CHAPTER 4

Model checking in the Cox regression model

After a model has been fitted to an observed set of survival data, the adequacy of the fitted model needs to be assessed. Indeed, the use of diagnostic procedures for model checking is an essential part of the modelling process.

In some situations, careful inspection of an observed set of data may leads to the identification of certain features, such as individuals with unusually large or small survival times. However, unless there are only one or two explanatory variables, a visual examination of the data may not be very revealing. The situation is further complicated by censoring, in that the occurrence of censored survival times make it difficult to judge aspects of model adequacy, even in the simplest of situations. Visual inspection of the data has therefore to be supplemented by diagnostic procedures for detecting inadequacies in a fitted model. Because methods used in assessing the adequacy of survival models have to cope with the occurrence of censored survival times, they are a little more complicated than the corresponding methods used in linear regression modelling. However, many of the procedures are easily carried out using computer software for survival analysis.

Once a model has been fitted, there are a number of aspects of the fit of a model that need to be studied. For example, the model must include an appropriate set of explanatory variables from those measured in the study, and we will need to check that the correct functional form of these variables has been used. It might be important to identify observed survival times that are greater than would have been anticipated, or individuals whose explanatory variables have an undue impact on particular hazard ratios. Also, some means of checking the assumption of proportional hazards might be required.

Many model-checking procedures are based on quantities known as *residuals*. These are values that can be calculated for each individual in the study, and have the feature that their behaviour is known, at least approximately, when the fitted model is satisfactory. A number of residuals have been proposed for use in connection with the Cox regression model, and this chapter begins with a review of some of these. The use of residuals in assessing specific aspects of model adequacy is then discussed in subsequent sections.

4.1 Residuals for the Cox regression model

Throughout this section, we will suppose that the survival times of n individuals are available, where r of these are death times and the remaining $n - r$

are right-censored. We further suppose that a Cox regression model has been fitted to the survival times, and that the linear component of the model contains p explanatory variables, $X_1, X_2, \ldots, X_p$. The fitted hazard function for the ith individual, $i = 1, 2, \ldots, n$, is therefore

$$\hat{h}_i(t) = \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i)\hat{h}_0(t),$$

where $\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_p x_{pi}$ is the value of the fitted component, or linear predictor, of the model for that individual and $\hat{h}_0(t)$ is the estimated baseline hazard function.

4.1.1 Cox-Snell residuals

The residual that is most widely used in the analysis of survival data is the *Cox-Snell residual*, so called because it is a particular example of the general definition of residuals given by Cox and Snell (1968).

The Cox-Snell residual for the ith individual, $i = 1, 2, \ldots, n$, is given by

$$r_{Ci} = \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i)\hat{H}_0(t_i), \tag{4.1}$$

where $\hat{H}_0(t_i)$ is an estimate of the baseline cumulative hazard function at time t_i, the observed survival time of that individual. In practice, the Nelson-Aalen estimate given in equation (3.25) is generally used. Note that from equation (3.21), the Cox-Snell residual, r_{Ci}, is the value of $\hat{H}_i(t_i) = -\log \hat{S}_i(t_i)$, where $\hat{H}_i(t_i)$ and $\hat{S}_i(t_i)$ are the estimated values of the cumulative hazard and survivor functions of the ith individual at t_i.

This residual can be derived from a general result in mathematical statistics on the distribution of a function of a random variable. According to this result, if T is the random variable associated with the survival time of an individual, and $S(t)$ is the corresponding survivor function, then the random variable $Y = -\log S(T)$ has an exponential distribution with unit mean, irrespective of the form of $S(t)$. The proof of this result is outlined in the following paragraph, which can be omitted without loss of continuity.

According to a general result, if $f_X(x)$ is the probability density function of the random variable X, the density of the random variable $Y = g(X)$ is given by

$$f_Y(y) = f_X\{g^{-1}(y)\} \Big/ \left|\frac{\mathrm{d}y}{\mathrm{d}x}\right|,$$

where $f_X\{g^{-1}(y)\}$ is the density of X expressed in terms of y. Using this result, the probability density function of the random variable $Y = -\log S(T)$ is given by

$$f_Y(y) = f_T\left\{S^{-1}(e^{-y})\right\} \Big/ \left|\frac{\mathrm{d}y}{\mathrm{d}t}\right|, \tag{4.2}$$

where $f_T(t)$ is the probability density function of T. Now,

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \frac{\mathrm{d}\{-\log S(t)\}}{\mathrm{d}t} = \frac{f_T(t)}{S(t)},$$

and when the absolute value of this function is expressed in terms of y, the derivative becomes

$$\frac{f_T\left\{S^{-1}(e^{-y})\right\}}{S\left\{S^{-1}(e^{-y})\right\}} = \frac{f_T\left\{S^{-1}(e^{-y})\right\}}{e^{-y}}.$$

Finally, on substituting for the derivative in equation (4.2), we find that

$$f_Y(y) = e^{-y},$$

which, from equation (5.6), is the probability density function of an exponential random variable with unit mean.

The next and crucial step in the argument is as follows. If the model fitted to the observed data is satisfactory, then a model-based estimate of the survivor function for the ith individual at t_i, the survival time of that individual, will be close to the corresponding true value $S_i(t_i)$. This suggests that if the correct model has been fitted, the values $\hat{S}_i(t_i)$ will have properties similar to those of $S_i(t_i)$. Then, the negative logarithms of the estimated survivor functions, $-\log \hat{S}_i(t_i)$, $i = 1, 2, \ldots, n$, will behave as n observations from a unit exponential distribution. These estimates are the Cox-Snell residuals.

If the observed survival time for an individual is right-censored, then the corresponding value of the residual is also right-censored. The residuals will therefore be a censored sample from the unit exponential distribution, and a test of this assumption provides a test of model adequacy, to which we return in Section 4.2.1.

The Cox-Snell residuals, r_{Ci}, have properties that are quite dissimilar to those of residuals used in linear regression analysis, for example. In particular, they will not be symmetrically distributed about zero, and in fact they cannot be negative. Furthermore, since the Cox-Snell residuals are assumed to have an exponential distribution when an appropriate model has been fitted, they have a highly skew distribution and the mean and variance of the ith residual will both be unity.

4.1.2 Modified Cox-Snell residuals

Censored observations lead to residuals that cannot be regarded on the same footing as residuals derived from uncensored observations. We might therefore seek to modify the Cox-Snell residuals so that explicit account can be taken of censoring.

Suppose that the ith survival time is a censored observation, t_i^*, and let t_i be the actual, but unknown, survival time, so that $t_i > t_i^*$. The Cox-Snell residual for this individual, evaluated at the censored survival time, is then given by

$$r_{Ci} = \hat{H}_i(t_i^*) = -\log \hat{S}_i(t_i^*),$$

where $\hat{H}_i(t_i^*)$ and $\hat{S}_i(t_i^*)$ are the estimated cumulative hazard and survivor functions, respectively, for the ith individual at the censored survival time.

If the fitted model is correct, then the values r_{Ci} can be taken to have a unit exponential distribution. The cumulative hazard function of this distribution

increases linearly with time, and so the greater the value of the survival time t_i for the ith individual, the greater the value of the Cox-Snell residual for that individual. It then follows that the residual for the ith individual at the actual (unknown) failure time, $\hat{H}_i(t_i)$, will be greater than the residual evaluated at the observed censored survival time.

To take account of this, Cox-Snell residuals can be modified by the addition of a positive constant Δ, which can be called the *excess residual.* Modified Cox-Snell residuals are therefore of the form

$$r'_{Ci} = \begin{cases} r_{Ci} & \text{for uncensored observations,} \\ r_{Ci} + \Delta & \text{for censored observations,} \end{cases}$$

where r_{Ci} is the Cox-Snell residual for the ith observation, defined in equation (4.1). It now remains to identify a suitable value for Δ. For this, we use the *lack of memory property* of the exponential distribution.

To demonstrate this property, suppose that the random variable T has an exponential distribution with mean λ^{-1}, and consider the probability that T exceeds $t_0 + t_1$, $t_1 \geqslant 0$, conditional on T being at least equal to t_0. From the standard result for conditional probability given in Section 3.3.1, this probability is

$$\mathrm{P}(T \geqslant t_0 + t_1 \mid T \geqslant t_0) = \frac{\mathrm{P}(T \geqslant t_0 + t_1 \text{ and } T \geqslant t_0)}{\mathrm{P}(T \geqslant t_0)}.$$

The numerator of this expression is simply $\mathrm{P}(T \geqslant t_0 + t_1)$, and so the required probability is the ratio of the probability of survival beyond $t_0 + t_1$ to the probability of survival beyond t_0, that is $S(t_0 + t_1)/S(t_0)$. The survivor function for the exponential distribution is given by $S(t) = e^{-\lambda t}$, as in equation (5.5) of Chapter 5, and so

$$\mathrm{P}(T \geqslant t_0 + t_1 \mid T \geqslant t_0) = \frac{\exp\{-\lambda(t_0 + t_1)\}}{\exp(-\lambda t_0)} = e^{-\lambda t_1},$$

which is the survivor function of an exponential random variable at time t_1, that is $\mathrm{P}(T \geqslant t_1)$. This result means that, conditional on survival to time t_0, the excess survival time beyond t_0 also has an exponential distribution with mean λ^{-1}. In other words, the probability of survival beyond time t_0 is not affected by the knowledge that the individual has already survived to time t_0.

From this result, since r_{Ci} has a unit exponential distribution, the excess residual, Δ, will also have a unit exponential distribution. The expected value of Δ is therefore unity, suggesting that Δ may be taken to be unity, and this leads to modified Cox-Snell residuals, given by

$$r'_{Ci} = \begin{cases} r_{Ci} & \text{for uncensored observations,} \\ r_{Ci} + 1 & \text{for censored observations.} \end{cases} \tag{4.3}$$

The ith modified Cox-Snell residual can be expressed in an alternative form by introducing an event indicator, δ_i, which takes the value zero if the observed survival time of the ith individual is censored and unity if it is uncensored.

Then the modified Cox-Snell residual is given by

$$r'_{Ci} = 1 - \delta_i + r_{Ci}. \tag{4.4}$$

Note that from the definition of this type of residual, r'_{Ci} must be greater than unity for a censored observation. Also, as for the unmodified residuals, the r'_{Ci} can take any value between zero and infinity, and they will have a skew distribution.

On the basis of empirical evidence, Crowley and Hu (1977) found that the addition of unity to a Cox-Snell residual for a censored observation inflated the residual to too great an extent. They therefore suggested that the median value of the excess residual be used rather than the mean. For the unit exponential distribution, the survivor function is $S(t) = e^{-t}$, and so the median, $t(50)$, is such that $e^{-t(50)} = 0.5$, whence $t(50) = \log 2 = 0.693$. Thus a second version of the modified Cox-Snell residual has

$$r''_{Ci} = \begin{cases} r_{Ci} & \text{for uncensored observations,} \\ r_{Ci} + 0.693 & \text{for censored observations.} \end{cases} \tag{4.5}$$

However, if the proportion of censored observations is not too great, the set of residuals obtained from each of these two forms of modification will not appear too different.

4.1.3 Martingale residuals

The modified residuals r'_{Ci} defined in equation (4.4) have a mean of unity for uncensored observations. Accordingly, these residuals might be further refined by relocating the r'_{Ci} so that they have a mean of zero when an observation is uncensored. If in addition the resulting values are multiplied by -1, we obtain the residuals

$$r_{Mi} = \delta_i - r_{Ci}. \tag{4.6}$$

These residuals are known as *martingale residuals*, since they can also be derived using what are known as martingale methods. In this derivation, the r_{Ci} are based on the Nelson-Aalen estimate of the cumulative hazard function. Because these methods rely heavily on probability theory and stochastic processes, this approach will not be discussed in this book. However, a comprehensive account of the martingale approach to the analysis of survival data has been presented by a number of authors, including Andersen *et al.* (1993), Fleming and Harrington (1991) and Therneau and Grambsch (2000).

Martingale residuals take values between $-\infty$ and unity, with the residuals for censored observations, where $\delta_i = 0$, being negative. It can also be shown that these residuals sum to zero and, in large samples, the martingale residuals are uncorrelated with one another and have an expected value of zero. In this respect, they have properties similar to those possessed by residuals encountered in linear regression analysis.

Another way of looking at the martingale residuals is to note that the quantity r_{Mi} in equation (4.6) is the difference between the observed number of deaths for the ith individual in the interval $(0, t_i)$ and the corresponding

estimated expected number on the basis of the fitted model. To see this, note that the observed number of deaths is unity if the survival time t_i is uncensored, and zero if censored, that is δ_i. The second term in equation (4.6) is an estimate of $H_i(t_i)$, the cumulative hazard or cumulative probability of death for the ith individual over the interval $(0, t_i)$. Since we are dealing with just one individual, this can be viewed as the expected number of deaths in that interval. This shows another similarity between the martingale residuals and residuals from other areas of data analysis.

4.1.4 Deviance residuals

Although martingale residuals share many of the properties possessed by residuals encountered in other situations, such as in linear regression analysis, they are not symmetrically distributed about zero, even when the fitted model is correct. This skewness makes plots based on the residuals difficult to interpret. The deviance residuals, which were introduced by Therneau *et al.* (1990), are much more symmetrically distributed about zero. They are defined by

$$r_{Di} = \text{sgn}(r_{Mi}) \left[-2 \left\{r_{Mi} + \delta_i \log(\delta_i - r_{Mi})\right\}\right]^{\frac{1}{2}}, \tag{4.7}$$

where r_{Mi} is the martingale residual for the ith individual, and the function $\text{sgn}(\cdot)$ is the sign function. This is the function that takes the value $+1$ if its argument is positive and -1 if negative. Thus $\text{sgn}(r_{Mi})$ ensures that the deviance residuals have the same sign as the martingale residuals.

The original motivation for these residuals is that they are components of the *deviance*. The deviance is a statistic that is used to summarise the extent to which the fit of a model of current interest deviates from that of a model which is a perfect fit to the data. This latter model is called the *saturated* or *full* model, and is a model in which the β-coefficients are allowed to be different for each individual. The statistic is given by

$$D = -2\left\{\log \hat{L}_c - \log \hat{L}_f\right\},$$

where $\hat{L}_c$ is the maximised partial likelihood under the current model and $\hat{L}_f$ is the maximised partial likelihood under the full model. The smaller the value of the deviance, the better the model. The deviance can be regarded as a generalisation of the residual sum of squares used in modelling normal data to the analysis of non-normal data, and features prominently in generalised linear modelling. Note that differences in deviance between two alternative models are the same as differences in the values of the statistic $-2\log \hat{L}$ introduced in Chapter 3. The deviance residuals are then such that $D = \sum r_{Di}^2$, so that observations that correspond to relatively large deviance residuals are those that are not well fitted by the model.

Another way of viewing the deviance residuals is that they are martingale residuals that have been transformed to produce values that are symmetric about zero when the fitted model is appropriate. To see this, first recall that the martingale residuals r_{Mi} can take any value in the interval $(-\infty, 1)$. For large negative values of r_{Mi}, the term in square brackets in equation (4.7) is dominated by r_{Mi}. Taking the square root of this quantity has the effect of bringing the residual closer to zero. Thus martingale residuals in the range $(-\infty, 0)$ are shrunk toward zero. Now consider martingale residuals in the interval $(0, 1)$. The term $\delta_i \log(\delta_i - r_{Mi})$ in equation (4.7) will only be non-zero for uncensored observations, and will then have the value $\log(1 - r_{Mi})$. As r_{Mi} gets closer to unity, $1 - r_{Mi}$ gets closer to zero and $\log(1 - r_{Mi})$ takes large negative values. The quantity in square brackets in equation (4.7) is then dominated by this logarithmic term, and so the deviance residuals are expanded toward $+\infty$ as the martingale residual reaches its upper limit of unity.

One final point to note is that although these residuals can be expected to be symmetrically distributed about zero when an appropriate model has been fitted, they do not necessarily sum to zero.

4.1.5 Schoenfeld residuals*

Two disadvantages of the residuals described in Sections 4.1.1 to 4.1.4 are that they depend heavily on the observed survival time and require an estimate of the cumulative hazard function. Both of these disadvantages are overcome in a residual proposed by Schoenfeld (1982). These residuals were originally termed *partial residuals*, for reasons given in the sequel, but are now commonly known as *Schoenfeld residuals*. This residual differs from those considered previously in one other important respect. This is that there is not a single value of the residual for each individual, but a set of values, one for each explanatory variable included in the fitted Cox regression model.

The ith partial or Schoenfeld residual for X_j, the jth explanatory variable in the model, is given by

$$r_{Pji} = \delta_i\{x_{ji} - \hat{a}_{ji}\}, \tag{4.8}$$

where x_{ji} is the value of the jth explanatory variable, $j = 1, 2, \ldots, p$, for the ith individual in the study,

$$\hat{a}_{ji} = \frac{\sum_{l \in R(t_i)} x_{jl} \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l)}{\sum_{l \in R(t_i)} \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l)}, \tag{4.9}$$

and $R(t_i)$ is the set of all individuals at risk at time t_i.

Note that non-zero values of these residuals only arise for uncensored observations. Moreover, if the largest observation in a sample of survival times is uncensored, the value of $\hat{a}_{ji}$ for that observation, from equation (4.9), will be equal to x_{ji} and so $r_{Pji} = 0$. To distinguish residuals that are genuinely zero from those obtained from censored observations, the latter are usually expressed as missing values.

The ith Schoenfeld residual, for the explanatory variable X_j, is an estimate of the ith component of the first derivative of the logarithm of the partial

likelihood function with respect to β_j, which, from equation (3.5), is given by

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \delta_i \{x_{ji} - a_{ji}\}, \tag{4.10}$$

where

$$a_{ji} = \frac{\sum_l x_{jl} \exp(\boldsymbol{\beta}' \boldsymbol{x}_l)}{\sum_l \exp(\boldsymbol{\beta}' \boldsymbol{x}_l)} \tag{4.11}$$

The ith term in this summation, evaluated at $\hat{\boldsymbol{\beta}}$, is then the Schoenfeld residual for X_j, given in equation (4.8). Since the estimates of the β's are such that

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \beta_j}\Big|_{\hat{\boldsymbol{\beta}}} = 0,$$

the Schoenfeld residuals must sum to zero. These residuals also have the property that, in large samples, the expected value of r_{Pji} is zero, and they are uncorrelated with one another.

It turns out that a scaled version of the Schoenfeld residuals, proposed by Grambsch and Therneau (1994), is more effective in detecting departures from the assumed model. Let the vector of Schoenfeld residuals for the ith individual be denoted $\boldsymbol{r}_{Pi} = (r_{P1i}, r_{P2i}, \ldots, r_{Ppi})'$. The scaled, or weighted, Schoenfeld residuals, r^*_{Pji}, are then the components of the vector

$$\boldsymbol{r}^*_{Pi} = r \operatorname{var}(\hat{\boldsymbol{\beta}}) \boldsymbol{r}_{Pi},$$

where r is the number of deaths among the n individuals, and $\operatorname{var}(\hat{\boldsymbol{\beta}})$ is the variance-covariance matrix of the parameter estimates in the fitted Cox regression model. These scaled Schoenfeld residuals are therefore quite straightforward to compute.

4.1.6 Score residuals*

There is one other type of residual that is useful in some aspects of model checking, and which, like the Schoenfeld residual, is obtained from the first derivative of the logarithm of the partial likelihood function with respect to the parameter β_j, $j = 1, 2, \ldots, p$. However, the derivative in equation (4.10) is now expressed in a quite different form, namely

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \left\{ \delta_i (x_{ji} - a_{ji}) + \exp(\boldsymbol{\beta}' \boldsymbol{x}_i) \sum_{t_r \leqslant t_i} \frac{(a_{jr} - x_{ji})\delta_r}{\sum_{l \in R(t_r)} \exp(\boldsymbol{\beta}' \boldsymbol{x}_l)} \right\}, \tag{4.12}$$

where x_{ji} is the ith value of the jth explanatory variable, δ_i is the event indicator which is zero for censored observations and unity otherwise, a_{ji} is given in equation (4.11), and $R(t_r)$ is the risk set at time t_r. In this formulation, the contribution of the ith observation to the derivative only depends on information up to time t_i. In other words, if the study was actually concluded at time t_i, the ith component of the derivative would be unaffected.

Residuals are then obtained as the estimated value of the n components of the derivative. From Appendix A, the first derivative of the logarithm of the partial likelihood function, with respect to β_j, is the efficient score for β_j, and so these residuals are known as *score residuals*.

From equation (4.12), the ith score residual, $i = 1, 2, \ldots, n$, for the jth explanatory variable in the model, X_j, is given by

$$r_{Sji} = \delta_i (x_{ji} - \hat{a}_{ji}) + \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i) \sum_{t_r \leqslant t_i} \frac{(\hat{a}_{jr} - x_{ji})\delta_r}{\sum_{l \in R(t_r)} \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l)}.$$

Using equation (4.8), this may be written in the form

$$r_{Sji} = r_{Pji} + \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i) \sum_{t_r \leqslant t_i} \frac{(\hat{a}_{jr} - x_{ji})\delta_r}{\sum_{l \in R(t_r)} \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l)}, \tag{4.13}$$

which shows that the score residuals are modifications of the Schoenfeld residuals. As for the Schoenfeld residuals, the score residuals sum to zero, but will not necessarily be zero when an observation is censored.

In this section, a number of residuals have been defined. We conclude with an example that illustrates the calculation of these different types of residual and that shows similarities and differences between them. This example will be used in many illustrations in this chapter, mainly because the relatively small number of observations allows the values of the residuals and other diagnostics to be readily tabulated. However, the methods of this chapter are generally more informative in larger data sets.

Example 4.1 Infection in patients on dialysis

In the treatment of certain disorders of the kidney, dialysis may be used to remove waste materials from the blood. One problem that can occur in patients on dialysis is the occurrence of an infection at the site at which the catheter is inserted. If any such infection occurs, the catheter must be removed, and the infection cleared up. In a study to investigate the incidence of infection, the time from insertion of the catheter until infection was recorded for a group of kidney patients. Sometimes, the catheter has to be removed for reasons other than infection, giving rise to right-censored observations. The data in Table 4.1 give the number of days from insertion of the catheter until its removal following the first occurrence of an infection. The data set includes the values of a variable that indicates the infection status of an individual, which takes the value zero if the catheter was removed for a reason other than the occurrence of an infection, and unity otherwise. Also given is the age of each patient in years and a variable that denotes the sex of each patient (1 = male, 2 = female). These data are taken from McGilchrist and Aisbett (1991), and relate to the 13 patients suffering from diseases of the kidney coded as type 3 in their paper.

When a Cox regression model is fitted these data, the fitted hazard function

Table 4.1 *Times to removal of a catheter following a kidney infection.*

Patient	Time	Status	Age	Sex
1	8	1	28	1
2	15	1	44	2
3	22	1	32	1
4	24	1	16	2
5	30	1	10	1
6	54	0	42	2
7	119	1	22	2
8	141	1	34	2
9	185	1	60	2
10	292	1	43	2
11	402	1	30	2
12	447	1	31	2
13	536	1	17	2

for the ith patient, $i = 1, 2, \ldots, 13$, is found to be

$$\hat{h}_i(t) = \exp\{0.030\, Age_i - 2.711\, Sex_i\}\, \hat{h}_0(t), \tag{4.14}$$

where Age_i and Sex_i refer to the age and sex of the ith patient.

The variable *Sex* is certainly important, since when *Sex* is added to the model that contains *Age* alone, the decrease in the value of the statistic $-2\log\hat{L}$ is 6.445 on 1 d.f. This change is highly significant ($P = 0.011$). On the other hand, there is no statistical evidence for including the variable *Age* in the model, since the change in the value of the statistic $-2\log\hat{L}$ on adding *Age* to the model that contains *Sex* is 1.320 on 1 d.f. ($P = 0.251$). However, it can be argued that from the clinical viewpoint, the hazard of infection may well depend on age. Consequently, both variables will be retained in the model.

The values of different types of residual for the model in equation (4.14) are displayed in Table 4.2. In this table, r_{Ci}, r_{Mi} and r_{Di} are the Cox-Snell residuals, martingale residuals and deviance residuals, respectively. Also r_{P1i} and r_{P2i} are the values of Schoenfeld residuals for the variables *Age* and *Sex*, respectively, r^*_{P1i} and r^*_{P2i} are the corresponding scaled Schoenfeld residuals, and r_{S1i}, r_{S2i} are the score residuals.

The values in this table were computed using the Nelson-Aalen estimate of the baseline cumulative hazard function given in equation (3.25). Had the estimate $\hat{H}_0(t)$, in equation (3.19), been used, different values for all but the Schoenfeld residuals would be obtained. In addition, because the corresponding estimate of the survivor function is zero at the longest removal time, which is that for patient number 13, values of the Cox-Snell, martingale and deviance residuals would not then be defined for this patient, and the martingale residuals would no longer sum to zero.

Table 4.2 *Different types of residual after fitting a Cox regression model.*

Patient	r_{Ci}	r_{Mi}	r_{Di}	r_{P1i}	r_{P2i}	r^*_{P1i}	r^*_{P2i}	r_{S1i}	r_{S2i}
1	0.280	0.720	1.052	−1.085	−0.242	0.033	−3.295	−0.781	−0.174
2	0.072	0.928	1.843	14.493	0.664	0.005	7.069	13.432	0.614
3	1.214	−0.214	−0.200	3.129	−0.306	0.079	−4.958	−0.322	0.058
4	0.084	0.916	1.765	−10.222	0.434	−0.159	8.023	−9.214	0.384
5	1.506	−0.506	−0.439	−16.588	−0.550	−0.042	−5.064	9.833	0.130
6	0.265	−0.265	−0.728	–	–	–	–	−3.826	−0.145
7	0.235	0.765	1.168	−17.829	0.000	−0.147	3.083	−15.401	−0.079
8	0.484	0.516	0.648	−7.620	0.000	−0.063	1.318	−7.091	−0.114
9	1.438	−0.438	−0.387	17.091	0.000	0.141	−2.955	−15.811	−0.251
10	1.212	−0.212	−0.199	10.239	0.000	0.085	−1.770	1.564	−0.150
11	1.187	−0.187	−0.176	2.857	0.000	0.024	−0.494	6.575	−0.101
12	1.828	−0.828	−0.670	5.534	0.000	0.046	−0.957	4.797	−0.104
13	2.195	−1.195	−0.904	0.000	0.000	0.000	0.000	16.246	−0.068

In this data set, there is just one censored observation, which is for patient number 6. Therefore, the modified Cox-Snell residuals will be the same as the Cox-Snell residuals for all patients except number 6. For this patient, the values of the two forms of modified residuals are $r'_{C6} = 1.265$ and $r''_{C6} = 0.958$. Also, the Schoenfeld residuals are not defined for the patient with a censored removal time, and are zero for the patient that has the longest period of time before removal of the catheter.

The skewness of the Cox-Snell and martingale residuals is clearly shown in Table 4.2, as is the fact that the Cox-Snell residuals are centred on unity while the martingale and deviance residuals are centred on zero. Note also that the martingale, Schoenfeld and score residuals sum to zero, as they should do. One unusual feature about the residuals in Table 4.2 is the large number of zeros for the values of the Schoenfeld residual corresponding to *Sex*. The reason for this is that for infection times greater than 30 days, the value of the variable *Sex* is always equal to 2. This means that the value of the term $\hat{a}_{ji}$ for this variable, given in equation (4.9), is equal to 2 for a survival time greater than 30 days, and so the corresponding Schoenfeld residual defined in equation (4.8) is zero.

We now consider how residuals obtained after fitting a Cox regression model can be used to throw light on the extent to which the fitted model provides an appropriate description of the observed data. We will then be in a position to study the residuals obtained in Example 4.1 in greater detail.

4.2 Assessment of model fit

A number of plots based on residuals can be used in the graphical assessment of the adequacy of a fitted model. Unfortunately, many graphical procedures

that are analogues of residual plots used in linear regression analysis have not proved to be very helpful. This is because plots of residuals against quantities such as the observed survival times, or the rank order of these times, often exhibit a definite pattern, even when the correct model has been fitted. Traditionally, plots of residuals have been based on the Cox-Snell residuals, or adjusted versions of them described in Section 4.1.2. The use of these residuals is therefore reviewed in the next section, and this is followed by a description of how some other types of residuals may be used in the graphical assessment of the fit of a model.

4.2.1 Plots based on the Cox-Snell residuals

In Section 4.1.1, the Cox-Snell residuals were shown to have an exponential distribution with unit mean, if the fitted model is correct. They therefore have a mean and variance of unity, and are asymmetrically distributed about the mean. This means that simple plots of the residuals, such as plots of the residuals against the observation number, known as *index plots*, will not lead to a symmetric display. The residuals are also correlated with the survival times, and so plots of these residuals against quantities such as the observed survival times, or the rank order of these times are also unhelpful.

One particular plot of these residuals, that can be used to assess the overall fit of the model, leads to an assessment of whether the residuals are indeed a plausible sample from a unit exponential distribution. This plot is based on the fact that if a random variable T has an exponential distribution with unit mean, then the survivor function of T is e^{-t}; see Section 5.1.1 of Chapter 5. Accordingly, a plot of the cumulative hazard function $H(t) = -\log S(t)$ against t, known as a *cumulative hazard plot*, will give a straight line through the origin with unit slope.

This result can be used to examine whether the residuals have a unit exponential distribution. After computing the Cox-Snell residuals, r_{Ci}, the Kaplan-Meier estimate of the survivor function of these values is found. This estimate is computed in a similar manner to the Kaplan-Meier estimate of the survivor function of survival times, except that the data on which the estimate is based are now the residuals r_{Ci}. Residuals obtained from censored survival times are themselves taken to be censored. Denoting the estimate by $\hat{S}(r_{Ci})$, the values of $\hat{H}(r_{Ci}) = -\log \hat{S}(r_{Ci})$ are plotted against r_{Ci}. This gives a cumulative hazard plot of the residuals. A straight line with unit slope and zero intercept will then indicate that the fitted survival model is satisfactory. On the other hand, a plot that displays a systematic departure from a straight line, or yields a line that does not have approximately unit slope or zero intercept, might suggest that the model needs to be modified in some way. Equivalently, a *log-cumulative hazard plot* of the residuals, that is a plot of $\log \hat{H}(r_{Ci})$ against $\log r_{Ci}$ may be used. This plot is discussed in more detail in Section 4.4.1.

Example 4.2 Infection in patients on dialysis
Consider again the data on the time to the occurrence of an infection in kidney patients, described in Example 4.1. In this example, we first examine whether the Cox-Snell residuals are a plausible sample of observations from a unit exponential distribution. For this, the Kaplan-Meier estimate of the survivor function of the Cox-Snell residuals, $\hat{S}(r_{Ci})$, is obtained. The cumulative hazard function of the residuals, $\hat{H}(r_{Ci})$, derived from $-\log \hat{S}(r_{Ci})$, is then plotted against the corresponding residual to give a cumulative hazard plot of the residuals. The details of this calculation are summarised in Table 4.3, and the cumulative hazard plot is shown in Figure 4.1. The residual for patient number 6 is omitted from Table 4.3 because this observation is censored.

Table 4.3 *Calculation of the cumulative hazard function of the Cox-Snell residuals.*

r_{Ci}	$\hat{S}(r_{Ci})$	$\hat{H}(r_{Ci})$
0.072	0.9231	0.080
0.084	0.8462	0.167
0.235	0.7692	0.262
0.280	0.6838	0.380
0.484	0.5983	0.514
1.187	0.5128	0.668
1.212	0.4274	0.850
1.214	0.3419	1.073
1.438	0.2564	1.361
1.506	0.1709	1.767
1.828	0.0855	2.459
2.195	0.0000	–

The relatively small number of observations in this data set makes it difficult to interpret plots of residuals. However, the plotted points in Figure 4.1 are fairly close to a straight line through the origin, which has approximately unit slope. This could suggest that the model fitted to the data given in Table 4.1 is satisfactory.

On the face of it, this procedure would appear to have some merit, but cumulative hazard plots of the Cox-Snell residuals have not proved to be very useful in practice. In an earlier section it was argued that since the values $-\log S(t_i)$ have a unit exponential distribution, the Cox-Snell residuals, which are estimates of these quantities, should have an approximate unit exponential distribution when the fitted model is correct. This result is then used when interpreting a cumulative hazard plot of the residuals. Unfortunately this approximation is not very reliable, particularly in small samples. This is because estimates of the β's, and also of the baseline cumulative hazard function, $H_0(t)$, are needed in the computation of the r_{Ci}. The substitution of estimates means that the actual distribution of the residuals is not necessarily unit exponential, but their exact distribution is not known. In fact, the