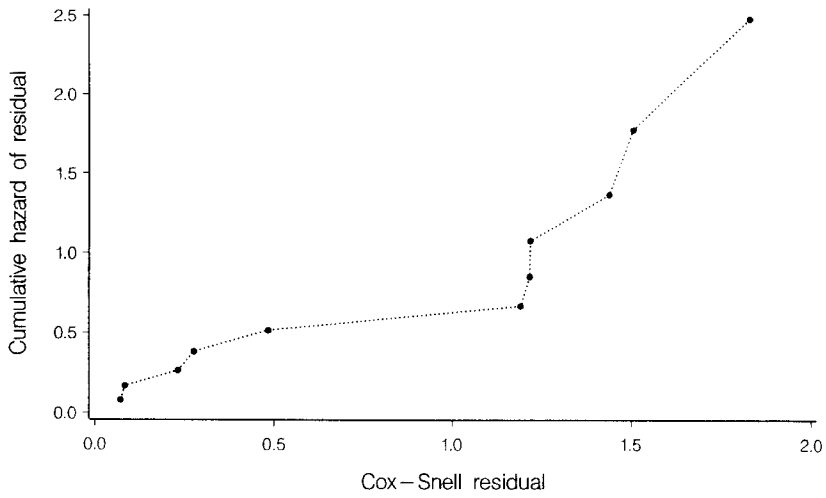

Figure 4.1 *Cumulative hazard plot of the Cox-Snell residuals.*

distribution of Cox-Snell residuals for $n = 3$ was shown by Lagakos (1981) to be quite dissimilar to a unit exponential sample.

On other occasions, a straight line plot may be obtained when the model fitted is known to be incorrect. Indeed, practical experience suggests that a fitted model has to be seriously wrong before anything other than a straight line of unit slope is seen in the cumulative hazard plot of the Cox-Snell residuals.

In the particular case of the null model, that is, the model that contains no explanatory variates, the cumulative hazard plot will be a straight line with unit slope and zero intercept, even if some explanatory variables should actually be included in the model. The reason for this is that when no covariates are included, the Cox-Snell residual for the ith individual reduces to $-\log \hat{S}_0(t_i)$. From equation (3.26) in Chapter 3, in the absence of ties this is approximately $\sum_{j=1}^{k} 1/n_j$ at the kth uncensored survival time, $k = 1, 2, \ldots, r-1$, where n_j is the number at risk at time t_j. This summation is simply $\sum_{j=1}^{k} 1/(n-j+1)$, which is the expected value of the kth order statistic in a sample of size n from a unit exponential distribution.

In view of the limitations of the Cox-Snell residuals in assessing model adequacy, diagnostic procedures based on other types of residuals, that are of practical use, are described in the following section.

4.2.2 Plots based on the martingale and deviance residuals

The martingale residuals, introduced in Section 4.1.3, can be interpreted as the difference between the observed and expected number of deaths in the time interval $(0, t_i)$, for the ith individual. Accordingly, these residuals highlight individuals who, on the basis of the assumed model, have died too soon

or lived too long. Large negative residuals will correspond to individuals who have a long survival time, but covariate values that suggest they should have died earlier. On the other hand, a residual close to unity, the upper limit of a martingale residual, will be obtained when an individual has an unexpectedly short survival time. An index plot of the martingale residuals will highlight individuals whose survival time is not well fitted by the model. Such observations may be termed *outliers*. The data from individuals for whom the residual is unusually large in absolute value, will need to be the subject of further scrutiny. Plots of these residuals against the survival time, the rank order of the survival times, or explanatory variables, may indicate whether there are particular times, or values of the explanatory variables, where the model does not fit well.

Since the deviance residuals are more symmetrically distributed than the martingale residuals, plots based on these residuals tend to be easier to interpret. Consequently, an index plot of the deviance residuals may also be used to identify individuals whose survival times are out of line.

In a fitted Cox regression model, the hazard of death for the ith individual at any time depends on the values of explanatory variables for that individual, $\boldsymbol{x}_i$, through the function $\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_i)$. This means that individuals for whom $\hat{\boldsymbol{\beta}}'\boldsymbol{x}_i$ has large negative values have a lower than average risk of death, and individuals for whom $\hat{\boldsymbol{\beta}}'\boldsymbol{x}_i$ has a large positive value have a higher than average risk. The quantity $\hat{\boldsymbol{\beta}}'\boldsymbol{x}_i$ is the risk score, introduced in Section 3.1 of Chapter 3, and provides information about whether an individual might be expected to survive for a short or long time. By reconciling information about individuals whose survival times are out of line, with the values of their risk score, useful information can be obtained about the characteristics of observations that are not well fitted by the model. In this context, a plot of the deviance residuals against the risk score is a particulary helpful diagnostic.

Example 4.3 Infection in patients on dialysis

Consider again the data on times to infection in kidney patients. From the values of the martingale and deviance residuals given in Table 4.2, we see that patient 2 has the largest positive residual, suggesting that the time to removal of the catheter is shorter for this patient than might have been expected on the basis of the fitted model. The table also shows that the two types of residual do not rank the observations in the same order. For example, the second largest negative martingale residual is found for patient 12, whereas patient 6 has the second largest negative deviance residual. However, the observations that have the most extreme values of the martingale and deviance residuals will tend to be the same, as in this example. Index plots of the martingale and deviance residuals are shown in Figure 4.2.

The plots are quite similar, but the distribution of the deviance residuals is seen to be more symmetric. The plots also show that there are no patients that have residuals that are unusually large in absolute value. Figure 4.3 gives

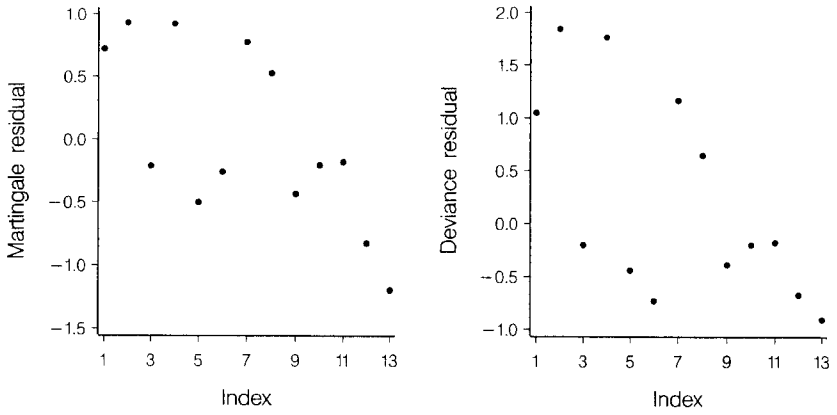

Figure 4.2 *Index plots of the martingale and deviance residuals.*

a plot of the deviance residuals against the risk scores, that are found from the values of $0.030\, Age_i - 2.711\, Sex_i$, for $i = 1, 2, \ldots, 13$.

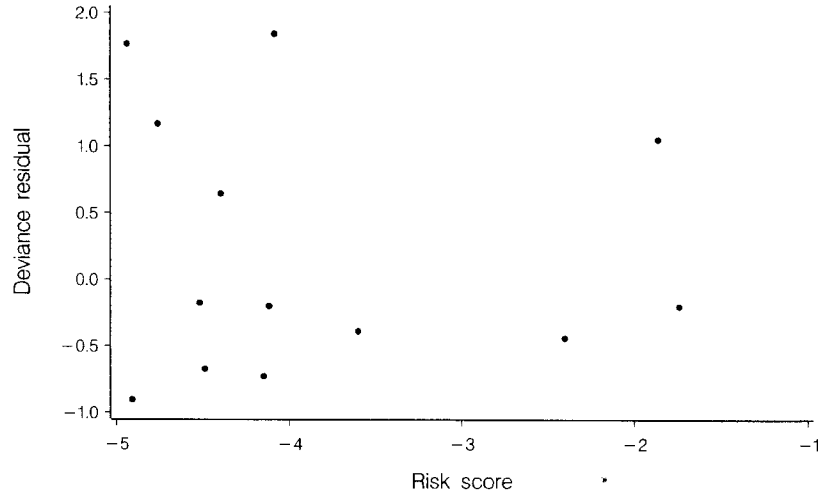

Figure 4.3 *Plot of the deviance residuals against the values of the risk score.*

This figure shows that patients with the largest deviance residuals have low risk scores. This indicates that these patients are at relatively low risk of an early catheter removal, and yet their removal time is sooner than expected.

4.2.3 Checking the functional form of covariates

Although the model-based approach to the analysis of survival data, described in Chapter 3, identifies a particular set of covariates on which the hazard function depends, it will be important to check that the correct functional form has been adopted for these variables. An improvement in the fit of a model may well be obtained by using some transformation of the values of a variable instead of the original values. For example, it might the that a better fitting model is obtained by using a non-linear function of the age of an individual at baseline, or the logarithm of a biochemical variable such as serum bilirubin level. Similarly, an explanatory variable such as serum cholesterol level may only begin to exert an effect on survival when it exceeds some threshold value, after which time the hazard of death might increase with increasing values of that variable.

A straightforward means of assessing this aspect of model adequacy is based on the martingale residuals obtained from fitting the null model, that is, the model that contains no covariates. These residuals are then plotted against the values of each covariate in the model. It has been shown by Therneau *et al.* (1990) that this plot should display the functional form required for the covariate. In particular, a straight line plot indicates that a linear term is needed.

As an extension to this approach, if the functional form of certain covariates can be assumed to be known, martingale residuals may be calculated from the fitted Cox regression model that contains these covariates alone. The resulting martingale residuals are then plotted against the covariates whose functional form needs to be determined.

The graphs obtained in this way are usually quite "noisy" and their interpretation is much helped by superimposing a smoothed curve that is fitted to the scatterplot. There are a number of such *smoothers* that can be obtained, including *smoothing splines*, but the one that is most commonly used is the *LOWESS* or *LOESS* smoother, proposed by Cleveland (1979). This algorithm is implemented in many software packages.

Even with a smoother, it can be difficult to discern a specific functional form when a non-linear pattern is seen in the plot. If a specific transformation is suggested, such as the logarithmic transformation, the covariate can be so transformed, and the martingale residuals for the null model plotted against the transformed variate. A straight line would then confirm that an appropriate transformation has been used.

Example 4.4 Infection in patients on dialysis

In this example, we illustrate the use of martingale residuals in assessing whether the age effect is linear in the Cox regression model fitted to the data of Example 4.1. First, the martingale residuals for the null model are obtained and these are plotted against the corresponding values of the age of a patient in Figure 4.4.

There is too little data to say much about this graph, but the smoothed

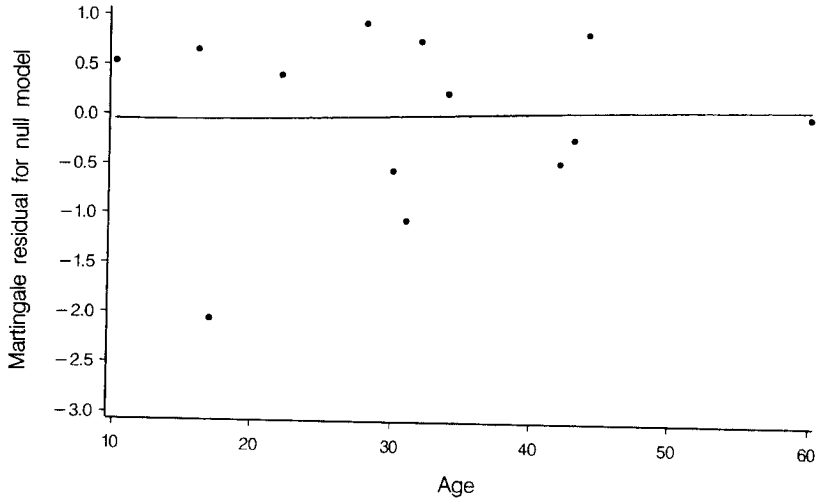

Figure 4.4 *Plot of the martingale residuals for the null model against Age, with a smoothed curve superimposed.*

curve indicates that there is no need for anything other than a linear term in *Age*. In fact, the age effect is not actually significant, and so it is not surprising that the smoothed curve is roughly horizontal.

We end this section with a further illustrative example.

Example 4.5 Survival of multiple myeloma patients
In this example we return to the data on the survival times of 48 patients with multiple myeloma, described in Example 1.3. In Example 3.5, a Cox regression model that contained the explanatory variables *Hb* (serum haemoglobin) and *Bun* (blood urea nitrogen) was found to be a suitable model for the hazard function. We now perform an analysis of the residuals in order to study the adequacy of this fitted model.

First, a cumulative hazard plot of the Cox-Snell residuals is shown in Figure 4.5. The line made by the plotted points in Figure 4.5 is reasonably straight, and has a unit slope and zero intercept. On the basis of this plot, there is no reason to doubt the adequacy of the fitted model. However, as pointed out in Section 4.2.1, this plot is not at all sensitive to departures from the fitted model.

To further assess the fit of the model, the deviance residuals are plotted against the corresponding risk scores in Figure 4.6. This plot shows that patients 41 and 38 have the largest values of the deviance residuals, but these are not much separated from values of the residuals for some of the other patients. Patients with the three largest risk scores have residuals that are close to zero, suggesting that these observations are well fitted by the model. Again, there is no reason to doubt the validity of the fitted model.

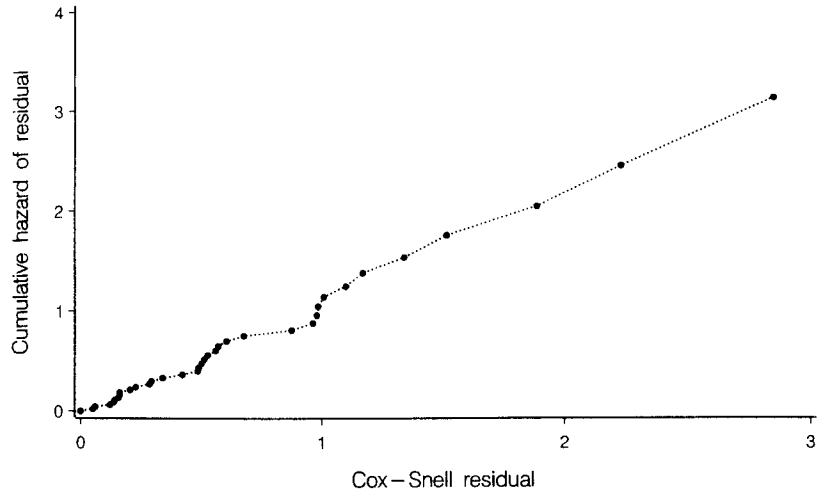

Figure 4.5 *Log-cumulative hazard plot of the Cox-Snell residuals.*

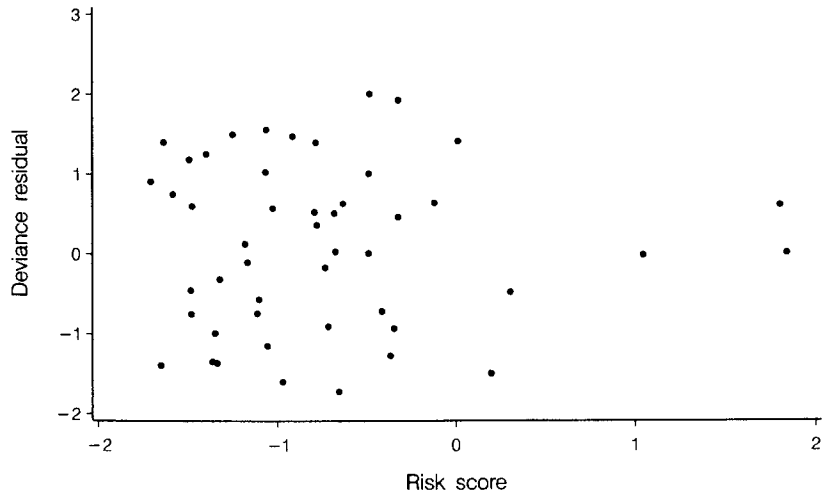

Figure 4.6 *Deviance residuals plotted against the risk score.*

In order to investigate whether the correct functional form for the variates *Hb* and *Bun* has been used, martingale residuals are calculated for the null model and plotted against the values of these variables. The resulting plots, with a smoothed curve superimposed to aid in their interpretation, are shown in Figures 4.7 and 4.8.

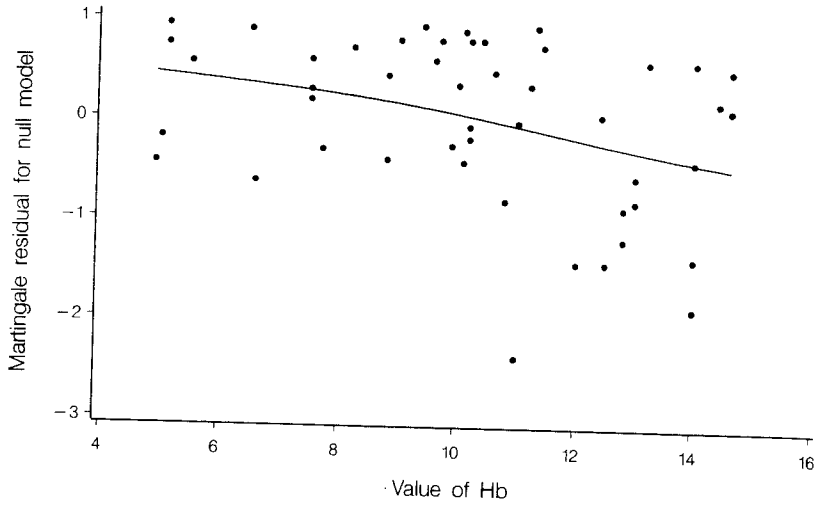

Figure 4.7 *Plot of the martingale residuals for the null model against the values of Hb, with a smoothed curve superimposed.*

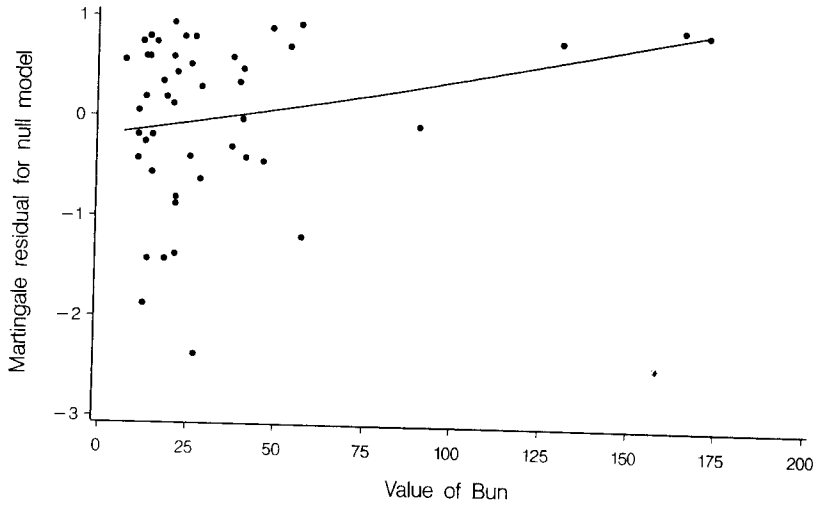

Figure 4.8 *Plot of the martingale residuals for the null model against the values of Bun, with a smoothed curve superimposed.*

The plots for *Hb* and *Bun* confirm that linear terms in each variable are required in the model. Note that the slope of the plot for *Hb* in Figure 4.7 is negative, corresponding to the negative coefficient of *Hb* in the fitted model, while the plot for *Bun* in Figure 4.8 has a positive slope.

In this data set, the values of *Bun* range from 6 to 172, and the distribution of their values across the 48 subjects is positively skewed. In order to guard against the extreme values of this variate having an undue impact on the coefficient of *Bun*, logarithms of this variable might be used in the modelling process. Although there is no suggestion of this in Figure 4.8, for illustrative purposes, we will use this type of plot to investigate whether a model containing log *Bun* rather than *Bun* is acceptable. Figure 4.9 shows the martingale residuals for the null model plotted against the values of log *Bun*.

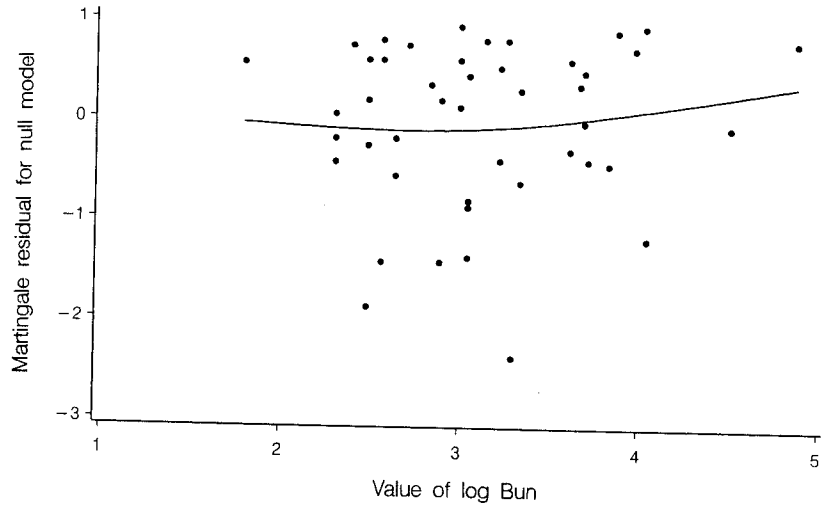

Figure 4.9 *Plot of the martingale residuals for the null model against the values of* log *Bun, with a smoothed curve superimposed.*

The smoothed curve in this figure does suggest that it is not appropriate to use a linear term in log *Bun*. Indeed, if it were decided to use log *Bun* in the model, Figure 4.9 indicates that a quadratic term in log *Bun* may be needed. In fact, adding this quadratic term to a model that includes *Hb* and log *Bun* leads to a significant reduction in the value of $-2 \log \hat{L}$, but the resulting value of this statistic, 201.458, is then only slightly less than the corresponding value for the model containing *Hb* and *Bun*, which is 202.938. This analysis confirms that the model should contain linear terms in the variables *Hb* and *Bun*.

4.3 Identification of influential observations

In the assessment of model adequacy, it is important to determine whether any particular observation has an undue impact on inferences made on the

basis of a model fitted to an observed set of survival data. Observations that do have an effect on model-based inferences are said to be *influential*.

As an example, consider a survival study in which a new treatment is to be compared with a standard. In such a comparison, it would be important to determine if the hazard of death on the new treatment, relative to that on the standard, was substantially affected by any one individual. In particular, it might be that when the data record for one individual is removed from the data base, the relative hazard is increased or reduced by a substantial amount. If this happens, the data from such an individual would need to be subject to particular scrutiny.

Conclusions from a survival analysis are often framed in terms of estimates of quantities such as the relative hazard and median survival time, which depend on the estimated values of the β-parameters in the fitted Cox regression model. It is therefore of particular interest to examine the influence of each observation on these estimates. We can do this by examining the extent to which the estimated parameters in the fitted model are affected by omitting in turn the data record for each individual in the study. In some circumstances, the estimates of a subset of the parameters may be of special importance, such as parameters associated with treatment effects. The study of influence may then be limited to just these parameters. On many occasions, the influence that each observation has on the estimated hazard function will be of interest, and it would then be important to identify observations that influence the complete set of parameter estimates under the model. These two aspects of influence are discussed in the following sections.

In contrast to models encountered in the analysis of other types of data, such as the general linear model, the effect of removing one observation from a set of survival data is not easy to study. This is mainly because the log-likelihood function for the Cox regression model cannot be expressed as the sum of a number of terms, in which each term is the contribution to the log-likelihood made by each observation. Instead, the removal of one observation affects the risk sets over which quantities of the form $\exp(\boldsymbol{\beta}'\boldsymbol{x})$ are summed. This means that influence diagnostics are quite difficult to derive and so the following sections of this chapter simply give the relevant results. References to the articles that contain derivations of the quoted formulae are included in the final section of this chapter.

4.3.1 Influence of observations on a parameter estimate

Suppose that we wish to determine whether any particular observation has an untoward effect on $\hat{\beta}_j$, the jth parameter estimate, $j = 1, 2, \ldots, p$, in a fitted Cox regression model. One way of doing this would be to fit the model to all n observations in the data set, and to then fit the same model to the sets of $n - 1$ observations obtained by omitting each of the n observations in turn. The actual effect that omitting each observation has on the parameter estimate could then be determined. This procedure is computationally expensive, unless the number of observations is not too large, and so we use

instead an approximation to the amount by which $\hat{\beta}_j$ changes when the ith observation is omitted, for $i = 1, 2, \ldots, n$. Suppose that the value of the jth parameter estimate on omitting the ith observation is denoted by $\hat{\beta}_{j(i)}$. Cain and Lange (1984) showed that an approximation to $\hat{\beta}_j - \hat{\beta}_{j(i)}$ is based on the score residuals, described in Section 4.1.6.

Let $\boldsymbol{r}_{Si}$ denote the vector of values of the score residuals for the ith observation, so that $\boldsymbol{r}'_{Si} = (r_{S1i}, r_{S2i}, \ldots, r_{Spi})$, where r_{Sji}, $j = 1, 2, \ldots, p$, is the ith score residual given in equation (4.13). An approximation to $\hat{\beta}_j - \hat{\beta}_{j(i)}$, the change in $\hat{\beta}_j$ on omitting the ith observation, is then the jth component of the vector

$$\boldsymbol{r}'_{Si} \operatorname{var}(\hat{\boldsymbol{\beta}}),$$

$\operatorname{var}(\hat{\boldsymbol{\beta}})$ being the variance-covariance matrix of the vector of parameter estimates in the fitted Cox regression model. The jth element of this vector, which is called a *delta-beta*, will be denoted by $\Delta_i\hat{\beta}_j$, so that $\Delta_i\hat{\beta}_j \approx \hat{\beta}_j - \hat{\beta}_{j(i)}$. Use of this approximation means that the values of $\Delta_i\hat{\beta}_j$ can be computed from quantities available after fitting the model to the full data set.

Observations that influence a particular parameter estimate, the jth say, will be such that the values of $\Delta_i\hat{\beta}_j$, the delta-betas for these observations, are larger in absolute value than for other observations in the data set. Index plots of the delta-betas for each explanatory variable in the model will then reveal whether there are observations that have an undue impact on the parameter estimate for any particular explanatory variable. In addition, a plot of the values of $\Delta_i\hat{\beta}_j$ against the rank order of the survival times yields information about the relation between survival time and influence.

The delta-betas may be standardised by dividing $\Delta_i\hat{\beta}_j$ by the standard error of $\hat{\beta}_j$ to give a *standardised delta-beta*. The standardised delta-beta can be interpreted as the change in the value of the statistic $\hat{\beta}/\operatorname{se}(\hat{\beta})$, on omitting the ith observation. Since this statistic can be used in assessing whether a particular parameter has a value significantly different from zero (see Section 3.4 of Chapter 3), the standardised delta-beta can be used to provide information on how the significance of the parameter estimate is affected by the removal of the ith observation from the data base. Again, an index plot is the most useful way of displaying the standardised delta-betas.

The statistic $\Delta_i\hat{\beta}_j$ is an approximation to the actual change in the parameter estimate when the ith observation is omitted from the fit. The approximation is generally adequate in the sense that observations that have an influence on a parameter estimate will be highlighted. However, the actual effect of omitting any particular observation on model-based inferences will need to be studied. The agreement between the actual and approximate delta-betas in a particular situation is illustrated in Example 4.6.

Example 4.6 Infection in patients on dialysis

In this example, we return to the data on the times to infection following commencement of dialysis. To investigate the influence that the data from each of the 13 patients in the study has on the estimated value of the coefficients

of the variables *Age* and *Sex* in the linear component of the fitted Cox regression model, the approximate unstandardised delta-betas, $\Delta_i\beta_1$ and $\Delta_i\beta_2$, are obtained. These are given in Table 4.4.

Table 4.4 *Approximate delta-betas for Age* ($\hat{\beta}_1$)*, and Sex* ($\hat{\beta}_2$).

Observation	$\Delta_i\hat{\beta}_1$	$\Delta_i\hat{\beta}_2$
1	0.0020	−0.1977
2	0.0004	0.5433
3	−0.0011	0.0741
4	−0.0119	0.5943
5	0.0049	0.0139
6	−0.0005	−0.1192
7	−0.0095	0.1270
8	−0.0032	−0.0346
9	−0.0073	−0.0734
10	0.0032	−0.2023
11	0.0060	−0.2158
12	0.0048	−0.1939
13	0.0122	−0.3157

The largest delta-beta for *Age* occurs for patient number 13, but there are other delta-betas with similar values. The actual change in the parameter estimate on omitting the data for this patient is 0.0195, and so omission of this observation increases the hazard of infection relative to the baseline hazard. The standard error of the parameter estimate for *Age* in the full data set is 0.026, and so the maximum amount by which this estimate is changed when one observation is deleted is about three-quarters of a standard error. When the data from patient 13 is omitted, the age effect becomes slightly more significant, but the difference is unlikely to be of practical importance.

There are two large delta-betas for *Sex* that are quite close to one another. These correspond to the observations from patients 2 and 4. The actual change in the parameter estimate when each observation is omitted in turn is 0.820 and 0.818, and so the approximate delta-betas underestimate the actual change. The standard error of the estimated coefficient of *Sex* in the full data set is 1.096, and so again the change in the estimate on deleting an observation is less than one standard error. The effect of deleting either of these two observations is to increase the relative hazard, but again this increase is not great.

To compare the approximate delta-betas with the actual values, a plot of their values against the rank of the time to infection is given in Figures 4.10 and 4.11. These figures show that the agreement is generally quite good, although there is a tendency for the actual changes in the parameter estimates to be underestimated by the approximation. The largest difference between the actual and approximate value of the delta-beta for *Age* is 0.010, which

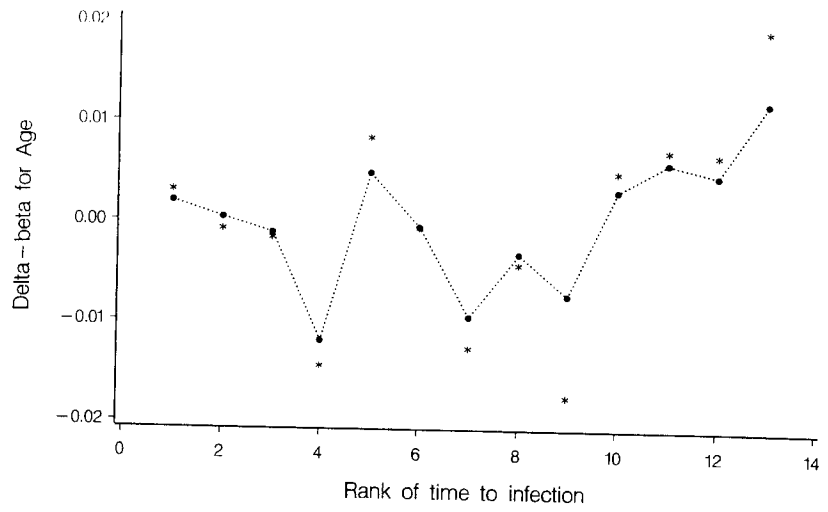

Figure 4.10 *Plot of the exact (∗) and approximate (•) delta-betas for Age against rank order of time to infection.*

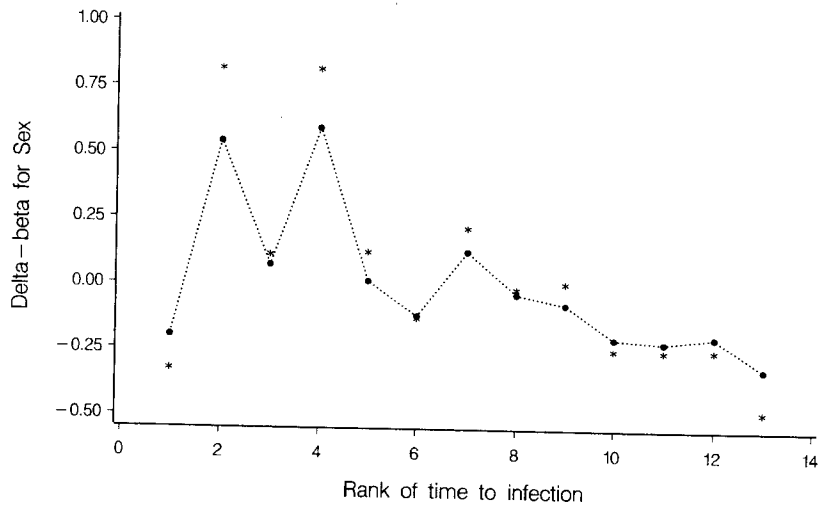

Figure 4.11 *Plot of the exact (∗) and approximate (•) delta-betas for Sex against rank order of time to infection.*