_p$.

3.4 Confidence intervals and hypothesis tests for the β's

When a statistical package is used to fit a proportional hazards model, the parameter estimates that are provided are usually accompanied by their standard errors. These standard errors can be used to obtain approximate confidence intervals for the unknown β-parameters. In particular, a $100(1-\alpha)$% confidence interval for a parameter β is the interval with limits $\hat{\beta} \pm z_{\alpha/2}\,\mathrm{se}\,(\hat{\beta})$, where $\hat{\beta}$ is the estimate of β, and $z_{\alpha/2}$ is the upper $\alpha/2$-point of the standard normal distribution.

If a $100(1-\alpha)$% confidence interval for β does not include zero, this is evidence that the value of β is non-zero. More specifically, the null hypothesis that $\beta = 0$ can be tested by calculating the value of the statistic $\hat{\beta}/\,\mathrm{se}\,(\hat{\beta})$. The observed value of this statistic is then compared to percentage points of the standard normal distribution in order to obtain the corresponding P-value. Equivalently, the square of this statistic can be compared with percentage points of a chi-squared distribution on one degree of freedom. This procedure is sometimes called a *Wald test.* Indeed, the P-values for this test are often given alongside parameter estimates and their standard errors in computer output.

When attempting to interpret the P-value for a given parameter, β_j, say, it is important to recognise that the hypothesis that is being tested is that $\beta_j = 0$ in the presence of all other terms that are in the model. For example, suppose that a model contains the three explanatory variables X_1, X_2, X_3, and that their coefficients are $\beta_1, \beta_2, \beta_3$. The test statistic $\hat{\beta}_2 / \mathrm{se}\,(\hat{\beta}_2)$ is then used to test the null hypothesis that $\beta_2 = 0$ in the presence of β_1 and β_3. If there was no evidence to reject this hypothesis, we would conclude that X_2 was not needed in the model in the presence of X_1 and X_3.

In general, the individual estimates of the β's in a proportional hazards model are not all independent of one another. This means that the results of testing separate hypotheses about the β-parameters in a model may not be easy to interpret. For example, consider again the situation where there are three explanatory variables, X_1, X_2, X_3. If $\hat{\beta}_1$ and $\hat{\beta}_2$ were not found to be significantly different from zero, when compared with their standard errors, we could not conclude that only X_3 need be included in the model. This is because the coefficient of X_1, for example, could well change when X_2 is excluded from the model, and vice versa. This would certainly happen if X_1 and X_2 were correlated.

Because of the difficulty in interpreting the results of tests concerning the coefficients of the explanatory variables in a model, alternative methods for comparing different proportional hazards models are required. It turns out that the methods to be described in Section 3.5 are much more satisfactory than the Wald tests. Little attention should therefore be paid to the results of these tests given in computer-based analyses of survival data.

3.4.1 Standard errors and confidence intervals for hazard ratios

We have seen that in situations where there are two groups of survival data, the parameter β is the logarithm of the ratio of the hazard of death at time t for individuals in one group relative to those in the other. Hence the hazard ratio itself is $\psi = e^{\beta}$. The corresponding estimate of the hazard ratio is $\hat{\psi} = \exp(\hat{\beta})$, and the standard error of $\hat{\psi}$ can be obtained from the standard error of $\hat{\beta}$ using the result given as equation (2.9) in Chapter 2. From this result, the approximate variance of $\hat{\psi}$, a function of $\hat{\beta}$, is

$$\left\{\exp(\hat{\beta})\right\}^2 \,\mathrm{var}\,(\hat{\beta}),$$

that is, $\hat{\psi}^2\,\mathrm{var}\,(\hat{\beta})$, and so the standard error of $\hat{\psi}$ is given by

$$\mathrm{se}\,(\hat{\psi}) = \hat{\psi}\,\mathrm{se}\,(\hat{\beta}). \tag{3.12}$$

Generally speaking, a confidence interval for the true hazard ratio will be more informative than the standard error of the estimated hazard ratio. A $100(1-\alpha)\%$ confidence interval for the true hazard ratio, ψ, can be found simply by exponentiating the confidence limits for β. An interval estimate obtained in this way is preferable to one found using $\hat{\psi} \pm z_{\alpha/2}\,\mathrm{se}\,(\hat{\psi})$. This is because the distribution of the logarithm of the estimated hazard ratio will be more closely approximated by a normal distribution than that of the hazard ratio itself.

The construction of a confidence interval for a hazard ratio is illustrated in Example 3.1 below. Fuller details on the interpretation of the parameters in the linear component of a proportional hazards model are given in Section 3.7.

3.4.2 Two examples

In this section, the results of fitting a proportional hazards model to data from two of the examples introduced in Chapter 1 are given.

Example 3.1 Prognosis for women with breast cancer

Data on the survival times of breast cancer patients, classified according to whether or not sections of their tumours were positively stained, were first given in Example 1.2. The variable that indexes the result of the staining process can be regarded as a factor with two levels. From the arguments given in Section 3.2.1, this factor can be fitted by using an indicator variable X to denote the staining result, where $X = 0$ corresponds to negative staining and $X = 1$ to positive staining. Under the proportional hazards model, the hazard of death at time t for the ith woman, for whom the value of the indicator variable is x_i, is

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

where x_i is zero or unity. The baseline hazard function $h_0(t)$ is then the hazard function for a women with a negatively stained tumour. This is essentially the model considered in Section 3.1.1, and given in equation (3.2).

In the group of women whose tumours were positively stained, there are two who die at 26 months. To cope with this tie, the Breslow approximation to the likelihood function will be used. This model is fitted by finding that value of β, $\hat{\beta}$, which maximises the likelihood function in equation (3.9). The maximum likelihood estimate of β is $\hat{\beta} = 0.908$. The standard error of this estimate is also obtained from statistical packages for fitting the Cox regression model, and turns out to be given by $\mathrm{se}\,(\hat{\beta}) = 0.501$.

The quantity e^{β} is the ratio of the hazard function for a woman with $X = 1$ to that for a woman with $X = 0$, so that β is the logarithm of the ratio of the hazard of death at time t for positively stained relative to negatively stained women. The estimated value of this hazard ratio is $e^{0.908} = 2.48$. Since this is greater than unity, we conclude that a woman who has a positively stained tumour will have a greater risk of death at any given time than a comparable women whose tumour was negatively stained. Positive staining therefore indicates a poorer prognosis for a breast cancer patient.

The standard error of the hazard ratio can be found from the standard error of $\hat{\beta}$, using the result in equation (3.12). Since the estimated relative hazard is $\hat{\psi} = \exp(\hat{\beta}) = 2.480$, and the standard error of $\hat{\beta}$ is 0.501, the standard error of $\hat{\psi}$ is given by

$$\text{se}(\hat{\psi}) = 2.480 \times 0.501 = 1.242.$$

We can go further and construct a confidence interval for this hazard ratio. The first step is to obtain a confidence interval for the logarithm of the hazard ratio, β. For example, a 95% confidence interval for β is the interval from $\hat{\beta} - 1.96\,\text{se}(\hat{\beta})$ to $\hat{\beta} + 1.96\,\text{se}(\hat{\beta})$, that is, the interval from -0.074 to 1.890. Exponentiating these confidence limits gives $(0.93, 6.62)$ as a 95% confidence interval for the hazard ratio itself. Notice that this interval barely includes unity, suggesting that there is evidence that the two groups of women have a different survival experience.

Example 3.2 Survival of multiple myeloma patients

Data on the survival times of 48 patients suffering from multiple myeloma were given in Example 1.3. The data base also contains the values of seven other variables that were recorded for each patient. For convenience, the values of the variable that describes the sex of a patient have been redefined to be zero and unity for males and females respectively. The variables are then as follows:

Age:	Age of the patient,
Sex:	Sex of the patient (0 = male, 1 = female),
Bun:	Blood urea nitrogen,
Ca:	Serum calcium,
Hb:	Serum haemoglobin,
Pcells:	Percentage of plasma cells,
Protein:	Bence-Jones protein (0 = absent, 1 = present).

The sex of the patient and the variable associated with the occurrence of Bence-Jones protein are factors with two levels. These terms are fitted using the indicator variables *Sex* and *Protein*. The proportional hazards model for the ith individual is then

$$\begin{aligned} h_i(t) = {} & \exp(\beta_1 Age_i + \beta_2 Sex_i + \beta_3 Bun_i + \beta_4 Ca_i + \beta_5 Hb_i \\ & + \beta_6 Pcells_i + \beta_7 Protein_i) h_0(t), \end{aligned}$$

where the subscript i on an explanatory variable denotes the value of that variable for the ith individual. The baseline hazard function is the hazard function for an individual for whom the values of all seven of these variables are zero. This function therefore corresponds to a male aged zero, who has zero values of *Bun*, *Ca*, *Hb* and *Pcells*, and no Bence-Jones protein. In view of the obvious difficulty in interpreting this function, it might be more sensible to redefine the variables *Age*, *Bun*, *Ca*, *Hb* and *Pcells* by subtracting values for an average patient. For example, if we took *Age* − 60 in place of *Age*, the baseline hazard would correspond to a male aged 60 years. This procedure also avoids the introduction of a function that describes the hazard of individuals whose ages are rather different from the age range of patients in the study. Although this leads to a baseline hazard function that has a more natural interpretation, it will not affect inference about the influence of the explanatory variables on the hazard of death. For this reason, the untransformed variables will be used in this example. On fitting the model, the estimates of the coefficients of the explanatory variables and their standard errors are found to be those shown in Table 3.1.

Table 3.1 *Estimated values of the coefficients of the explanatory variables on fitting a proportional hazards model to the data from Example 1.3.*

Variable	$\hat{\beta}$	se $(\hat{\beta})$
Age	−0.019	0.028
Sex	−0.251	0.402
Bun	0.021	0.006
Ca	0.013	0.132
Hb	−0.135	0.069
Pcells	−0.002	0.007
Protein	−0.640	0.427

We see from Table 3.1 that some of the estimates are close to zero. Indeed, if individual 95% confidence intervals are calculated for the coefficients of the seven variables, only those for *Bun* and *Hb* exclude zero. This suggests that the hazard function does not depend on all seven explanatory variables.

However, we cannot deduce from this that *Bun* and *Hb* are the relevant variables, since the estimates of the coefficients of the seven explanatory variables in the fitted model are not independent of one another. This means that if one of the seven explanatory variables were excluded from the model, the coefficients of the remaining six might be different from those in Table 3.1. For example, if *Bun* is omitted, the estimated coefficients of the six remaining explanatory variables, *Age*, *Sex*, *Ca*, *Hb*, *Pcells* and *Protein*, turn out to be −0.009, −0.301, −0.036, −0.140, −0.001, and −0.420, respectively. Comparison with the values shown in Table 3.1 shows that there are differences in the estimated coefficients of each of these six variables, although in this case the differences are not very great.

In general, to determine on which of the seven explanatory variables the hazard function depends, a number of different models will need to be fitted, and the results compared. Methods for comparing the fit of alternative models, and strategies for model building are considered in subsequent sections of this chapter.

3.5 Comparing alternative models

In a modelling approach to the analysis of survival data, a model is developed for the dependence of the hazard function on one or more explanatory variables. In this development process, proportional hazards models with linear components that contain different sets of terms are fitted, and comparisons made between them.

As a specific example, consider the situation where there are two groups of survival times, corresponding to individuals who receive either a new treatment or a standard. The common hazard function under the model for no treatment difference can be taken to be $h_0(t)$. This model is a special case of the general proportional hazards model in equation (3.3), in which there are no explanatory variables in the linear component of the model. This model is therefore referred to as the *null model.*

Now let X be an indicator variable that takes the value zero for individuals receiving the standard treatment and unity otherwise. Under a proportional hazards model, the hazard function for an individual for whom X takes the value x is $e^{\beta x}h_0(t)$. The hazard functions for individuals on the standard and new treatments are then $h_0(t)$ and $e^{\beta}h_0(t)$, respectively. The difference between this model and the null model is that the linear component of the latter contains the additional term βx. Since $\beta = 0$ corresponds to no treatment effect, the extent of any treatment difference can be investigated by comparing these two proportional hazards models for the observed survival data.

More generally, suppose that two models are contemplated for a particular data set, Model (1) and Model (2), say, where Model (1) contains a subset of the terms in Model (2). Model (1) is then said to be *parametrically nested* within Model (2). Specifically, suppose that the p explanatory variables, $X_1, X_2, \ldots, X_p$, are fitted in Model (1), so that the hazard function under this model can be written as

$$\exp\{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\}h_0(t).$$

Also suppose that the $p+q$ explanatory variables $X_1, X_2, \ldots, X_p, X_{p+1}, \ldots,$ X_{p+q} are fitted in Model (2), so that the hazard function under this model is

$$\exp\{\beta_1 x_1 + \cdots + \beta_p x_p + \beta_{p+1}x_{p+1} + \cdots + \beta_{p+q}x_{p+q}\}h_0(t).$$

Model (2) then contains the q additional explanatory variables $X_{p+1}, X_{p+2},$ $\ldots, X_{p+q}$. Because Model (2) has a larger number of terms than Model (1), Model (2) must be a better fit to the observed data. The statistical problem is then to determine whether the additional q terms in Model (2) significantly improve the explanatory power of the model. If not, they might be omitted, and Model (1) would be deemed to be adequate.

In the discussion of Example 3.2, we saw that when there are a number of explanatory variables of possible relevance, the effect of each term cannot be studied independently of the others. The effect of any given term therefore depends on the other terms currently included in the model. For example, in Model (1), the effect of any of the p explanatory variables on the hazard function depends on the $p-1$ variables that have already been fitted, and so the effect of X_p is said to be *adjusted* for the remaining $p-1$ variables. In particular, the effect of X_p is adjusted for $X_1, X_2, \ldots, X_{p-1}$, but we also speak of the effect of X_p *eliminating* or *allowing for* $X_1, X_2, \ldots, X_{p-1}$. Similarly, when the q variables $X_{p+1}, X_{p+2}, \ldots, X_{p+q}$ are added to Model (1), the effect of these variables on the hazard function is said to be adjusted for the p variables that have already been fitted, $X_1, X_2, \ldots, X_p$.

3.5.1 The statistic $-2\log\hat{L}$

In order to compare alternative models fitted to an observed set of survival data, a statistic that measures the extent to which the data are fitted by a particular model is required. Since the likelihood function summarises the information that the data contain about the unknown parameters in a given model, a suitable summary statistic is the value of the likelihood function when the parameters are replaced by their maximum likelihood estimates. This is the maximised likelihood under an assumed model, and can be computed from equation (3.4) by replacing the β's by their maximum likelihood estimates under the model. For a given set of data, the larger the value of the maximised likelihood, the better is the agreement between the model and the observed data.

For reasons given in the sequel, it is more convenient to use minus twice the logarithm of the maximised likelihood in comparing alternative models. If the maximised likelihood for a given model is denoted by $\hat{L}$, the summary measure of agreement between the model and the data is $-2\log\hat{L}$. From Section 3.3.1, $\hat{L}$ is in fact the product of a series of conditional probabilities, and so this statistic will be less than unity. In consequence, $-2\log\hat{L}$ will always be positive, and for a given data set, the smaller the value of $-2\log\hat{L}$, the better the model.

The statistic $-2\log\hat{L}$ cannot be used on its own as a measure of model adequacy. The reason for this is that the value of $\hat{L}$, and hence of $-2\log\hat{L}$, is dependent upon the number of observations in the data set. Thus if, after fitting a model to a set of data, additional data became available to which the fit of the model was the same as that to the original data, the value of $-2\log\hat{L}$ for the enlarged data set would be different from that of the original data. Accordingly the value of $-2\log\hat{L}$ is only useful when making comparisons between models fitted to the same data.

3.5.2 Comparing nested models

Consider again Model (1) and Model (2) defined above, and let the value of the maximised log-likelihood function for each model be denoted by $\hat{L}(1)$ and $\hat{L}(2)$, respectively. The two models can then be compared on the basis of the difference between the values of $-2\log\hat{L}$ for each model. In particular, a large difference between $-2\log\hat{L}(1)$ and $-2\log\hat{L}(2)$ would lead to the conclusion that the q variates in Model (2), that are additional to those in Model (1), do improve the adequacy of the model. Naturally, the amount by which the value of $-2\log\hat{L}$ changes when terms are added to a model will depend on which terms have already been included. In particular, the difference in the values of $-2\log\hat{L}(1)$ and $-2\log\hat{L}(2)$, that is, $-2\log\hat{L}(1) + 2\log\hat{L}(2)$, will reflect the combined effect of adding the variables $X_{p+1}, X_{p+2}, \ldots, X_{p+q}$ to a model that already contains $X_1, X_2, \ldots, X_p$. This is said to be the change in the value of $-2\log\hat{L}$ due to fitting $X_{p+1}, X_{p+2}, \ldots, X_{p+q}$, adjusted for $X_1, X_2, \ldots, X_p$.

The statistic $-2\log\hat{L}(1) + 2\log\hat{L}(2)$, can be written as

$$-2\log\{\hat{L}(1)/\hat{L}(2)\},$$

and this is the log-likelihood ratio statistic for testing the null hypothesis that the q parameters $\beta_{p+1}, \beta_{p+2}, \ldots, \beta_{p+q}$ in Model (2) are all zero. From results associated with the theory of likelihood ratio testing (see Appendix A), this statistic has an asymptotic chi-squared distribution, under the null hypothesis that the coefficients of the additional variables are zero. The number of degrees of freedom of this chi-squared distribution is equal to the difference between the number of independent β-parameters being fitted under the two models. Hence, in order to compare the value of $-2\log\hat{L}$ for Model (1) and Model (2), we use the fact that the statistic $-2\log\hat{L}(1) + 2\log\hat{L}(2)$ has a chi-squared distribution on q degrees of freedom, under the null hypothesis that $\beta_{p+1}, \beta_{p+2}, \ldots, \beta_{p+q}$ are all zero. If the observed value of the statistic is not significantly large, the two models will be adjudged to be equally suitable. Then, other things being equal, the more simple model, that is, the one with fewer terms, would be preferred. On the other hand, if the values of $-2\log\hat{L}$ for the two models are significantly different, we would argue that the additional terms are needed and the more complex model would be adopted.

Some texts, and some software packages, ascribe degrees of freedom to the quantity $-2\log\hat{L}$. However, the value of $-2\log\hat{L}$ for a particular model does not have a chi-squared distribution, and so the quantity cannot be considered to have an associated number of degrees of freedom. Additionally, the quantity $-2\log\hat{L}$ is sometimes referred to as a *deviance*. This is also inappropriate, since unlike the deviance used in the context of generalised linear modelling, $-2\log\hat{L}$ does not measure deviation from a model that is a perfect fit to the data.

Example 3.3 Prognosis for women with breast cancer

Consider again the data from Example 1.2 on the survival times of breast cancer patients. On fitting a proportional hazards model that contains no explanatory variables, that is, the null model, the value of $-2\log\hat{L}$ is 173.968. As in Example 3.1, the indicator variable X, will be used to represent the result of the staining procedure, so that X is zero for women whose tumours are negatively stained and unity otherwise. When the variable X is included in the linear component of the model, the value of $-2\log\hat{L}$ decreases to 170.096. The values of $-2\log\hat{L}$ for alternative models are conveniently summarised in tabular form, as illustrated in Table 3.2.

The difference between the values of $-2\log\hat{L}$ for the null model and the model that contains X can be used to assess the significance of the difference between the hazard functions for the two groups of women. Since one model contains one more β-parameter than the other, the difference in the values of $-2\log\hat{L}$ has a chi-squared distribution on one degree of freedom. The difference in the two values of $-2\log\hat{L}$ is 173.968 − 170.096 = 3.872, which is just significant at the 5% level ($P = 0.049$). We may therefore conclude that there is evidence, significant at the 5% level, that the hazard functions for the two groups of women are different.

Table 3.2 *Values of* $-2\log\hat{L}$ *on fitting proportional hazards models to the data from Example 1.2.*

Variables in model	$-2\log\hat{L}$
none	173.968
X	170.096

In Example 2.12, the extent of the difference between the survival times of the two groups of women was investigated using the log-rank test. The chi-squared value for this test was found to be 3.515 ($P = 0.061$). This value is not very different from the figure of 3.872 ($P = 0.049$) obtained above. The similarity of these two P-values means that essentially the same conclusions are drawn about the extent to which the data provide evidence against the null hypothesis of no group difference. From the practical viewpoint, the fact that one result is just significant at the 5% level, while the other is not quite significant at that level, is immaterial.

Although the model-based approach used in this example is operationally different from the log-rank test, the two procedures are in fact closely related. This relationship will be explored in greater detail in Section 3.9.

Example 3.4 Treatment of hypernephroma

In a study carried out at the University of Oklahoma Health Sciences Center, data were obtained on the survival times of 36 patients with a malignant tumour in the kidney, or hypernephroma. The patients had all been treated with a combination of chemotherapy and immunotherapy, but additionally a nephrectomy, the surgical removal of the kidney, had been carried out on some of the patients. Of particular interest is whether the survival time of the patients depends on their age at the time of diagnosis and on whether or not they had received a nephrectomy. The data obtained in the study were given in Lee and Wang (2003). In the data set to be used as a basis for this example, the age of a patient has been classified according to whether the patient is less than 60, between 60 and 70 or greater than 70. Table 3.3 gives the survival times of the patients in months, where an asterisk denotes a censored observation.

In this example, there is a factor, age group, with three levels (< 60, 60–70, > 70), and a factor associated with whether or not a nephrectomy was performed. There are a number of possible models for these data depending on whether the hazard function is related to neither, one or both of these factors. Suppose that the effect due to the jth age group is denoted by α_j, $j = 1, 2, 3$,