The coefficient of x_i in this model can then be interpreted as the logarithm of a hazard ratio. Now consider the ratio of the hazard of death for an individual for whom the value $x+1$ is recorded on X, relative to one for whom the value x is obtained. This is

$$\frac{\exp\{\beta(x+1)\}}{\exp(\beta x)} = e^{\beta},$$

and so $\hat{\beta}$ in the fitted proportional hazards model is the estimated change in the logarithm of the hazard ratio when the value of X is increased by one unit.

Using a similar argument, the estimated change in the log-hazard ratio when the value of the variable X is increased by r units is $r\hat{\beta}$, and the corresponding estimate of the hazard ratio is $\exp(r\hat{\beta})$. The standard error of the estimated log-hazard ratio will be $r\,\mathrm{se}\,(\hat{\beta})$, from which confidence intervals for the true hazard ratio can be derived.

The above argument shows that when a continuous variable X is included in a proportional hazards model, the hazard ratio when the value of X is changed by r units does not depend on the actual value of X. For example, if X refers to the age of an individual, the hazard ratio for an individual aged 70, relative to one aged 65, would be the same as that for an individual aged 20, relative to one aged 15. This feature is a direct result of fitting X as a linear term in the proportional hazards model. If there is doubt about the assumption of linearity, a factor whose levels correspond to different sets of values of X can be fitted. The linearity assumption can then be checked using the procedure described in Section 3.6.2.

3.7.2 Models with a factor

When individuals fall into one of m groups, $m \geqslant 2$, which correspond to categories of an explanatory variable, the groups can be indexed by the levels of a factor. Under a proportional hazards model, the hazard function for an individual in the jth group, $j = 1, 2, \ldots, m$, is given by

$$h_j(t) = \exp(\gamma_j) h_0(t),$$

where γ_j is the effect due to the jth level of the factor, and $h_0(t)$ is the baseline hazard function. This model is overparameterised, and so, as in Section 3.2.2, we take $\gamma_1 = 0$. The baseline hazard function then corresponds to the hazard of death at time t for an individual in the first group. The ratio of the hazards at time t for an individual in the jth group, $j \geqslant 2$, relative to an individual in the first group, is then $\exp(\gamma_j)$. Consequently, the parameter γ_j is the logarithm of this relative hazard, that is,

$$\gamma_j = \log\{h_j(t)/h_0(t)\}.$$

A model that contains the terms γ_j, $j = 1, 2, \ldots, m$, with $\gamma_1 = 0$, can be fitted by defining $m-1$ indicator variables, $X_2, X_3, \ldots, X_m$, as shown in [illegible] model leads to estimates $\hat{\gamma}_2, \hat{\gamma}_3, \ldots, \hat{\gamma}_m$, and their

standard errors. The estimated logarithm of the relative hazard for an individual in group j, relative to an individual in group 1, is then $\hat{\gamma}_j$.

A $100(1-\alpha)\%$ confidence interval for the true log-hazard ratio is the interval from $\hat{\gamma}_j - z_{\alpha/2}\,\mathrm{se}\,(\hat{\gamma}_j)$ to $\hat{\gamma}_j + z_{\alpha/2}\,\mathrm{se}\,(\hat{\gamma}_j)$, where $z_{\alpha/2}$ is the upper $\alpha/2$-point of the standard normal distribution. A corresponding confidence interval for the hazard ratio itself is obtained by exponentiating these confidence limits.

Example 3.8 Treatment of hypernephroma

Data on the survival times of patients with hypernephroma were given in Table 3.3. In this example, we will only consider the data from those patients on whom a nephrectomy has been performed, given in columns 4 to 6 of Table 3.3. The survival times of this set of patients are classified according to their age group. If the effect due to the jth age group is denoted by α_j, $j = 1, 2, 3$, the proportional hazards model for the hazard at time t for a patient in the jth age group is such that

$$h_j(t) = \exp(\alpha_j) h_0(t).$$

This model can be fitted by defining two indicator variables, A_2 and A_3, where A_2 is unity if the patient is aged between 60 and 70, and A_3 is unity if the patient is more than 70 years of age, as in Example 3.4. This corresponds to taking $\alpha_1 = 0$.

The value of $-2\log\hat{L}$ for the null model is 128.901, and when the term α_j is added, the value of this statistic reduces to 122.501. This reduction of 6.400 on 2 d.f. is significant at the 5% level ($P = 0.041$), and so we conclude that the hazard function does depend on which age group the patient is in.

The coefficients of the indicator variables A_2 and A_3 are estimates of α_2 and α_3, respectively, and are given in Table 3.7. Since the constraint $\alpha_1 = 0$ has been used, $\hat{\alpha}_1 = 0$.

Table 3.7 *Parameter estimates and their standard errors on fitting a proportional hazards model to data from Example 3.4.*

Parameter	Estimate	s.e.
α_2	−0.065	0.498
α_3	1.824	0.682

The hazard ratio for a patient aged 60–70, relative to one aged less than 60, is $e^{-0.065} = 0.94$, while that for a patient whose age is greater than 70, relative to one aged less than 60, is $e^{1.824} = 6.20$. These results suggest that the hazard of death at any given time is greatest for patients who are older than 70, but that there is little difference in the hazard functions for patients in the other two age groups.

The standard error of the parameter estimates in Table 3.7 can be used to

obtain confidence intervals for the true hazard ratios. A 95% confidence interval for the log-hazard ratio for a patient whose age is between 60 and 70, relative to one aged less than 60, is the interval with limits $-0.065 \pm (1.96 \times 0.498)$, that is, the interval $(-1.041, 0.912)$. The corresponding 95% confidence interval for the hazard ratio itself is $(0.35, 2.49)$. This confidence interval includes unity, which suggests that the hazard function for an individual whose age is between 60 and 70 is similar to that of a patient aged less than 60. Similarly, a 95% confidence interval for the hazard for a patient aged greater than 70, relative to one aged less than 60, is found to be $(1.63, 23.59)$. This interval does not include unity, and so an individual whose age is greater than 70 has a significantly greater hazard of death, at any given time, than patients aged less than 60.

In some applications, the hazard ratio relative to the level of a factor other than the first may be required. In these circumstances, the levels of the factor, and associated indicator variables, could be redefined so that some other level of the factor corresponds to the required baseline level, and the model refitted. The required estimates can also be found directly from the estimates obtained when the first level of the original factor is taken as the baseline, although this is more difficult.

The hazard functions for individuals at levels j and j' of the factor are $\exp(\alpha_j)h_0(t)$ and $\exp(\alpha_{j'})h_0(t)$, respectively, and so the hazard ratio for an individual at level j, relative to one at level j', is $\exp(\alpha_j - \alpha_{j'})$. The log-hazard ratio is then $\alpha_j - \alpha_{j'}$, which is estimated by $\hat{\alpha}_j - \hat{\alpha}_{j'}$.

To obtain the standard error of this estimate, we use the result that the variance of the difference $\hat{\alpha}_j - \hat{\alpha}_{j'}$ is given by

$$\text{var}\,(\hat{\alpha}_j - \hat{\alpha}_{j'}) = \text{var}\,(\hat{\alpha}_j) + \text{var}\,(\hat{\alpha}_{j'}) - 2\,\text{cov}\,(\hat{\alpha}_j, \hat{\alpha}_{j'}).$$

In view of this, an estimate of the covariance between $\hat{\alpha}_j$ and $\hat{\alpha}_{j'}$, as well as estimates of their variance, will be needed to compute $\text{se}\,(\hat{\alpha}_j - \hat{\alpha}_{j'})$. The calculations are illustrated in Example 3.9.

Example 3.9 Treatment of hypernephroma
Consider again the subset of the data from Example 3.4, corresponding to those patients who have had a nephrectomy. Suppose that an estimate of the hazard ratio for an individual aged greater than 70, relative to one aged between 60 and 70, is required. Using the estimates in Table 3.7, the estimated log-hazard ratio is $\hat{\alpha}_3 - \hat{\alpha}_2 = 1.824 + 0.065 = 1.889$, and so the estimated hazard ratio is $e^{1.889} = 6.61$. This suggests that the hazard of death at any given time for someone aged greater than 70 is more than six and a half times that for someone aged between 60 and 70.

The variance of $\hat{\alpha}_3 - \hat{\alpha}_2$ is

$$\text{var}\,(\hat{\alpha}_3) + \text{var}\,(\hat{\alpha}_2) - 2\,\text{cov}\,(\hat{\alpha}_3, \hat{\alpha}_2),$$

and the variance-covariance matrix of the parameter estimates gives the required variances and covariance. This matrix can be obtained from statistical

packages used to fit the Cox regression model, and is found to be

$$\begin{array}{c} A_2 \\ A_3 \end{array} \begin{pmatrix} 0.2484 & 0.0832 \\ 0.0832 & 0.4649 \end{pmatrix}, \quad \begin{array}{cc} A_2 & A_3 \end{array}$$

from which $\text{var}\,(\hat{\alpha}_2) = 0.2484$, $\text{var}\,(\hat{\alpha}_3) = 0.4649$, and $\text{cov}\,(\hat{\alpha}_2, \hat{\alpha}_3) = 0.0832$. Of course, the variances are simply the squares of the standard errors in Table 3.7. It then follows that

$$\text{var}\,(\hat{\alpha}_3 - \hat{\alpha}_2) = 0.4649 + 0.2484 - (2 \times 0.0832) = 0.5469,$$

and so the standard error of $\hat{\alpha}_2 - \hat{\alpha}_3$ is 0.740. Consequently a 95% confidence interval for the log-hazard ratio is $(0.440, 3.338)$ and that for the hazard ratio itself is $(1.55, 8.18)$.

An easier way of obtaining the estimated value of the hazard ratio for an individual who is aged greater than 70, relative to one aged between 60 and 70, and the standard error of the estimate, is to redefine the levels of the factor associated with age group. Suppose that the data are now arranged so that the first level of the factor corresponds to the age range 60–70, level 2 corresponds to patients aged greater than 70 and level 3 to those aged less than 60. Choosing indicator variables to be such that the effect due to the first level of the redefined factor is set equal to zero leads to the variables B_2 and B_3 defined in the table below.

Age group	B_2	B_3
<60	0	1
60–70	0	0
>70	1	0

The estimated log-hazard ratio is now simply the estimated coefficient of B_2, and its standard error can be read directly from standard computer output.

The manner in which the coefficients of indicator variables are interpreted is crucially dependent upon the coding that has been used for them. This means that when a proportional hazards model is fitted using a statistical package that enables factors to be fitted directly, it is essential to know how indicator variables used within the package have been defined.

As a further illustration, suppose that individuals fall into one of m groups and that the coding used for the $m-1$ indicator variables, $X_2, X_3, \ldots, X_m$, is such that the sum of the main effects of A, $\sum_{j=1}^{m} \alpha_j$, is equal to zero. The values of the indicator variables corresponding to an m-level factor A, are then as shown in the following table.

Level of A	X_2	X_3	. . .	X_m
1	−1	−1	. . .	−1
2	1	0	. . .	0
3	0	1	. . .	0
. . .	. . .	. . .	. . .	. . .
m	0	0	. . .	1

With this choice of indicator variables, A proportional hazards model that contains this factor can be expressed in the form

$$h_j(t) = \exp(\alpha_2 x_2 + \alpha_3 x_3 + \cdots + \alpha_m x_m) h_0(t),$$

where x_j is the value of X_j for an individual for whom the factor A is at the jth level, $j = 2, 3, \ldots, m$. The hazard of death at a given time for an individual at the first level of the factor is

$$\exp\{-(\alpha_2 + \alpha_3 + \cdots + \alpha_m)\},$$

while that for an individual at the jth level of the factor is

$$\exp(\alpha_j),$$

for $j \geqslant 2$. The ratio of the hazard for an individual in group j, $j \geqslant 2$, relative to that of an individual in the first group, is then

$$\exp(\alpha_j + \alpha_2 + \alpha_3 + \cdots + \alpha_m).$$

For example, if $m = 4$ and $j = 3$, the hazard ratio is $\exp(\alpha_2 + 2\alpha_3 + \alpha_4)$, and the variance of the corresponding estimated log-hazard ratio is

$$\begin{aligned}&\operatorname{var}(\hat{\alpha}_2) + 4\operatorname{var}(\hat{\alpha}_3) + \operatorname{var}(\hat{\alpha}_4) + 4\operatorname{cov}(\hat{\alpha}_2, \hat{\alpha}_3)\\ &\quad + 4\operatorname{cov}(\hat{\alpha}_3, \hat{\alpha}_4) + 2\operatorname{cov}(\hat{\alpha}_2, \hat{\alpha}_4).\end{aligned}$$

Each of the terms in this expression can be found from the variance-covariance matrix of the parameter estimates after fitting a proportional hazards model, and a confidence interval for the hazard ratio obtained. Although this is reasonably straightforward, this particular coding of the indicator variables does make it much more complicated to interpret the individual parameter estimates in a fitted model.

3.7.3 Models with combinations of terms

In previous sections, we have only considered the interpretation of parameter estimates in proportional hazards models that contain a single term. More generally, a fitted model will contain terms corresponding to a number of variates, factors or combinations of the two. With suitable coding of indicator variables corresponding to factors in the model, the parameter estimates can again be interpreted as logarithms of hazard ratios.

When a model contains more than one variable, the parameter estimate

associated with a particular effect is said to be adjusted for the other variables in the model, and so the estimates are log-hazard ratios, adjusted for the other terms in the model. The proportional hazards model can therefore be used to estimate hazard ratios, taking account of other variables included in the model.

When interactions between factors, or mixed terms involving factors and variates, are fitted, the estimated log-hazard ratios for a particular factor will differ according to the level of any factor, or the value of any variate with which it interacts. In this situation, the value of any such factor level or variate will need to be made clear when the estimated hazard ratios for the factor of primary interest are presented.

Instead of giving algebraic details on how hazard ratios can be estimated after fitting models with different combinations of terms, the general approach will be illustrated in two examples. The first of these involves both factors and variates, while the second includes an interaction between two factors.

Example 3.10 Comparison of two treatments for prostatic cancer
In Example 3.6, the most important prognostic variables in the study on the survival of prostatic cancer patients were found to be size of tumour (*Size*) and the Gleason index of tumour stage (*Index*). The indicator variable *Treat*, which represents the treatment effect, is also included in a proportional hazards model, since the aim of the study is to quantify the treatment effect. The model for the ith individual can then be expressed in the form

$$h_i(t) = \exp\{\beta_1 \mathit{Size}_i + \beta_2 \mathit{Index}_i + \beta_3 \mathit{Treat}_i\} h_0(t),$$

for $i = 1, 2, \ldots, 38$. Estimates of the β-coefficients and their standard errors on fitting this model are given in Table 3.8.

Table 3.8 *Estimated coefficients of the explanatory variables on fitting a proportional hazards model to the data from Example 1.4.*

Variable	$\hat{\beta}$	se $(\hat{\beta})$
Size	0.083	0.048
Index	0.710	0.338
Treat	−1.113	1.203

The estimated log-hazard ratio for an individual on the active treatment, DES, (*Treat* = 1) relative to an individual on the placebo (*Treat* = 0), with the same values of *Size* and *Index* as the individual on DES, is $\hat{\beta}_3 = -1.113$. Consequently the estimated hazard ratio is $e^{-1.113} = 0.329$. The value of this hazard ratio is unaffected by the actual values of *Size* and *Index*. However, since these two explanatory variables were included in the model, the estimated hazard ratio is adjusted for these variables.

For comparison, if a model that only contains *Treat* is fitted, the estimated coefficient of *Treat* is -1.978. The estimated hazard ratio for an individual on DES, relative to one on the placebo, unadjusted for *Size* and *Index*, is now $e^{-1.978} = 0.14$. This shows that unless proper account is taken of the effect of size of tumour and index of tumour grade, the extent of the treatment effect is overestimated.

Now consider the hazard ratio for an individual on a particular treatment with a given value of the variable *Index* and a tumour of a given size, relative to an individual on the same treatment with the same value of *Index*, but whose size of tumour is one unit less. This is $e^{0.083} = 1.09$. Since this is greater than unity, we conclude that, other things being equal, the greater the size of the tumour, the greater that hazard of death at any given time. Similarly, the hazard ratio for an individual on a given treatment with a given value of *Size*, relative to one on the same treatment with the same value of *Size*, whose value of *Index* is one unit less, is $e^{0.710} = 2.03$. This again means that the greater the value of the Gleason index, the greater is the hazard of death at any given time. In particular, an increase of one unit in the value of *Index* doubles the hazard of death.

Example 3.11 Treatment of hypernephroma

Consider again the full set of data on survival times following treatment for hypernephroma, given in Table 3.3. In Example 3.4, the most appropriate proportional hazards model was found to contain terms α_j, $j = 1, 2, 3$, corresponding to age group, and terms ν_k, $k = 1, 2$, corresponding to whether or not a nephrectomy was performed. For illustrative purposes, in this example we will consider the model that also contains the interaction between these two factors, even though it was found not to be significant. Under this model, the hazard function for an individual in the jth age group and the kth level of nephrectomy status is

$$h(t) = \exp\{\alpha_j + \nu_k + (\alpha\nu)_{jk}\}h_0(t), \tag{3.13}$$

where $(\alpha\nu)_{jk}$ is the term corresponding to the interaction.

Consider the ratio of the hazard of death at time t for a patient in the jth age group, $j = 1, 2, 3$, and the kth level of nephrectomy status, $k = 1, 2$, relative to an individual in the first age group who has not had a nephrectomy, which is

$$\frac{\exp\{\alpha_j + \nu_k + (\alpha\nu)_{jk}\}}{\exp\{\alpha_1 + \nu_1 + (\alpha\nu)_{11}\}}.$$

As in Example 3.4, the model in equation (3.13) is fitted by including the indicator variables A_2, A_3, and N in the model, together with the products A_2N and A_3N. The estimated coefficients of these variables are then $\hat{\alpha}_2$, $\hat{\alpha}_3$, $\hat{\nu}_2$, $\widehat{(\alpha\nu)}_{22}$, and $\widehat{(\alpha\nu)}_{32}$, respectively. From the coding of the indicator variables that has been used, the estimates $\hat{\alpha}_1$, $\hat{\nu}_1$, $\widehat{(\alpha\nu)}_{11}$ and $\widehat{(\alpha\nu)}_{12}$ are all zero. The estimated hazard ratio for an individual in the jth age group, $j = 1, 2, 3$, and the kth level of nephrectomy status, $k = 1, 2$, relative to one in the first age group who has not had a nephrectomy, is then just

$$\exp\{\hat{\alpha}_j + \hat{\nu}_k + \widehat{(\alpha\nu)}_{jk}\}.$$

The non-zero parameter estimates are $\hat{\alpha}_2 = 0.005$, $\hat{\alpha}_3 = 0.065$, $\hat{\nu}_2 = -1.943$, $\widehat{(\alpha\nu)}_{22} = -0.051$, and $\widehat{(\alpha\nu)}_{32} = 2.003$. The estimated hazard ratios are summarised in Table 3.9.

Table 3.9 *Estimated hazard ratios on fitting a model that contains an interaction to the data from Example 3.4.*

Age group	No nephrectomy	Nephrectomy
<60	1.000	0.143
60–70	1.005	0.137
>70	1.067	1.133

Inclusion of the combination of factor levels for which the estimated hazard ratio is 1.00, in tables such as Table 3.9, emphasises that the hazards are relative to those for individuals in the first age group who have not had a nephrectomy. This table shows that individuals aged less than or equal to 70, who have had a nephrectomy, have a much reduced hazard of death, compared to those in the other age group and those who have not had a nephrectomy.

Confidence intervals for the corresponding true hazard ratios can be found using the method described in Section 3.7.2. As a further illustration, a confidence interval will be obtained for the hazard ratio for individuals who have had a nephrectomy in the second age group relative to those in the first. The log-hazard ratio is now $\hat{\alpha}_2 + \widehat{(\alpha\nu)}_{22}$, and so the estimated hazard ratio is 0.955. The variance of this estimate is given by

$$\text{var}\,(\hat{\alpha}_2) + \text{var}\,\{\widehat{(\alpha\nu)}_{22}\} + 2\,\text{cov}\,\{\hat{\alpha}_2, \widehat{(\alpha\nu)}_{22}\}.$$

From the variance-covariance matrix of the parameter estimates after fitting the model in equation (3.13), $\text{var}\,(\hat{\alpha}_2) = 0.697$, $\text{var}\,\{\widehat{(\alpha\nu)}_{22}\} = 0.942$, and the covariance term is $\text{cov}\,\{\hat{\alpha}_2, \widehat{(\alpha\nu)}_{22}\} = -0.695$. Consequently, the variance of the estimated log-hazard ratio is 0.248, and so a 95% confidence interval for the true log-hazard ratio ranges from -0.532 to 0.441. The corresponding confidence interval for the true hazard ratio is $(0.59, 1.55)$. This interval includes unity, and so the hazard ratio of 0.955 is not significantly different from unity at the 5% level. Confidence intervals for the hazard ratios in Table 3.9 can be found in a similar manner.

3.8* Estimating the hazard and survivor functions

So far in this chapter, we have only considered the estimation of the β-parameters in the linear component of a proportional hazards model. As we

have seen, this is all that is required in order to draw inferences about the effect of explanatory variables in the model on the hazard function. Once a suitable model for a set of survival data has been identified, the hazard function, and the corresponding survivor function, can be estimated. These estimates can then be used to summarise the survival experience of individuals in the study.

Suppose that the linear component of a proportional hazards model contains p explanatory variables, $X_1, X_2, \ldots, X_p$, and that the estimated coefficients of these variables are $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$. The estimated hazard function for the ith of n individuals in the study is then given by

$$\hat{h}_i(t) = \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i)\hat{h}_0(t), \tag{3.14}$$

where $\boldsymbol{x}_i$ is the vector of values of the explanatory variables for the ith individual, $i = 1, 2, \ldots, n$, $\hat{\boldsymbol{\beta}}$ is the vector of estimated coefficients, and $\hat{h}_0(t)$ is the estimated baseline hazard function. Using this equation, the hazard function for an individual can be estimated once an estimate of $h_0(t)$ has been found. The relationship between the hazard, cumulative hazard and survivor functions can then be used to give estimates of the cumulative hazard function and the survivor function.

An estimate of the baseline hazard function was derived by Kalbfleisch and Prentice (1973) using an approach based on the method of maximum likelihood. Suppose that there are r distinct death times which, when arranged in increasing order, are $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$, and that there are d_j deaths and n_j individuals at risk at time $t_{(j)}$. The estimated baseline hazard function at time $t_{(j)}$ is then given by

$$\hat{h}_0(t_{(j)}) = 1 - \hat{\xi}_j, \tag{3.15}$$

where $\hat{\xi}_j$ is the solution of the equation

$$\sum_{l \in D(t_{(j)})} \frac{\exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l)}{1 - \hat{\xi}_j^{\exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l)}} = \sum_{l \in R(t_{(j)})} \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l), \tag{3.16}$$

for $j = 1, 2, \ldots, r$. In equation (3.16), $D(t_{(j)})$ is the set of all d_j individuals who die at the jth ordered death time, $t_{(j)}$, and as in Section 3.3, $R(t_{(j)})$ is the set of all n_j individuals at risk at time $t_{(j)}$. The estimates of the β's, which form the vector $\hat{\boldsymbol{\beta}}$, are those which maximise the likelihood function in equation (3.4). The derivation of this estimate of $h_0(t)$ is quite complex, and so it will not be reproduced here.

In the particular case where there are no tied death times, that is, where $d_j = 1$ for $j = 1, 2, \ldots, r$, the left-hand side of equation (3.16) will be a single term. This equation can then be solved to give

$$\hat{\xi}_j = \left(1 - \frac{\exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l)}\right)^{\exp(-\hat{\boldsymbol{\beta}}' \boldsymbol{x}_{(j)})},$$

where $\boldsymbol{x}_{(j)}$ is the vector of explanatory variables for the individual who dies at time $t_{(j)}$.

When there are tied observations, that is, when one or more of the d_j are greater than unity, the summation on the left-hand side of equation (3.16) is the sum of a series of fractions in which $\hat{\xi}_j$ occurs in the denominators, raised to different powers. Equation (3.16) cannot then be solved explicitly, and an iterative scheme is required.

We now make the assumption that the hazard of death is constant between adjacent death times. An appropriate estimate of the baseline hazard function in this interval is then obtained by dividing the estimated hazard in equation (3.15) by the time interval, to give the step function

$$\hat{h}_0(t) = \frac{1 - \hat{\xi}_j}{t_{(j+1)} - t_{(j)}}, \tag{3.17}$$

for $t_{(j)} \leqslant t < t_{(j+1)}$, $j = 1, 2, \ldots, r-1$, with $\hat{h}_0(t) = 0$ for $t < t_{(1)}$.

The quantity $\hat{\xi}_j$ can be regarded as an estimate of the probability that an individual survives through the interval from $t_{(j)}$ to $t_{(j+1)}$. The baseline survivor function can then be estimated by

$$\hat{S}_0(t) = \prod_{j=1}^{k} \hat{\xi}_j, \tag{3.18}$$

for $t_{(k)} \leqslant t < t_{(k+1)}$, $k = 1, 2, \ldots, r-1$, and so this estimate is also a step-function. The estimated value of the baseline survivor function is unity for $t < t_{(1)}$, and zero for $t \geqslant t_{(r)}$, unless there are censored survival times greater than $t_{(r)}$. If this is the case, $\hat{S}_0(t)$ can be taken to be $\hat{S}_0(t_{(r)})$ until the largest censored time, but the estimated survivor function is undefined beyond that time.

The baseline cumulative hazard function is, from equation (1.7), given by $H_0(t) = -\log S_0(t)$, and so an estimate of this function is

$$\hat{H}_0(t) = -\log \hat{S}_0(t) = -\sum_{j=1}^{k} \log \hat{\xi}_j, \tag{3.19}$$

for $t_{(k)} \leqslant t < t_{(k+1)}$, $k = 1, 2, \ldots, r-1$, with $\hat{H}_0(t) = 0$ for $t < t_{(1)}$.

The estimates of the baseline hazard, survivor and cumulative hazard functions in equations (3.17), (3.18) and (3.19) can be used to obtain the corresponding estimates for an individual with vector of explanatory variables $\boldsymbol{x}_i$. In particular, from equation (3.14), the hazard function is estimated by $\exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i)\hat{h}_0(t)$. Next, integrating both sides of equation (3.14), we get

$$\int_0^t \hat{h}_i(u)\,\mathrm{d}u = \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i) \int_0^t \hat{h}_0(u)\,\mathrm{d}u, \tag{3.20}$$

so that the estimated cumulative hazard function for the ith individual is