Table 4.5 *Values of the approximate likelihood displacement, LD_i, and the elements of $|\boldsymbol{l}_{\max}|$.*

Observation	LD_i	$\lvert\boldsymbol{l}_{\max}\rvert$
1	0.033	0.161
2	0.339	0.309
3	0.005	0.068
4	0.338	0.621
5	0.050	0.104
6	0.019	0.058
7	0.136	0.291
8	0.027	0.054
9	0.133	0.124
10	0.035	0.193
11	0.061	0.264
12	0.043	0.224
13	0.219	0.464

pair of parameter estimates. The elements corresponding to patients 2 and 13 are also large relative to the other values, suggesting that the data for these patients are also influential. The sum of the squares of elements 2, 4 and 13 of $\boldsymbol{l}_{\max}$ is 0.70. The total of the sums of squares of the elements is 1.00, and so cases 2, 4 and 13 account for nearly three-quarters of the variability in the elements of $\boldsymbol{l}_{\max}$. Note that the analysis of the delta-betas in Example 4.6 showed that the observations from patients 2 and 4 most influence the parameter estimate for *Sex*, while the observation for patient 13 has a greater effect on the estimate for *Age*.

In summary, the observations from patients 2, 4 and 13 affect the form of the hazard function to the greatest extent. Omitting each of these in turn gives the following estimates of the linear component in the hazard functions for the ith individual:

Omitting patient number 2: $0.031\, Age_i - 3.530\, Sex_i$,

Omitting patient number 4: $0.045\, Age_i - 3.529\, Sex_i$,

Omitting patient number 13: $0.011\, Age_i - 2.234\, Sex_i$.

For comparison, the linear component for the full data set is

$$0.030\, Age_i - 2.711\, Sex_i.$$

To illustrate the magnitude of the change in estimated hazard ratios, consider the relative hazard of infection at time t for a patient aged 50 years relative to one aged 40 years. For the full data set, this is $e^{0.304} = 1.355$. This value is

increased to 1.365 and 1.564 when patients 2 and 4, respectively, are omitted, and decreased to 1.114 when patient 13 is omitted. The effect on the hazard function of removing these patients from the data base is therefore not particularly marked.

In the same way, the hazard of infection at time t for a male patient ($Sex = 1$) relative to a female ($Sex = 2$) is $e^{2.711}$, that is, 5.041 for the full data set. When observations 2, 4, and 13 are omitted in turn, the hazard for males relative to females is 4.138, 4.097 and 9.334, respectively. Omission of the data from patient number 13 appears to have a great effect on the estimated hazard ratio. However, some caution is needed in interpreting this result. Since there are very few males in the data set, the estimated hazard ratio is imprecisely estimated. In fact, a 95% confidence interval for the hazard ratio, when the data from patient 13 are omitted, ranges from 0.012 to 82.96!

4.3.3 Treatment of influential observations

Once observations have been found to be unduly influential, it is difficult to offer any firm advice on what should be done about them. So much depends on the scientific background to the study.

When possible, the origin of influential observations should be checked. Errors in transcribing and recording categorical and numerical data frequently occur. If any mistakes are found, the data need to be corrected and the analysis repeated. If the observed value of a survival time, or other explanatory variables, is impossible, and correction is not possible, the corresponding observation should be omitted from the data base before repeating the analysis.

In many situations it will not be possible to confirm that the data corresponding to an influential observation are valid. Certainly, influential observations should not then be rejected outright. In these circumstances, the most appropriate course of action will be to establish the actual effect on the inferences to be drawn from the analysis. For example, if a relative hazard or median survival time is being used in quantifying the size of a treatment effect, the values of these statistics with and without the influential values can be contrasted. If the difference between the results is so small as to not be of practical importance, the queried observations can be retained. On the other hand, if the effect of removing the influential observations is large enough to be of practical importance, analyses based on both the full and reduced data sets will need to be reported. The outcome of consultations with the scientists involved in the study will then be a vital ingredient in the process of deciding on the course of future action.

Example 4.8 Survival of multiple myeloma patients
The effect of individual observations on the estimated values of the parameters of a Cox regression model fitted to the data from Example 1.3 will now be investigated. Plots of the approximate unstandardised delta-betas for *Hb* and *Bun* against the rank order of the survival times are shown in Figures 4.12 and 4.13.

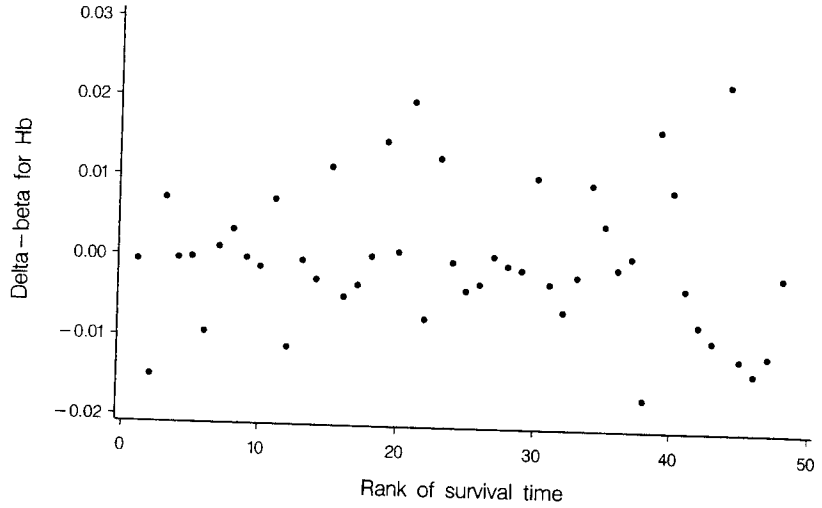

Figure 4.12 *Plot of the delta-betas for Hb against rank order of survival time.*

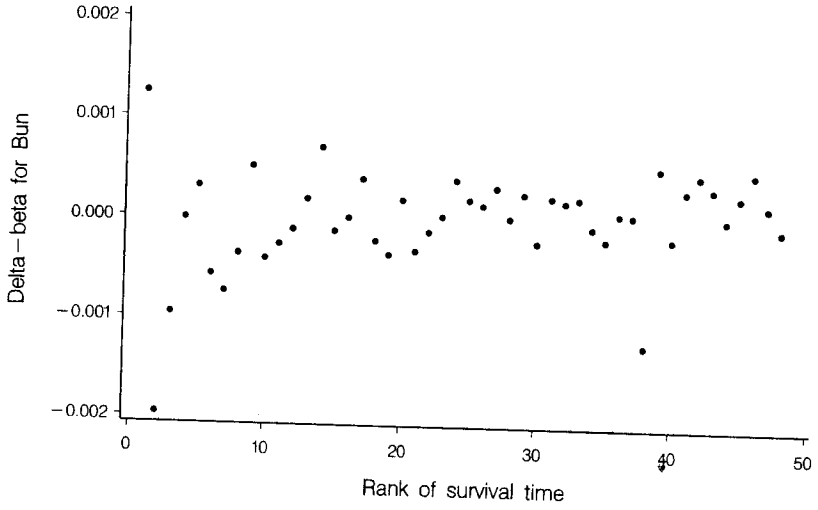

Figure 4.13 *Plot of the delta-betas for Bun against rank order of survival time.*

From Figure 4.12, no one observation stands out as having a delta-beta for *Hb* that is different from the rest. However, Figure 4.13 shows that the two observations with the shortest survival times have relatively large positive or large negative delta-betas for *Bun*. These correspond to patients 32 and 38 in the data given in Table 1.3. Patient 32 has a survival time of just one month, and the second largest value of *Bun*. Deletion of this observation from the data base decreases the parameter estimate for *Bun*. Patient number 38 also survived for just one month after trial entry, but has a value of *Bun* that is rather low for someone surviving for such a short time. If the data from this patient are omitted, the coefficient of *Bun* in the model is increased.

To identify observations that influence the set of parameter estimates, a plot of the absolute values of the elements of the diagnostic $l_{\max}$ against the rank order of the survival times is shown in Figure 4.14.

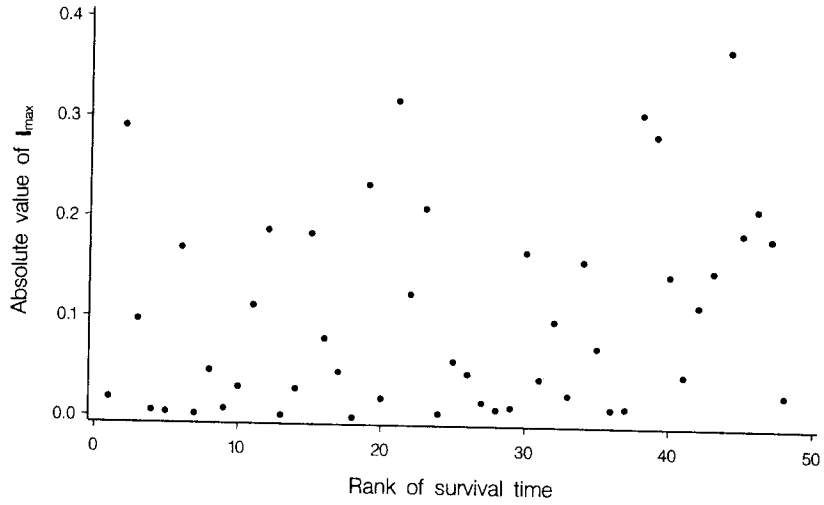

Figure 4.14 *Plot of the absolute values of the elements of* $l_{\max}$ *against rank order of survival time.*

The observation with the largest value of $|l_{\max}|$ corresponds to patient 13. This patient has an unusually small value of *Hb*, and a value of *Bun* that is a little high, for someone who has survived as long as 65 months. If this observation is omitted from the data set, the coefficient of *Bun* remains the same, but that of *Hb* is reduced from -0.134 to -0.157. The effect of *Hb* on the hazard of death is then a little more significant. In summary, the record for patient 13 has little effect on the form of the estimated hazard function.

4.4 Testing the assumption of proportional hazards

So far in this chapter we have concentrated on how the adequacy of the linear component of a survival model can be examined. A crucial assumption

made when using the Cox regression model is that of proportional hazards. If hazards are not proportional, this means that the linear component of the model varies with time in some manner. We must therefore consider how the validity of this assumption can be examined. In this section, a straightforward plot that can be used in advance of model fitting is first described, and this is followed by a description of how diagnostics derived from a fitted model can be used in examining the proportional hazards assumption.

4.4.1 *The log-cumulative hazard plot*

According to the Cox regression model, the hazard of death at any time t for the ith individual is given by

$$h_i(t) = \exp(\boldsymbol{\beta}'\boldsymbol{x}_i)h_0(t), \tag{4.16}$$

where $\boldsymbol{x}_i$ is the vector of values of explanatory variables for that individual, $\boldsymbol{\beta}$ is the corresponding vector of coefficients, and $h_0(t)$ is the baseline hazard function. Integrating both sides of this equation over t gives

$$\int_0^t h_i(u)\mathrm{d}u = \exp(\boldsymbol{\beta}'\boldsymbol{x}_i)\int_0^t h_0(u)\mathrm{d}u,$$

and so, using equation (1.6),

$$H_i(t) = \exp(\boldsymbol{\beta}'\boldsymbol{x}_i)H_0(t),$$

where $H_i(t)$ and $H_0(t)$ are the cumulative hazard functions. Taking logarithms of each side of this equation, we get

$$\log H_i(t) = \boldsymbol{\beta}'\boldsymbol{x}_i + \log H_0(t),$$

from which it follows that differences in the log-cumulative hazard functions do not depend on time. This means that if the log-cumulative hazard functions for individuals with different values of their explanatory variables are plotted against time, the curves so formed will be parallel if the proportional hazards model in equation (4.16) is valid. This provides the basis of a widely used diagnostic for assessing the validity of the proportional hazards assumption. It turns out that plotting the log-cumulative hazard functions against the logarithm of t, rather than t itself, is a useful diagnostic in parametric modelling, and so this form of plot is generally used; see Section 5.4.1 of Chapter 5 for further details on the use of this log-cumulative hazard plot.

To use this plot, the survival data are first grouped according to the levels of one or more factors. If continuous variables are to feature in this analysis, their values will first need to be grouped in some way to give a categorical variable. The Kaplan-Meier estimate of the survivor function of the data in each group is then obtained. A log-cumulative hazard plot, that is, a plot of the logarithm of the estimated cumulative hazard function against the logarithm of the survival time, will yield parallel curves if the hazards are proportional across the different groups. This method is informative, and simple to operate when there is a small number of factors, and a reasonable number of observations at each level. On the other hand, the plot will be based on very few observations at the later survival times, and in more highly structured data sets, a different approach needs to be taken.

Example 4.9 Survival of multiple myeloma patients

We again use the data on the survival times of 48 patients with multiple myeloma, to illustrate the log-cumulative hazard plot. In particular we will investigate whether the assumption of proportional hazards is valid in respect of the variable *Hb*, which is associated with the serum haemoglobin level. Because this is a continuous variable, we first need to categorise the values of *Hb*. This will be done in the same manner as in Example 3.7 of Chapter 3, where four groups were defined with values of *Hb* which are such that $Hb \leqslant 7$, $7 < Hb \leqslant 10$, $10 < Hb \leqslant 13$ and $Hb > 13$. The patients are then grouped according to their haemoglobin level, and the Kaplan-Meier estimate of the survivor function is obtained for each of the four groups. From this estimate, the estimated log-cumulative hazard is formed using the relation $\hat{H}(t) = -\log \hat{S}(t)$, from equation (1.7) of Chapter 1, and plotted against the values of $\log t$. The resulting log-cumulative hazard plot is shown in Figure 4.15.

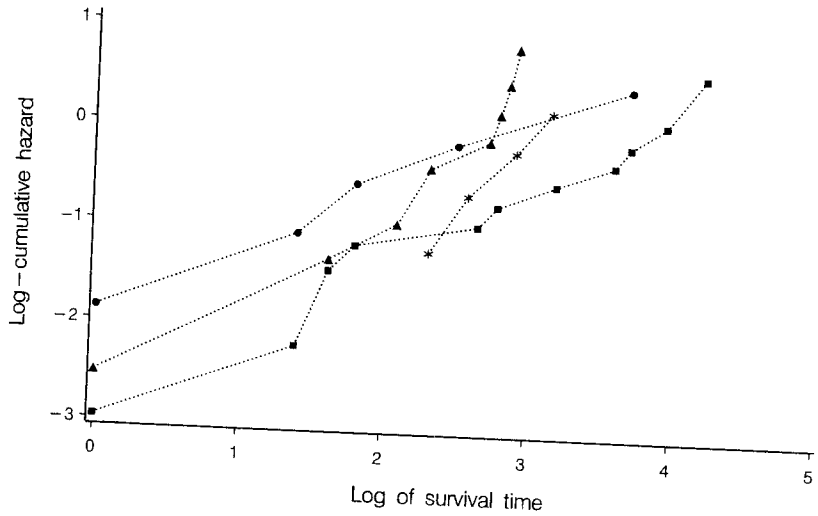

Figure 4.15 *Log-cumulative hazard plot for multiple myeloma patients in four groups defined by* $Hb \leqslant 7$ (•), $7 < Hb \leqslant 10$ (■), $10 < Hb \leqslant 13$ (▲) *and* $Hb > 13$ (*).

This figure indicates that the plots for $Hb \leqslant 7$, $7 < Hb \leqslant 10$, and $Hb > 13$ are roughly parallel. The plot for $10 < Hb \leqslant 13$ is not in line with the others, although this impression results from relatively large cumulative hazard estimates at the longest survival times experienced by patients in this group. This plot takes no account of the values of the other variable *Bun*, and it

could be that the survival times of the individuals in the third *Hb* group have been affected by their *Bun* values. Overall, there is little reason to doubt the proportional hazards assumption.

4.4.2 Use of Schoenfeld residuals*

Hazards are said to be proportional if ratios of hazards are independent of time. If there are one or more explanatory variables in the model whose coefficients vary with time, or if there are explanatory variables that are time-dependent, the proportional hazards assumption will be violated. We therefore require a method that can be used to detect whether there is some form of time dependency in particular covariates, after allowing for the effects of explanatory variables that are known, or expected to be, independent of time.

The Schoenfeld residuals, defined in Section 4.1.5, are particularly useful in evaluating the assumption of proportional hazards after fitting a Cox regression model. Grambsch and Therneau (1994) have shown that the expected value of the ith scaled Schoenfeld residual, for the jth explanatory variable, X_j, in the model, r^*_{Pji}, is given by $\mathrm{E}(r^*_{Pji}) \approx \beta_j(t_i) - \hat{\beta}_j$, where $\beta_j(t)$ is taken to be a time-varying coefficient of X_j, $\beta_j(t_i)$ is the value of the coefficient at the ith death time, t_i, and $\hat{\beta}_j$ is the estimated value of β_j in the fitted Cox regression model. Consequently, a plot of the values of $r^*_{Pji} + \hat{\beta}_j$ against the death times should give information about the form of the time-dependent coefficient of X_j, $\beta_j(t)$. In particular, a horizontal line will suggest that the coefficient of X_j is constant, and the proportional hazards assumption is satisfied. A smoothed curve can be superimposed on this plot to aid interpretation, as in the plots of martingale residuals against the values of explanatory variables in Section 4.2.3. This plot can also be supplemented by fitting a straight line, and formally testing if the slope of this line is zero. However, this procedure has its limitations, since a slope that is not significantly different from zero may be found when there is, in fact, a non-linear relationship between the coefficient and time.

Example 4.10 Infection in patients on dialysis

The data on catheter removal times for patients on dialysis is now used to illustrate the use of the scaled Schoenfeld residuals in assessing non-proportional hazards. The scaled Schoenfeld residuals for the variables *Age* and *Sex* were given in Table 4.2. Adding the values of the estimated coefficients of these two variables, that is 0.030 and −2.711, respectively, to these two sets of residuals, and plotting their values against time, gives the graphs shown in Figures 4.16 and 4.17.

In neither plot is there any suggestion of non-proportional hazards. In fact, on fitting a straight line relationship between the values of $r^*_{Pji} + \hat{\beta}_j$ and time, using simple linear regression, the P-values for testing whether the estimated slope is significantly different from zero are 0.391 and 0.694 for *Age* and *Sex*, respectively.

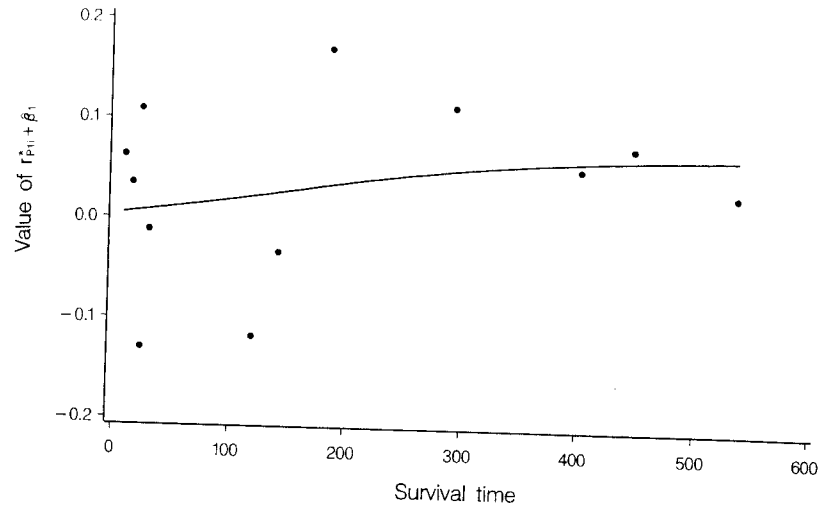

Figure 4.16 *Plot of values of $r^*_{P1i} + \hat{\beta}_1$ against time for Age with a smoothed curve superimposed.*

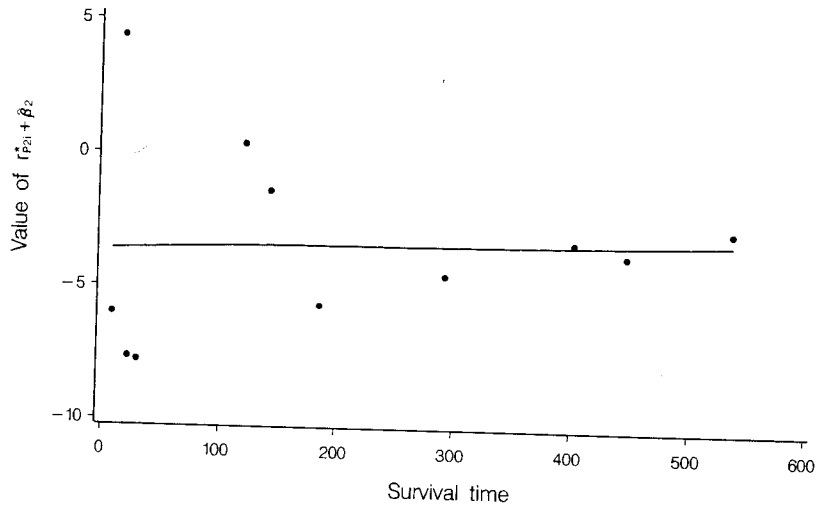

Figure 4.17 *Plot of values of $r^*_{P2i} + \hat{\beta}_2$ against time for Sex with a smoothed curve superimposed.*

4.4.3* *Adding a time-dependent variable*

To examine the assumption of proportional hazards in the Cox regression model, a *time-dependent variable* can be added to the model. Fuller details on the use of time-dependent variables in modelling survival data are given in Chapter 8, but in this section, the procedure is described in a particular context.

Consider a survival study in which each patient has been allocated to one of two groups, corresponding to a standard treatment and a new treatment. Interest may then centre on whether the ratio of the hazard of death at time t in one treatment group, relative to the other, is independent of survival time. A proportional hazards model for the hazard function of the ith individual in the study is then

$$h_i(t) = \exp(\beta_1 x_{1i}) h_0(t), \tag{4.17}$$

where x_{1i} is the value of an indicator variable X_1 that is zero for the standard treatment and unity for the new treatment. The relative hazard of death at any time for a patient on the new treatment, relative to one on the standard, is then e^{β_1}, which is independent of the survival time.

Now define a time-dependent explanatory variable X_2, where $X_2 = X_1 t$. If this variable is added to the model in equation (4.17), the hazard of death at time t for the ith individual becomes

$$h_i(t) = \exp(\beta_1 x_{1i} + \beta_2 x_{2i}) h_0(t), \tag{4.18}$$

where $x_{2i} = x_{1i} t$ is the value of $X_1 t$ for the ith individual. The relative hazard at time t is now

$$\exp(\beta_1 + \beta_2 t), \tag{4.19}$$

since $X_2 = t$ under the new treatment, and zero otherwise. This hazard ratio depends on t, and the model in equation (4.18) is no longer a proportional hazards model. In particular, if $\beta_2 < 0$, the relative hazard decreases with time. This means that the hazard of death on the new treatment, relative to that on the standard, decreases with time. If $\beta_1 < 0$, the interpretation of this would be that the superiority of the new treatment becomes more apparent as time goes on. On the other hand, if $\beta_2 > 0$, the relative hazard of death on the new treatment increases with time, reflecting an increasing risk of death on the new treatment relative to the standard. In the particular case where $\beta_2 = 0$, the relative hazard is constant at e^{β_1}. This means that a test of the hypothesis that $\beta_2 = 0$ is a test of the assumption of proportional hazards. The situation is illustrated in Figure 4.18.

In order to aid in both the computation and interpretation of the parameters in the model of equation (4.18), the variable X_2 can be defined in terms of the deviation from some time, t_0. The estimated values of β_1 and β_2 will then tend to be less highly correlated, and maximisation of the appropriate likelihood function will be less difficult. If X_2 is taken to be such that $X_2 = X_1(t - t_0)$, the value of X_2 is $t - t_0$ for the new treatment and zero for the standard. The

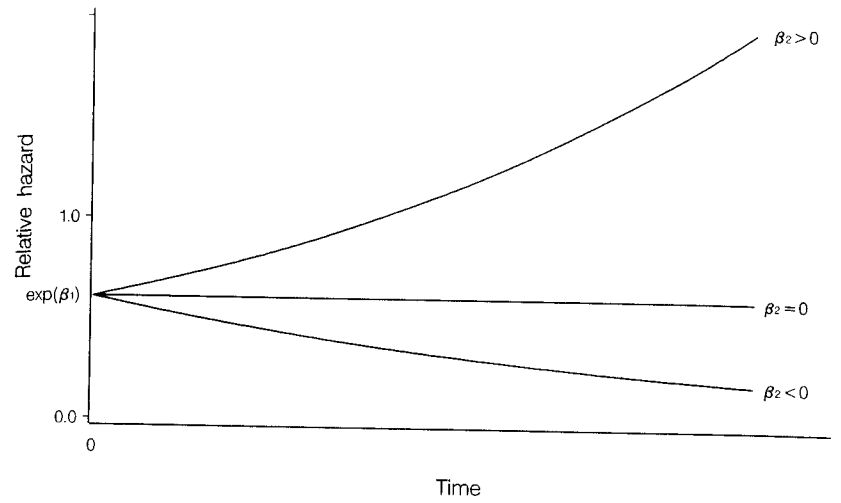

Figure 4.18 *Plot of the relative hazard,* $\exp\{\beta_1 + \beta_2 t\}$*, against* t*, for different values of* β_2*.*

relative hazard now becomes

$$\exp\{\beta_1 + \beta_2(t - t_0)\}.$$

In the model of equation (4.18), the quantity e^{β_1} is the hazard of death at time t_0 for an individual on the new treatment relative to one on the standard. In practical applications, t_0 will generally be chosen to provide a convenient interpretation for the time at which this relative hazard is applicable. For example, taking t_0 to be the mean or median survival time means that $\exp(\hat{\beta}_1)$ is the estimated relative hazard of death at this time.

A similar model can be used to detect whether the coefficient of a continuous variate has a coefficient that depends on time. Suppose that X is such a variate, and we wish to examine whether there is any evidence that the coefficient of X is linearly dependent on time. To do this, the term Xt is added to the model that includes X. The hazard of death at time t for the ith individual is then

$$h_i(t) = \exp(\beta_1 x_i + \beta_2 x_i t) h_0(t),$$

where x_i is the value of X for that individual. The hazard of death at time t for an individual for whom $X = x_i + 1$, relative to an individual for whom $X = x_i$, is then $\exp(\beta_1 + \beta_2 t)$, as in equation (4.19).

The time-dependent variables considered in this section are such that their coefficients are linearly dependent on time. A similar approach can be used when a coefficient that is a non-linear function of time is anticipated. For example, $\log t$ might be used in place of t in the definition of the time-dependent variable X_2, used in equation (4.18). In this version of the model, a test of the hypothesis that $\beta_2 = 0$ is a test of proportional hazards, where the alternative

hypothesis is that the hazard ratio is dependent on the logarithm of time. Using $\log t$ in the definition of a time-dependent variable is also helpful when the numerical values of the survival times are large, such as when survival in a long-term study is measured in days. There may then be computational problems associated with calculating the value of $\exp(\beta_2 x_{2i})$ in equation (4.18), which are resolved by using $\log t$ in place of t in the definition of X_2.

Models that include the time-dependent variable X_2 cannot be fitted by treating X_2 in the same manner as other explanatory variables in the model. The reason for this is that this variable will have different values at different death times, complicating the calculation of the denominator of the partial likelihood function in equation (3.4). Full details on the fitting process will be deferred to Chapter 8. However, inferences about the effect of time-dependent variables on the hazard function can be evaluated as for other variables. In particular, the change in the value of the $-2\log\hat{L}$ statistic can be compared to percentage points of the chi-squared distribution to test the significance of the variable. This is therefore a formal test of proportional hazards.

Example 4.11 Infection in patients on dialysis
An informal assessment of non-proportional hazards in respect of the variables *Age* and *Sex* was given in Example 4.10. We now add variables whose coefficients are linear functions of time in order to provide a formal test of the proportional hazards assumption.

We begin by fitting the Cox regression model containing just *Age* and *Sex*, which leads to a value of $-2\log\hat{L}$ of 34.468. We now define terms that are the products of these variables with time, namely $Tage = Age \times t$ and $Tsex = Sex \times t$. These variables are then added to the model. Note that we cannot simply form these products from the observed survival times of the patients, since the model-fitting process requires that these values be computed for different values of t; see Chapter 8 for details on this.

When the variable *Tage* is added to the model that contains *Age* and *Sex*, the value of $-2\log\hat{L}$ reduces to 32.006, but this reduction is not significant at the 5% level ($P = 0.117$). The reduction in $-2\log\hat{L}$ when *Tsex* is added to the model that has *Age* and *Sex* is only 0.364 ($P = 0.546$). This analysis confirms that there is no reason to doubt the assumption of proportional hazards in respect of the variables *Age* and *Sex*.

4.5 Recommendations

In this chapter, a range of diagnostics have been presented. Which should be used on a routine basis and which are needed when a more thorough assessment of model adequacy is required?

In terms of assessing the overall fit of a model, a plot of the deviance residuals against the risk score gives information on observations that are not well fitted by the model, and their relation to the set of values of the explanatory variables. This diagnostic is generally more informative than the cumulative, or log-cumulative, hazard plot of the Cox-Snell residuals. Plots of residuals against the survival times, the rank order of the survival times, or explanatory variables may also be useful.

Plots of residuals might be supplemented by influence diagnostics. When the inference to be drawn from a model centres on one or two particular parameters, the delta-beta statistic for those parameters, will be the most relevant. Plots of these values against the rank order of survival times will then be useful. To investigate whether there are observations that have an influence on the set of parameter estimates, or risk score, the diagnostic based on the absolute values of the elements of $\boldsymbol{l}_{\max}$ is probably the most suitable. Plots of these values against the rank order of survival times will be informative, but plots against particular explanatory variables might also be revealing. An initial assessment of the validity of the proportional hazards assumption can be made from log-cumulative hazard plots. However, plots based on the scaled Schoenfeld residuals can be more helpful. Formal tests of the assumption of proportional hazards may be based on time-dependent variables.

4.6 Further reading

General introductions to the ideas of model checking in linear models are included in Draper and Smith (1998) and Montgomery *et al.* (2001). Cook and Weisberg (1982) give a more detailed account of the theory underlying residuals and influence diagnostics in a number of situations. Atkinson (1985) describes model checking in linear models from a practical viewpoint, and McCullagh and Nelder (1989) and Aitkin *et al.* (1989) discuss this topic in the context of generalised linear models.

Many textbooks devoted to the analysis of survival data, and particularly those of Cox and Oakes (1984), Hosmer and Lemeshow (1999), Lawless (2002), and Kalbfleisch and Prentice (2002), include sections on the use of residuals. Hinkley *et al.* (1991) and Hastie and Tibshirani (1990) also include brief discussions on methods for assessing the adequacy of models fitted to survival data.

Early articles on the use of residuals in checking the adequacy of survival models include Kay (1977) and Crowley and Hu (1977). These papers include a discussion on the Cox-Snell residuals, which are based on the general definition of residuals given by Cox and Snell (1968). Crowley and Storer (1983) showed empirically that the cumulative hazard plot of the residuals is not particularly good at identifying inadequacies in the fitted model. See also Crowley and Storer (1983) for a practical application of the methods. Reviews of diagnostic procedures in survival analysis were given in the mid-1980s by Kay (1984) and Day (1985).

Martingale residuals were proposed by Barlow and Prentice (1988). Essentially the same residuals were proposed by Lagakos (1981) and their use is discussed by Therneau, Grambsch and Fleming (1990) and Henderson and Milner (1991). Deviance residuals were also introduced in Therneau, Grambsch and Fleming (1990). The Schoenfeld residuals for the Cox model were proposed by Schoenfeld (1982). In accounts of survival analysis based on the

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).

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

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