proach to inference is covered in many statistical texts. See, for example, Altman (1991) and Armitage *et al.* (2001) for non-technical presentations of the ideas in a medical context.

The log-rank test results from the work of Mantel and Haenszel (1959), Mantel (1966) and Peto and Peto (1972). See Lawless (2002) for details of the rank test formulation. A thorough review of the hypergeometric distribution, used in the derivation of the log-rank test in Section 2.6.2, is included in Johnson and Kotz (1969).

The log-rank test for trend is derived from the test for trend in a $2 \times k$ contingency table, given in Armitage *et al.* (2001). The test is also described by Altman (1991). Peto *et al.* (1976, 1977) give a non-mathematical account of the log-rank test and its extensions.

CHAPTER 3

Modelling survival data

The non-parametric methods described in Chapter 2 can be useful in the analysis of a single sample of survival data, or in the comparison of two or more groups of survival times. However, in most medical studies that give rise to survival data, supplementary information will also be recorded on each individual. A typical example would be a clinical trial to compare the survival times of patients who receive one or other of two treatments. In such a study, demographic variables such as the age and sex of the patient, the values of physiological variables such as serum haemoglobin level and heart rate, and factors that are associated with the lifestyle of the patient, such as smoking history and dietary habits, may all have an impact on the time that the patient survives. Accordingly, the values of these variables, which are referred to as *explanatory variables*, would be recorded at the outset of the study. The resulting data set would then be more complex than those considered in Chapter 2, and the methods described in that chapter would generally be unsuitable.

In order to explore the relationship between the survival experience of a patient and explanatory variables, an approach based on statistical modelling can be used. Indeed, the particular model that is developed in this chapter both unifies and extends the non-parametric procedures of Chapter 2.

3.1 Modelling the hazard function

Through a modelling approach to the analysis of survival data, we can explore how the survival experience of a group of patients depends on the values of one or more explanatory variables, whose values have been recorded for each patient at the time origin. For example, in the study on multiple myeloma, given as Example 1.3, the aim is to determine which of seven explanatory variables have an impact on the survival time of the patients. In Example 1.4 on the survival times of patients in a clinical trial involving two treatments for prostatic cancer, the primary aim is to identify whether patients in the two treatment groups have a different survival experience. Because additional variables such as the age of the patient and the size of their tumour are likely to influence survival time, it will be important to take account of these variables when assessing the extent of any treatment difference.

In the analysis of survival data, interest centres on the risk or hazard of death at any time after the time origin of the study. As a consequence, the hazard function is modelled directly in survival analysis. The resulting models are somewhat different in form from linear models encountered in regression analysis and in the analysis of data from designed experiments, where the dependence of the mean response, or some function of it, on certain explanatory variables is modelled. However, many of the principles and procedures used in linear modelling carry over to the modelling of survival data.

There are two broad reasons for modelling survival data. One objective of the modelling process is to determine which combination of potential explanatory variables affect the form of the hazard function. In particular, the effect that the treatment has on the hazard of death can be studied, as can the extent to which other explanatory variables affect the hazard function. Another reason for modelling the hazard function is to obtain an estimate of the hazard function itself for an individual. This may be of interest in its own right, but in addition, from the relationship between the survivor function and hazard function described by equation (1.5), an estimate of the survivor function can be found. This will in turn lead to an estimate of quantities such as the median survival time, which will be a function of the explanatory variables in the model. The median survival time could then be estimated for current or future patients with particular values of these explanatory variables. The resulting estimate could be particularly useful in devising a treatment regimen, or in counselling the patient about their prognosis.

The basic model for survival data to be considered in this chapter is the *proportional hazards model*. This model was proposed by Cox (1972) and has also come to be known as the *Cox regression model*. Although the model is based on the assumption of proportional hazards, introduced in Section 2.6.4, no particular form of probability distribution is assumed for the survival times. The model is therefore referred to as a *semi-parametric model*. We now go on to develop the model for the comparison of the hazard functions for individuals in two groups.

3.1.1 A model for the comparison of two groups

Suppose that patients are randomised to receive either a standard treatment or a new treatment, and let $h_S(t)$ and $h_N(t)$ be the hazards of death at time t for patients on the standard treatment and new treatment, respectively. According to a simple model for the survival times of the two groups of patients, the hazard at time t for a patient on the new treatment is proportional to the hazard at that same time for a patient on the standard treatment. This *proportional hazards model* can be expressed in the form

$$h_N(t) = \psi h_S(t), \tag{3.1}$$

for any non-negative value of t, where ψ is a constant. An implication of this assumption is that the corresponding true survivor functions for individuals on the new and standard treatments do not cross, as previously shown in Section 2.6.4.

The value of ψ is the ratio of the hazards of death at any time for an individual on the new treatment relative to an individual on the standard treatment, and so ψ is known as the *relative hazard* or *hazard ratio*. If $\psi < 1$, the hazard of death at t is smaller for an individual on the new drug, relative to an individual on the standard. The new treatment is then an improvement on the standard. On the other hand, if $\psi > 1$, the hazard of death at t is greater for an individual on the new drug, and the standard treatment is superior.

An alternative way of expressing the model in equation (3.1) leads to a model that can more easily be generalised. Suppose that survival data are available on n individuals and denote the hazard function for the ith of these by $h_i(t)$, $i = 1, 2, \ldots, n$. Also, write $h_0(t)$ for the hazard function for an individual on the standard treatment. The hazard function for an individual on the new treatment is then $\psi h_0(t)$. The relative hazard ψ cannot be negative, and so it is convenient to set $\psi = \exp(\beta)$. The parameter β is then the logarithm of the hazard ratio, that is, $\beta = \log \psi$, and any value of β in the range $(-\infty, \infty)$ will lead to a positive value of ψ. Note that positive values of β are obtained when the hazard ratio, ψ, is greater than unity, that is, when the new treatment is inferior to the standard.

Now let X be an *indicator variable*, which takes the value zero if an individual is on the standard drug, and unity if an individual is on the new drug. If x_i is the value of X for the ith individual in the study, $i = 1, 2, \ldots, n$, the hazard function for this individual can be written as

$$h_i(t) = e^{\beta x_i} h_0(t), \tag{3.2}$$

where $x_i = 1$ if the ith individual is on the new treatment and $x_i = 0$ otherwise. This is the proportional hazards model for the comparison of two treatment groups.

3.1.2 The general proportional hazards model

The model of the previous section is now generalised to the situation where the hazard of death at a particular time depends on the values $x_1, x_2, \ldots, x_p$ of p explanatory variables, $X_1, X_2, \ldots, X_p$. The values of these variables will be assumed to have been recorded at the time origin of the study. An extension of the model to cover the situation where the values of one or more of the explanatory variables change over time will be considered in Chapter 8.

The set of values of the explanatory variables in the proportional hazards model will be represented by the vector $\boldsymbol{x}$, so that $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)'$. Let $h_0(t)$ be the hazard function for an individual for whom the values of all the explanatory variables that make up the vector $\boldsymbol{x}$ are zero. The function $h_0(t)$ is called the *baseline hazard function*. The hazard function for the ith individual can then be written as

$$h_i(t) = \psi(\boldsymbol{x}_i) h_0(t),$$

where $\psi(\boldsymbol{x}_i)$ is a function of the values of the vector of explanatory variables for the ith individual. The function $\psi(\cdot)$ can be interpreted as the hazard at time t for an individual whose vector of explanatory variables is $\boldsymbol{x}_i$, relative to the hazard for an individual for whom $\boldsymbol{x} = \boldsymbol{0}$.

Again, since the relative hazard, $\psi(\boldsymbol{x}_i)$, cannot be negative, it is convenient to write this as $\exp(\eta_i)$, where η_i is a linear combination of the p explanatory variables in $\boldsymbol{x}_i$. Therefore,

$$\eta_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi},$$

so that $\eta_i = \sum_{j=1}^{p} \beta_j x_{ji}$. In matrix notation, $\eta_i = \boldsymbol{\beta}' \boldsymbol{x}_i$, where $\boldsymbol{\beta}$ is the vector of coefficients of the explanatory variables $x_1, x_2, \ldots, x_p$ in the model. The quantity η_i is called the *linear component* of the model, but it is also known as the *risk score* or *prognostic index* for the ith individual. The general proportional hazards model then becomes

$$h_i(t) = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}) h_0(t). \tag{3.3}$$

Since this model can be re-expressed in the form

$$\log \left\{ \frac{h_i(t)}{h_0(t)} \right\} = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi},$$

the proportional hazards model may also be regarded as a linear model for the logarithm of the hazard ratio. There are other possible forms for $\psi(\boldsymbol{x}_i)$, but the choice $\psi(\boldsymbol{x}_i) = \exp(\boldsymbol{\beta}' \boldsymbol{x}_i)$ leads to the most commonly used model for survival data.

Notice that there is no constant term in the linear component of the proportional hazards model. If a constant term β_0, say, were included, the baseline hazard function could simply be rescaled by dividing $h_0(t)$ by $\exp(\beta_0)$, and the constant term would cancel out. Moreover, we have made no assumptions concerning the actual form of the baseline hazard function $h_0(t)$. Indeed, we will see later that the β-coefficients in this proportional hazards model can be estimated without making any such assumptions. Of course, we will often need to estimate $h_0(t)$ itself, and we will see how this can be done in Section 3.8.

3.2 The linear component of the proportional hazards model

There are two types of variable on which a hazard function may depend, namely *variates* and *factors*. A variate is a variable that takes numerical values that are often on a continuous scale of measurement, such as age or systolic blood pressure. A factor is a variable that takes a limited set of values, which are known as the *levels* of the factor. For example, sex is a factor with two levels, and type of tumour might be a factor whose levels correspond to different histologies, such as squamous, adeno or small cell.

We now consider how variates, factors, and terms that combine factors and variates, can be incorporated in the linear component of a proportional hazards model.

3.2.1 Including a variate

Variates, either alone or in combination, are readily incorporated in a proportional hazards model. Each variate appears in the model with a corresponding β-coefficient. As an illustration, consider a situation in which the hazard function depends on two variates X_1 and X_2. The value of these variates for the ith individual will be x_{1i} and x_{2i}, respectively, and the proportional hazards model for the ith of n individuals is written as

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

In models such as this, the baseline hazard function, $h_0(t)$, is the hazard function for an individual for whom all the variates included in the model take the value zero.

3.2.2 Including a factor

Suppose that the dependence of the hazard function on a single factor, A, is to be modelled, where A has a levels. The model for an individual for whom the level of A is j will then need to incorporate the term α_j which represents the effect due to the jth level of the factor. The terms $\alpha_1, \alpha_2, \ldots, \alpha_a$ are known as the *main effects* of the factor A. According to the proportional hazards model, the hazard function for an individual with factor A at level j is $\exp(\alpha_j) h_0(t)$. Now, the baseline hazard function $h_0(t)$ has been defined to be the hazard for an individual with values of all explanatory variables equal to zero. To be consistent with this definition, one of the α_j must be taken to be zero. One possibility is to adopt the constraint $\alpha_1 = 0$, which corresponds to taking the baseline hazard to be the hazard for an individual for whom A is at the first level. This is the constraint that will be used in the sequel.

Models that contain terms corresponding to factors can be expressed as linear combinations of explanatory variables by defining *indicator* or *dummy variables* for each factor. This procedure will be required when using computer software for survival analysis that does not allow factors to be fitted directly. If the constraint $\alpha_1 = 0$ is adopted, the term α_j can be included in the model by defining $a - 1$ indicator variables, $X_2, X_3, \ldots, X_a$, that take the values shown in the table below.

Level of A	X_2	X_3	. . .	X_a
1	0	0	. . .	0
2	1	0	. . .	0
3	0	1	. . .	0
. . .	. . .	. . .	. . .	. . .
a	0	0	. . .	1

The term α_j can then be incorporated in the linear part of the proportional hazards model by including the $a - 1$ explanatory variables $X_2, X_3, \ldots, X_a$ with coefficients $\alpha_2, \alpha_3, \ldots, \alpha_a$. In other words, the term α_j in the model is replaced by $\alpha_2 x_2 + \alpha_3 x_3 + \cdots + \alpha_a x_a$, where x_j is the value of X_j for an individual for whom A is at level j, $j = 2, 3, \ldots, a$. There are then $a - 1$

parameters associated with the main effect of the factor A, and A is said to have $a-1$ *degrees of freedom*.

3.2.3 Including an interaction

When terms corresponding to more than one factor are to be included in the model, sets of indicator variables can be defined for each factor in a manner similar to that shown above. In this situation, it may also be appropriate to include a term in the model that corresponds to individual effects for each combination of levels of two or more factors. Such effects are known as *interactions*.

For example, suppose that the two factors are the sex of a patient and grade of tumour. If the effect of grade of tumour on the hazard of death is different in patients of each sex, we would say that there is an interaction between these two factors. The hazard function would then depend on the combination of levels of these two factors.

In general, if A and B are two factors, and the hazard of death depends on the combination of levels of A and B, then A and B are said to *interact*. If A and B have a and b levels, respectively, the term that represents an interaction between these two factors is denoted by $(\alpha\beta)_{jk}$, for $j = 1, 2, \ldots, a$ and $k = 1, 2, \ldots, b$.

In statistical modelling, the effect of an interaction can only be investigated by adding the interaction term to a model that already contains the corresponding main effects. If either α_j or β_k are excluded from the model, the term $(\alpha\beta)_{jk}$ represents the effect of one factor *nested* within the other. For example, if α_j is included in the model, but not β_k, then $(\alpha\beta)_{jk}$ is the effect of B nested within A. If both α_j and β_k are excluded, the term $(\alpha\beta)_{jk}$ represents the effect of the combination of level i of A and level j of B on the response variable. This means that $(\alpha\beta)_{jk}$ can only be interpreted as an interaction effect when included in a model that contains both α_j and β_k, which correspond to the main effects of A and B. We will return to this point when we consider model-building strategy in Section 3.5.

In order to include the term $(\alpha\beta)_{jk}$ in the model, products of indicator variables associated with the main effects are calculated. For example, if A and B have 2 and 3 levels respectively, indicator variables U_2 and V_2, V_3 are defined as in the following tables.

Level of A	U_2
1	0
2	1

Level of B	V_2	V_3
1	0	0
2	1	0
3	0	1

Let u_j and v_k be the values of U_j and V_k for a given individual, for $j = 2$, $k = 2, 3$. The term $(\alpha\beta)_{jk}$ is then fitted by including variates formed from the products of U_j and V_k in the model. The corresponding value of the product for a given individual is $u_j v_k$. The coefficient of this product is denoted $(\alpha\beta)_{jk}$, and so the term $(\alpha\beta)_{jk}$ is fitted as

$$(\alpha\beta)_{22}u_2v_2 + (\alpha\beta)_{23}u_2v_3.$$

There are therefore two parameters associated with the interaction between A and B. In general, if A and B have a and b levels, respectively, the two-factor interaction AB has $(a-1)(b-1)$ parameters associated with it, in other words AB has $(a-1)(b-1)$ degrees of freedom. Furthermore, the term $(\alpha\beta)_{jk}$ is equal to zero whenever either A or B are at the first level, that is, when either $j = 1$ or $k = 1$.

3.2.4 Including a mixed term

Another type of term that might be needed in a model is a mixed term formed from a factor and a variate. Terms of this type would be used when the coefficient of a variate in a model was likely to be different for each level of a factor. For example, consider a contraceptive trial in which the time to the onset of a period of amenorrhoea, the prolonged absence of menstrual bleeding, is being modelled. The hazard of an amenorrhoea may be related to the weight of a woman, but the coefficient of this variate may differ according to the level of a factor associated with the number of previous pregnancies that the woman has experienced.

The dependence of the coefficient of a variate, X, on the level of a factor, A, would be depicted by including the term $\alpha_j x$ in the linear component of the proportional hazards model, where x is the value of X for a given individual for whom the factor A is at the jth level, $j = 1, 2, \ldots, a$. To include such a term, indicator variables U_j, say, are defined for the factor A, and each of these is multiplied by the value of X for each individual. The resulting values of the products U_jX are u_jx, and the coefficient of u_jx in the model is α_j, where j indexes the level of the factor A.

If the same definition of indicator variables in the previous discussion were used, α_1, the coefficient of X for individuals at the first level of A, would be zero. It is then essential to include the variate X in the model as well as the products, for otherwise the dependence on X for individuals at the first level of A would not be modelled. An illustration should make this clearer.

Suppose that there are nine individuals in a study, on each of whom the value of a variate, X, and the level of a factor, A, have been recorded. We will take A to have three levels, where A is at the first level for the first three individuals, at the second level for the next three, and at the third level for the final three. In order to model the dependence of the coefficient of the variate X on the level of A, two indicator variables, U_2 and U_3 are defined as in the following table.

Individual	Level of A	X	U_2	U_3	U_2X	U_3X
1	1	x_1	0	0	0	0
2	1	x_2	0	0	0	0
3	1	x_3	0	0	0	0
4	2	x_4	1	0	x_4	0
5	2	x_5	1	0	x_5	0
6	2	x_6	1	0	x_6	0
7	3	x_7	0	1	0	x_7
8	3	x_8	0	1	0	x_8
9	3	x_9	0	1	0	x_9

Explanatory variables formed as the products U_2X and U_3X, given in the last two columns of this table, would then be included in the linear component of the model, together with the variate X. Let the coefficients of the values of the products U_2X and U_3X be α_2' and α_3', respectively, and let the coefficient of the value of the variate X in the model be β. Then, the model contains the terms $\beta x + \alpha_2'(u_2 x) + \alpha_3'(u_3 x)$. From the above table, $u_2 = 0$ and $u_3 = 0$ for individuals at level 1 of A, and so the coefficient of x for these individuals is just β. For those at level 2 of A, $u_2 = 1$ and $u_3 = 0$, and the coefficient of x is $\beta + \alpha_2'$. Similarly, at level 3 of A, $u_2 = 0$ and $u_3 = 1$, and the coefficient of x is $\beta + \alpha_3'$.

Notice that if the term βx is omitted from the model, the coefficient of x for individuals 1, 2 and 3 would be zero. There would then be no information about the relationship between the hazard function and the variate X for individuals at the first level of the factor A.

The manipulation described in the preceding paragraphs can be avoided by defining the indicator variables in a different way. If a factor A has a levels, and it is desired to include the term $\alpha_j x$ in a model, without necessarily including the term βx, a indicator variables $Z_1, Z_2, \ldots, Z_a$ can be defined for A, where $Z_j = 1$ at level j of A and zero otherwise. The corresponding values of these products for an individual, $z_1x, z_2x, \ldots, z_ax$, are included in the model with coefficients $\alpha_1, \alpha_2, \ldots, \alpha_a$. These are the coefficients of x for each level of A.

Now, if the variate X is included in the model, along with the a products of the form Z_jX, there will be $a + 1$ terms corresponding to the a coefficients. It will not then be possible to obtain unique estimates of each of these α-coefficients, and the model is said to be *overparameterised*. This overparameterisation can be dealt with by forcing one of the $a + 1$ coefficients to be zero. In particular, taking $\alpha_1 = 0$ would be equivalent to a redefinition of the indicator variables, in which Z_1 is taken to be zero. This then leads to the same formulation of the model that has already been discussed.

The application of these ideas in the analysis of actual data sets will be illustrated in Section 3.4, after we have seen how the proportional hazards model can be fitted.

3.3 Fitting the proportional hazards model

Fitting the proportional hazards model given in equation (3.3) to an observed set of survival data entails estimating the unknown coefficients of the explanatory variables, $X_1, X_2, \ldots, X_p$, in the linear component of the model, $\beta_1, \beta_2, \ldots, \beta_p$. The baseline hazard function, $h_0(t)$, may also need to be estimated. It turns out that these two components of the model can be estimated separately. The β's are estimated first and these estimates are then used to construct an estimate of the baseline hazard function. This is an important result, since it means that in order to make inferences about the effects of p explanatory variables, $X_1, X_2, \ldots, X_p$, on the relative hazard, $h_i(t)/h_0(t)$, we do not need an estimate of $h_0(t)$. Methods for estimating $h_0(t)$ will therefore be deferred until Section 3.8.

The β-coefficients in the proportional hazards model, which are the unknown parameters in the model, can be estimated using the *method of maximum likelihood*. To operate this method, we first obtain the *likelihood* of the sample data. This is the joint probability of the observed data, regarded as a function of the unknown parameters in the assumed model. For the proportional hazards model, this is a function of the observed survival times and the unknown β-parameters in the linear component of the model. Estimates of the β's are then those values that are the most likely on the basis of the observed data. These *maximum likelihood estimates* are therefore the values that maximise the likelihood function. From a computational viewpoint, it is more convenient to maximise the logarithm of the likelihood function. Furthermore, approximations to the variance of maximum likelihood estimates can be obtained from the second derivatives of the log-likelihood function. Details will not be given here, but Appendix A contains a summary of relevant results from the theory of maximum likelihood estimation.

Suppose that data are available for n individuals, among whom there are r distinct death times and $n - r$ right-censored survival times. We will for the moment assume that only one individual dies at each death time, so that there are no *ties* in the data. The treatment of ties will be discussed in Section 3.3.2. The r ordered death times will be denoted by $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$, so that $t_{(j)}$ is the jth ordered death time. The set of individuals who are at risk at time $t_{(j)}$ will be denoted by $R(t_{(j)})$, so that $R(t_{(j)})$ is the group of individuals who are alive and uncensored at a time just prior to $t_{(j)}$. The quantity $R(t_{(j)})$ is called the *risk set*.

Cox (1972) showed that the relevant likelihood function for the proportional hazards model in equation (3.3) is given by

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}'\boldsymbol{x}_l)}, \tag{3.4}$$

in which $\boldsymbol{x}_{(j)}$ is the vector of covariates for the individual who dies at the jth ordered death time, $t_{(j)}$. The summation in the denominator of this likelihood function is the sum of the values of $\exp(\boldsymbol{\beta}'\boldsymbol{x})$ over all individuals who are at risk at time $t_{(j)}$. Notice that the product is taken over the individuals for whom