

Texts in Statistical Science

# Modelling Survival Data in Medical Research

SECOND EDITION

**David Collett**



CHAPMAN & HALL/CRC

---

# Contents

---

## Preface to the second edition

## Preface to the first edition

<b>1</b>	<b>Survival analysis</b>	<b>1</b>
1.1	Special features of survival data	1
1.2	Some examples	5
1.3	Survivor function and hazard function	11
1.4	Further reading	13
<b>2</b>	<b>Some non-parametric procedures</b>	<b>15</b>
2.1	Estimating the survivor function	15
2.2	Standard error of the estimated survivor function	23
2.3	Estimating the hazard function	29
2.4	Estimating the median and percentiles of survival times	33
2.5	Confidence intervals for the median and percentiles	35
2.6	Comparison of two groups of survival data	37
2.7	Comparison of three or more groups of survival data	48
2.8	Stratified tests	49
2.9	Log-rank test for trend	51
2.10	Further reading	53
<b>3</b>	<b>Modelling survival data</b>	<b>55</b>
3.1	Modelling the hazard function	55
3.2	The linear component of the proportional hazards model	58
3.3	Fitting the proportional hazards model	63
3.4	Confidence intervals and hypothesis tests for the $\beta$ 's	69
3.5	Comparing alternative models	73
3.6	Strategy for model selection	80
3.7	Interpretation of parameter estimates	89
3.8	Estimating the hazard and survivor functions	97
3.9	Proportional hazards modelling and the log-rank test	106
3.10	Further reading	109

<b>4 Model checking in the Cox regression model</b>	<b>111</b>
4.1 Residuals for the Cox regression model	111
4.2 Assessment of model fit	121
4.3 Identification of influential observations	131
4.4 Testing the assumption of proportional hazards	141
4.5 Recommendations	148
4.6 Further reading	149
<b>5 Parametric proportional hazards models</b>	<b>151</b>
5.1 Models for the hazard function	151
5.2 Assessing the suitability of a parametric model	155
5.3 Fitting a parametric model to a single sample	158
5.4 A model for the comparison of two groups	168
5.5 The Weibull proportional hazards model	175
5.6 Comparing alternative Weibull models	183
5.7 The Gompertz proportional hazards model	190
5.8 Model choice	192
5.9 Further reading	193
<b>6 Accelerated failure time and other parametric models</b>	<b>195</b>
6.1 Probability distributions for survival data	195
6.2 Exploratory analyses	199
6.3 The accelerated failure time model for comparing two groups	200
6.4 The general accelerated failure time model	206
6.5 Parametric accelerated failure time models	209
6.6 Fitting and comparing accelerated failure time models	216
6.7 The proportional odds model	223
6.8 Some other distributions for survival data	227
6.9 Further reading	228
<b>7 Model checking in parametric models</b>	<b>231</b>
7.1 Residuals for parametric models	231
7.2 Residuals for particular parametric models	234
7.3 Comparing observed and fitted survivor functions	240
7.4 Identification of influential observations	242
7.5 Testing proportional hazards in the Weibull model	247
7.6 Further reading	248
<b>8 Time-dependent variables</b>	<b>251</b>
8.1 Types of time-dependent variables	251
8.2 A model with time-dependent variables	252
8.3 Model comparison and validation	258
8.4 Some applications of time-dependent variables	260
8.5 Three examples	262
8.6 Further reading	271

<b>9 Interval-censored survival data</b>	<b>273</b>
9.1 Modelling interval-censored survival data	273
9.2 Modelling the recurrence probability in the follow-up period	276
9.3 Modelling the recurrence probability at different times	279
9.4 Arbitrarily interval-censored survival data	286
9.5 Parametric models for interval-censored data	296
9.6 Discussion	297
9.7 Further reading	297

<b>10 Sample size requirements for a survival study</b>	<b>299</b>
10.1 Distinguishing between two treatment groups	299
10.2 Calculating the required number of deaths	300
10.3 Calculating the required number of patients	306
10.4 Further reading	311

<b>11 Some additional topics</b>	<b>313</b>
11.1 Non-proportional hazards	313
11.2 Informative censoring	318
11.3 Frailty models	320
11.4 Multistate models	323
11.5 Effect of covariate adjustment	327
11.6 Measures of explained variation	328
11.7 Modelling a cure probability	329
11.8 Some other designs in survival analysis	329

---

## Preface to the second edition

---

The aim of this book is to describe and illustrate the modelling approach to the analysis of survival data. This new edition is designed to reflect current statistical practice and take proper account of the widespread availability of computer software for survival analysis. The material in each chapter has been substantially revised and updated, and there has been some reorganisation of the material on parametric models and model checking. The opportunity has also been taken to provide a more detailed treatment of some topics, most notably accelerated failure time models and the analysis of interval-censored data. Over 80 references have been added to the sections at the end of each chapter that give suggestions for further reading.

I hope that this comprehensive practical account of statistical methods for use in modelling survival data will continue to meet the needs of statisticians in the pharmaceutical industry and medical research institutes. Much of the book should also be accessible to scientists who are analysing their own data, with or without the support of a statistician. The text is also designed for use by students following undergraduate and postgraduate courses in survival analysis.

Although it is anticipated that computer software will be used for the analysis of survival data, sufficient methodological details have been given to provide a sound understanding of the nature of the techniques, and corresponding computer output. The more technical sections continue to be indicated with an asterisk.

The main part of the book is formed by Chapters 1 to 7. After an introduction to survival analysis in Chapter 1, Chapter 2 describes methods for summarising survival data, and for comparing two or more groups of survival times. The modelling approach is introduced in Chapter 3, where the Cox proportional hazards model is presented in detail. Since model checking is such an important part of the modelling process, this chapter is followed by a description of methods for checking the adequacy of a fitted Cox regression model. Parametric proportional hazards models are covered in Chapter 5, with an emphasis on the Weibull model for survival data. In Chapter 6, parametric accelerated failure time models are described, and a detailed account is given of the log-linear representation of such models, used in most computer software for parametric modelling. Model-checking diagnostics for parametric models are then presented in Chapter 7.

The remaining chapters describe a number of extensions to the basic models. The use of time-dependent variables is covered in Chapter 8, which now

incorporates material on estimating the baseline functions in the presence of time-dependent variables. The analysis of interval-censored data is considered in Chapter 9, which gives details of how the Cox regression model, and fully parametric models, can be used in modelling arbitrarily interval-censored data. This is followed by a chapter on sample size requirements for survival studies. Chapter 11 provides a brief introduction to a number of additional topics, including non-proportional hazards, informative censoring, and frailty modelling.

Since the first edition was published, computer software for survival analysis has become a standard feature of many statistical packages. It is no longer feasible to review all of these, and so Chapter 12 focuses on the use of SAS. The output from this package is examined in detail, and the methods used by the software to generate the output are summarised. Because output from many other packages is quite similar, users of these packages should be able to reconcile their output with that of SAS, and the many illustrative examples in this book.

Some additional data sets that may be used to obtain a fuller appreciation of the methodology, or as student exercises, are given in Appendix D. All of the data sets used in this book are available in electronic form from the publisher's web site at

[www.crcpress.com/e\\_products/downloads/](http://www.crcpress.com/e_products/downloads/)

I am very grateful to all who took the trouble to let me know about errors and ambiguities in the first edition. While these have been corrected, the revision process may have led to the introduction of other errors. Naturally, I would be very pleased to be informed of any that are detected.

Dave Collett  
January 2003

## Preface to the first edition

In the course of medical research, data on the time to the occurrence of a particular event, such as the death of a patient, are frequently encountered. Such data are generically referred to as survival data. However, the event of interest need not necessarily be death, but could, for example, be the end of a period spent in remission from a disease, relief from symptoms, or the recurrence of a particular condition. Although there are a number of books devoted to the analysis of survival data, this book is designed to meet the need for an intermediate text that emphasises the application of the methodology to survival data arising from medical studies, which shows how widely available computer software can be used in survival analysis, and which will appeal to statisticians engaged in medical research.

This book is based on a course on the analysis of survival data from clinical trials that has been given annually by the Statistical Services Centre of the University of Reading, since 1986. Although it is written primarily for those working as statisticians in the pharmaceutical industry and in medical research institutes, much of the text should be accessible to numerate scientists and clinicians working alongside statisticians on the analysis of their own data sets. This book could also be used as a text to accompany undergraduate and postgraduate courses on survival analysis in universities and other institutes of higher education.

Many illustrative examples have been included in the text. In addition, sufficient methodological development is given to enable the reader to understand the assumptions on which particular techniques are based, and to help in adapting the methodology to deal with non-standard problems. A number of data sets are based on fewer observations than would normally be encountered in medical research programmes. This enables certain methods of analysis to be illustrated more easily, and means that tabular presentations of results are not too unwieldy. Naturally, the methods described in this book can be applied without modification to larger data sets.

The book begins with an introduction to survival analysis, and a description of four studies in which survival data were obtained. These data sets, and others besides, are then used to illustrate the techniques for analysing survival data presented in subsequent chapters. In Chapter 2, some methods for summarising survival data are introduced, and non-parametric methods for comparing the survival times of patients in two or more treatment groups are described.

A modelling approach to the analysis of survival data, based on the Cox proportional hazards model, is presented in Chapter 3. Models which assume

a Weibull distribution for survival times are developed in Chapter 4, and Chapter 5 gives a comprehensive account of diagnostics that can be used to check the adequacy of both the Cox and Weibull proportional hazards models. Some other parametric models for survival data, including the accelerated failure time model and the proportional odds model, are described in Chapter 6. This is followed by a chapter that shows how variables whose values change over time can be incorporated in models for survival data.

When the survival times of patients are not known exactly, methods used to analyse interval-censored data may be appropriate, and these are described in Chapter 8. The important issue of sample size requirements for a survival study is considered in Chapter 9, and this is followed by a chapter that contains a brief discussion of some additional topics in the analysis of survival data.

In order to implement many of the techniques for analysing survival data, appropriate computer software is needed. Accordingly, the final chapter of the book contains details on the use of some widely available statistical packages for survival analysis, particularly SAS, BMDP and SPSS. In this chapter, the facilities for survival analysis in these three packages are summarised, and illustrated using a particular data set.

Bibliographic notes and suggestions for further reading are given at the end of each chapter, but so as not to interrupt the flow, references in the text itself will be kept to a minimum. Some sections contain more mathematical details than others, and these have been denoted with an asterisk. These sections can be omitted without loss of continuity.

In writing this book, I have assumed that the reader has a basic knowledge of statistical methods, and has some familiarity with topics such as linear regression analysis and analysis of variance. Matrix algebra is used to give an expression for the standard error of a percentile of the Weibull proportional hazards model in Chapter 4 and Appendix B, and to express some diagnostics for model checking in Chapter 5. However, an understanding of matrices is not an essential requirement.

I have received help from a number of colleagues while writing this book. Mike Patefield and Anne Whitehead provided constructive comments on initial drafts of some of the chapters, and Marilyn Collins provided valuable assistance in writing the SAS macros described in Chapter 11. I would also like to thank Doug Altman for his many comments and suggestions on the first six chapters of the book. I owe a particular debt of gratitude to John Whitehead for many helpful discussions and for his comments on several draft chapters. In addition, the chapter on interval-censored survival data, and sections in some of the other chapters, are based heavily on material prepared by John for courses that he has given on survival analysis. However, I take full responsibility for any errors in the text. Finally, I would again like to thank my wife Janet for her support and encouragement over the period that this book was written.

Dave Collett  
October 1993

## CHAPTER 1

# Survival analysis

Survival analysis is the phrase used to describe the analysis of data in the form of times from a well-defined *time origin* until the occurrence of some particular event or *end-point*. In medical research, the time origin will often correspond to the recruitment of an individual into an experimental study, such as a clinical trial to compare two or more treatments. This in turn may coincide with the diagnosis of a particular condition, the commencement of a treatment regimen, or the occurrence of some adverse event. If the end-point is the death of a patient, the resulting data are literally survival times. However, data of a similar form can be obtained when the end-point is not fatal, such as the relief of pain, or the recurrence of symptoms. In this case, the observations are often referred to as *time to event* data. The methods for analysing survival data that are presented in this book can be used when the response variable is literally a survival time, but apply equally to data on the time to other end-points. The methodology can also be applied to data from other application areas, such as the survival times of animals in an experimental study, the time taken by an individual to complete a task in a psychological experiment, the storage times of seeds held in a seed bank, or the lifetimes of industrial or electronic components. The focus of this book is on the application of survival analysis to data arising from medical research, and for this reason much of the general discussion will be phrased in terms of the survival time of an individual patient from entry to a study until death.

### 1.1 Special features of survival data

We must first consider the reasons why survival data are not amenable to standard statistical procedures used in data analysis. One reason is that survival data are generally not symmetrically distributed. Typically, a histogram constructed from the survival times of a group of similar individuals will tend to be *positively skewed*, that is, the histogram will have a longer "tail" to the right of the interval that contains the largest number of observations. As a consequence, it will not be reasonable to assume that data of this type have a normal distribution. This difficulty could be resolved by first transforming the data to give a more symmetric distribution, for example by taking logarithms. However, a more satisfactory approach is to adopt an alternative distributional model for the original data.

The main feature of survival data that renders standard methods inappropriate is that survival times are frequently *censored*. The survival time of an

individual is said to be censored when the end-point of interest has not been observed for that individual. This may be because the data from a study are to be analysed at a point in time when some individuals are still alive. Alternatively, the survival status of an individual at the time of the analysis might not be known because that individual has been *lost to follow-up*. As an example, suppose that after being recruited to a clinical trial, a patient moves to another part of the country, or to a different country, and can no longer be traced. The only information available on the survival experience of that patient is the last date on which he or she was known to be alive. This date may well be the last time that the patient reported to a clinic for a regular check-up.

An actual survival time can also be regarded as censored when death is from a cause that is known to be unrelated to the treatment. However, it can be difficult to be sure that the death is not related to a particular treatment that the patient is receiving. For example, consider a patient in a clinical trial to compare alternative therapies for prostatic cancer who experiences a fatal road traffic accident. The accident could have resulted from an attack of dizziness, which might be a side effect of the treatment to which that patient has been assigned. If so, the death is not unrelated to the treatment. In circumstances such as these, the survival time until death from all causes, or the time to death from causes other than the primary condition for which the patient is being treated, might also be subjected to a survival analysis.

In each of these situations, a patient who entered a study at time  $t_0$  dies at time  $t_0 + t$ . However,  $t$  is unknown, either because the individual is still alive or because he or she has been lost to follow-up. If the individual was last known to be alive at time  $t_0 + c$ , the time  $c$  is called a censored survival time. This censoring occurs after the individual has been entered into a study, that is, to the right of the last known survival time, and is therefore known as *right censoring*. The right-censored survival time is then less than the actual, but unknown, survival time.

Another form of censoring is *left censoring*, which is encountered when the actual survival time of an individual is less than that observed. To illustrate this form of censoring, consider a study in which interest centres on the time to recurrence of a particular cancer following surgical removal of the primary tumour. Three months after their operation, the patients are examined to determine if the cancer has recurred. At this time, some of the patients may be found to have a recurrence. For such patients, the actual time to recurrence is less than three months, and the recurrence times of these patients is left-censored. Left censoring occurs far less commonly than right censoring, and so the emphasis of this book will be on the analysis of right-censored survival data.

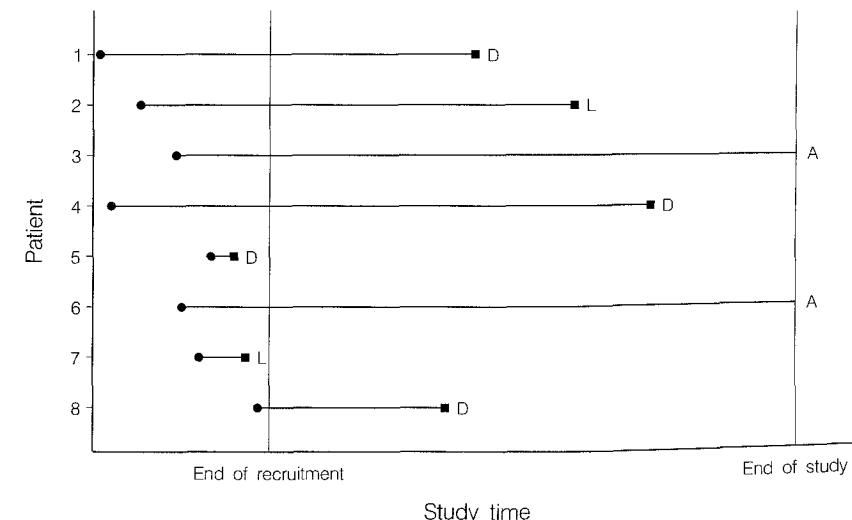
Yet another type of censoring is *interval censoring*. Here, individuals are known to have experienced an event within an interval of time. Consider again the example concerning the time to recurrence of a tumour used in the above discussion of left censoring. If a patient is observed to be free of the disease at three months, but is found to have had a recurrence when examined

six months after surgery, the actual recurrence time of that patient is known to be between three months and six months. The observed recurrence time is then said to be interval-censored. We will return to interval censoring later, in Chapter 9.

### 1.1.1 Patient time and study time

In a typical study, patients are not all recruited at exactly the same time, but accrue over a period of months or even years. After recruitment, patients are followed up until they die, or until a point in calendar time that marks the end of the study, when the data are analysed. Although the actual survival times will be observed for a number of patients, after recruitment some patients may be lost to follow-up, while others will still be alive at the end of the study. The calendar time period in which an individual is in the study is known as the *study time*.

The study time for eight individuals in a clinical trial is illustrated diagrammatically in Figure 1.1, in which the time of entry to the study is represented by a “•”. Individuals 1, 4, 5 and 8 die (D) during the course of the study, individuals 2 and 7 are lost to follow-up (L), and individuals 3 and 6 are still alive (A) at the end of the observation period.



**Figure 1.1** Study time for eight patients in a survival study.

As far as each patient is concerned, the trial begins at some time  $t_0$ . The corresponding survival times for the eight individuals depicted in Figure 1.1 are shown in order in Figure 1.2. The period of time that a patient spends in the study, measured from that patient's time origin, is often referred to as *patient time*. The period of time from the time origin to the death of a patient (D) is then the survival time, and this is recorded for individuals 1, 4,

5 and 8. The survival times of the remaining individuals are right-censored (C).

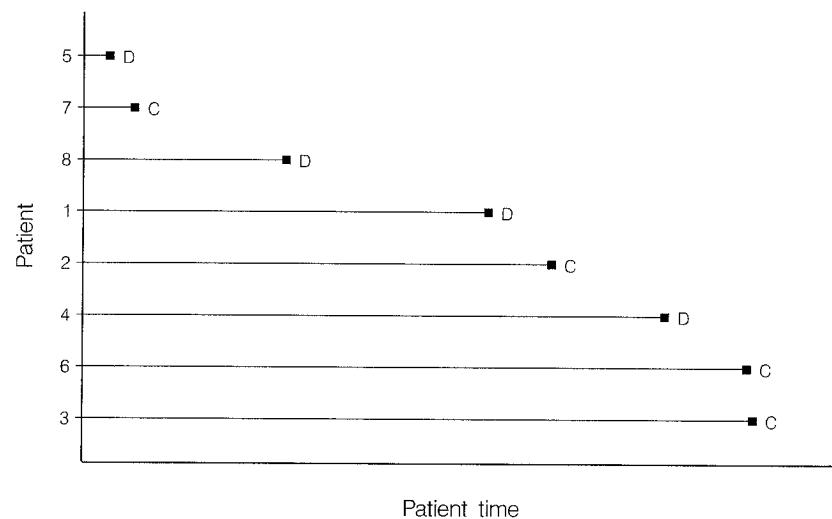


Figure 1.2 Patient time for eight patients in a survival study.

In practice, the actual data recorded will be the date on which each individual enters the study, and the date on which each individual dies or was last known to be alive. The survival time in days, weeks or months, whichever is the most appropriate, can then be calculated. Most computer software packages for survival analysis have facilities for performing this calculation from input data in the form of dates.

An important assumption that will be made in the analysis of censored survival data is that the actual survival time of an individual,  $t$ , is independent of any mechanism that causes that individual's survival time to be censored at time  $c$ , where  $c < t$ . This means that if we consider a group of individuals, all of whom have the same values of relevant prognostic variables, an individual whose survival time is censored at time  $c$  must be representative of all other individuals in that group who have survived to that time. A patient whose survival time is censored will be representative of those at risk at the censoring time if the censoring process operates randomly. Similarly, when survival data are to be analysed at a predetermined point in calendar time, or at a fixed interval of time after the time origin for each patient, the prognosis for individuals who are still alive can be taken to be independent of the censoring, so long as the time of analysis is specified before the data are examined. However, this assumption cannot be made if, for example, the survival time of an individual is censored through treatment being withdrawn as a result of a deterioration in their physical condition. This type of censoring is known as *informative censoring*. Great care should be taken to ensure that

any censoring is non-informative, for otherwise the methods presented in this book for the analysis of survival data are no longer valid.

## 1.2 Some examples

In this section, the essential features of survival data are illustrated through a number of examples. Data from these examples will then be used to illustrate some of the statistical techniques presented in subsequent chapters.

### Example 1.1 Time to discontinuation of the use of an IUD

In trials involving contraceptives, prevention of pregnancy is an obvious criterion for acceptability. However, modern contraceptives have very low failure rates, and so the occurrence of bleeding disturbances, such as amenorrhoea (the prolonged absence of bleeding), irregular or prolonged bleeding, become important in the evaluation of a particular method of contraception. To promote research into methods for analysing menstrual bleeding data from women in contraceptive trials, the World Health Organisation has made available data from clinical trials involving a number of different types of contraceptive (WHO, 1987). Part of this data set relates to the time from which a woman commences use of a particular method until discontinuation, with the discontinuation reason being recorded when known. The data in Table 1.1 refer to the number of weeks from the commencement of use of a particular type of intrauterine device (IUD), known as the Multiload 250, until discontinuation because of menstrual bleeding problems. Data are given for 18 women, all of whom were aged between 18 and 35 years and who had experienced two previous pregnancies. Discontinuation times that are censored are labelled with an asterisk.

Table 1.1 Time in weeks to discontinuation of the use of an IUD.

10	13*	18*	19	23*	30	36	38*	54*
56*	59	75	93	97	104*	107	107*	107*

In this example, the time origin corresponds to the first day in which a woman uses the IUD, and the end-point is discontinuation because of bleeding problems. Some women in the study ceased using the IUD because of the desire for pregnancy, or because they had no further need for a contraceptive, while others were simply lost to follow-up. These reasons account for the censored discontinuation times of 13, 18, 23, 38, 54 and 56 weeks. The study protocol called for the menstrual bleeding experience of each woman to be documented for a period of two years from the time origin. For practical reasons, each woman could not be examined exactly two years after recruitment to determine if they were still using the IUD, and this is why there are three discontinuation times greater than 104 weeks that are right-censored.

One objective in an analysis of these data would be to summarise the distribution of discontinuation times. We might then wish to estimate the median

time to discontinuation of the IUD, or the probability that a woman will stop using the device after a given period of time. Indeed, a graph of this estimated probability, as a function of time, will provide a useful summary of the observed data.

#### *Example 1.2 Prognosis for women with breast cancer*

Breast cancer is one of the most common forms of cancer occurring in women living in the Western world. However, the biological behaviour of the tumour is unpredictable, and there is at present no reliable method for determining whether or not a tumour is likely to have metastasised, or spread, to other organs in the body. Around 80% of women presenting with primary breast cancer are likely to have tumours that have already metastasised to other sites. If these patients could be identified, adjunctive treatment could be focused on them, while the remaining 20% could be reassured that their disease is surgically curable. The most accurate assessment of the prognosis of a patient is based on whether or not there is lymph node involvement. However, as a result of the trend toward more conservative forms of breast surgery, this indication will often be unknown. This has prompted research into alternative procedures for predicting the survival prospects of breast cancer patients.

The aim of an investigation carried out at the Middlesex Hospital, and documented in Leathem and Brooks (1987), was to evaluate a histochemical marker that discriminates between primary breast cancer that has metastasised and that which has not. The marker under study was a lectin from the albumin gland of the Roman snail, *Helix pomatia*, known as *Helix pomatia* agglutinin, or HPA. The marker binds to those breast cancer cells associated with metastasis to local lymph nodes, and the HPA stained cells can be identified by microscopic examination. In order to investigate whether HPA staining can be used to predict the survival experience of women who present with breast cancer, a retrospective study was carried out, based on the records of women who had received surgical treatment for breast cancer. Sections of the tumours of these women were treated with HPA and each tumour was subsequently classified as being positively or negatively stained, positive staining corresponding to a tumour with the potential for metastasis. The study was concluded in July 1987, when the survival times of those women who had died of breast cancer were calculated. For those women whose survival status in July 1987 was unknown, the time from surgery to the date on which they were last known to be alive is regarded as a censored survival time. The survival times of women who had died from causes other than breast cancer are also regarded as right-censored. The data given in Table 1.2 refer to the survival times in months of women who had received a simple or radical mastectomy to treat a tumour of Grade II, III or IV, between January 1969 and December 1971. In the table, the survival times of each woman are classified according to whether their tumour was positively or negatively stained. Censored survival times are labelled with an asterisk.

In the analysis of the data from this study, we will be particularly interested in whether or not there is a difference in the survival experience of the two

**Table 1.2** Survival times of women with tumours that were negatively or positively stained with HPA.

Negative staining	Positive staining
23	5
47	8
69	10
70*	13
71*	18
100*	24
101*	26
148	26
181	31
198*	35
208*	40
212*	41
224*	48
	50
	59
	61
	68
	71
	76*
	105*
	107*
	109*
	113
	116*
	118
	143
	154*
	162*
	188*
	212*
	217*
	225*

groups of women. If there were evidence that those women with negative HPA staining tended to live longer after surgery than those with a positive staining, we would conclude that the prognosis for a breast cancer patient was dependent on the result of the staining procedure.

#### *Example 1.3 Survival of multiple myeloma patients*

Multiple myeloma is a malignant disease characterised by the accumulation of abnormal plasma cells, a type of white blood cell, in the bone marrow. The proliferation of the abnormal plasma cells within the bone causes pain and the destruction of bone tissue. Patients with multiple myeloma also experience anaemia, haemorrhages, recurrent infections and weakness. Unless treated, the condition is invariably fatal. The aim of a study carried out at the Medical Center of the University of West Virginia, USA, was to examine the association between the values of certain *explanatory variables* or *covariates* and the survival time of patients. In the study, the primary response variable was the time, in months, from diagnosis until death from multiple myeloma.

The data in Table 1.3, which were obtained from Krall, Uthoff and Harley (1975), relate to 48 patients, all of whom were aged between 50 and 80 years. Some of these patients had not died by the time that the study was completed, and so these individuals contribute right-censored survival times. The coding of the survival status of an individual in the table is such that zero denotes a censored observation and unity death from multiple myeloma. At the time of diagnosis, the values of a number of explanatory variables were

recorded for each patient. These included the age of the patient in years, their sex (1 = male, 2 = female), the levels of blood urea nitrogen (*Bun*), serum calcium (*Ca*) and haemoglobin (*Hb*), the percentage of plasma cells in the bone marrow (*Pcells*) and an indicator variable (*Protein*) that denotes whether or not Bence-Jones protein was present in the urine (0 = absent, 1 = present).

The main aim of an analysis of these data would be to investigate the effect of the risk factors *Bun*, *Ca*, *Hb*, *Pcells* and *Protein* on the survival time of the multiple myeloma patients. The effects of these risk factors may be modified by the age or sex of a patient, and so the extent to which the relationship between survival and the important risk factors is consistent for each sex and for each of a number of age groups will also need to be studied.

#### *Example 1.4 Comparison of two treatments for prostatic cancer*

A randomised controlled clinical trial to compare treatments for prostatic cancer was begun in 1967 by the Veteran's Administration Cooperative Urological Research Group. The trial was double blind and two of the treatments used in the study were a placebo and 1.0 mg of diethylstilbestrol (DES). The treatments were administered daily by mouth. The time origin of the study is the date on which a patient was randomised to a treatment, and the end-point is the death of the patient from prostatic cancer.

The full data set is given in Andrews and Herzberg (1985), but the data used in this example are from patients presenting with Stage III cancer, that is, patients for whom there was evidence of a local extension of the tumour beyond the prostatic capsule, but without elevated serum prostatic acid phosphatase. Furthermore, the patients were those who had no history of cardiovascular disease, had a normal ECG result at trial entry, and who were not confined to bed during the daytime. In addition to recording the survival time of each patient in the study, information was recorded on a number of other prognostic factors. These included the age of the patient at trial entry, their serum haemoglobin level in gm/100 ml, the size of their primary tumour in cm<sup>2</sup>, and the value of a combined index of tumour stage and grade. This index is known as the Gleason index; the more advanced the tumour, the greater the value of the index.

Table 1.4 gives the data recorded for 38 patients, where the survival times are given in months. The survival times of patients who died from other causes, or who were lost during the follow-up process are regarded as censored. A variable associated with the status of an individual at the end of the study takes the value unity if the patient has died from prostatic cancer, and zero if the survival time is right-censored. The variable associated with the treatment group takes the value 2 when an individual is treated with DES and unity if an individual is on the placebo treatment.

The main aim of this study is to determine the extent of any evidence that patients treated with DES survive longer than those treated with the placebo. Since the data on which this example is based are from a randomised trial, one might expect that the distributions of the prognostic factors, that is the age of patient, serum haemoglobin level, size of tumour and Gleason index, will

**Table 1.3** *Survival times of patients in a study on multiple myeloma.*

Patient number	Survival time	Status	Age	Sex	Bun	Ca	Hb	Pcells	Protein
1	13	1	66	1	25	10	14.6	18	1
2	52	0	66	1	13	11	12.0	100	0
3	6	1	53	2	15	13	11.4	33	1
4	40	1	69	1	10	10	10.2	30	1
5	10	1	65	1	20	10	13.2	66	0
6	7	0	57	2	12	8	9.9	45	0
7	66	1	52	1	21	10	12.8	11	1
8	10	0	60	1	41	9	14.0	70	1
9	10	1	70	1	37	12	7.5	47	0
10	14	1	70	1	40	11	10.6	27	0
11	16	1	68	1	39	10	11.2	41	0
12	4	1	50	2	172	9	10.1	46	1
13	65	1	59	1	28	9	6.6	66	0
14	5	1	60	1	13	10	9.7	25	0
15	11	0	66	2	25	9	8.8	23	0
16	10	1	51	2	12	9	9.6	80	0
17	15	0	55	1	14	9	13.0	8	0
18	5	1	67	2	26	8	10.4	49	0
19	76	0	60	1	12	12	14.0	9	0
20	56	0	66	1	18	11	12.5	90	0
21	88	1	63	1	21	9	14.0	42	1
22	24	1	67	1	10	10	12.4	44	0
23	51	1	60	2	10	10	10.1	45	1
24	4	1	74	1	48	9	6.5	54	0
25	40	0	72	1	57	9	12.8	28	1
26	8	1	55	1	53	12	8.2	55	0
27	18	1	51	1	12	15	14.4	100	0
28	5	1	70	2	130	8	10.2	23	0
29	16	1	53	1	17	9	10.0	28	0
30	50	1	74	1	37	13	7.7	11	1
31	40	1	70	2	14	9	5.0	22	0
32	1	1	67	1	165	10	9.4	90	0
33	36	1	63	1	40	9	11.0	16	1
34	5	1	77	1	23	8	9.0	29	0
35	10	1	61	1	13	10	14.0	19	0
36	91	1	58	2	27	11	11.0	26	1
37	18	0	69	2	21	10	10.8	33	0
38	1	1	57	1	20	9	5.1	100	1
39	18	0	59	2	21	10	13.0	100	0
40	6	1	61	2	11	10	5.1	100	0
41	1	1	75	1	56	12	11.3	18	0
42	23	1	56	2	20	9	14.6	3	0
43	15	1	62	2	21	10	8.8	5	0
44	18	1	60	2	18	9	7.5	85	1
45	12	0	71	2	46	9	4.9	62	0
46	12	1	60	2	6	10	5.5	25	0
47	17	1	65	2	28	8	7.5	8	0
48	3	0	59	1	90	10	10.2	6	1

**Table 1.4** Survival times of prostatic cancer patients in a clinical trial to compare two treatments.

Patient number	Treatment	Survival time	Status	Age	Serum haem.	Size of tumour	Gleason index
1	1	65	0	67	13.4	34	8
2	2	61	0	60	14.6	4	10
3	2	60	0	77	15.6	3	8
4	1	58	0	64	16.2	6	9
5	2	51	0	65	14.1	21	9
6	1	51	0	61	13.5	8	8
7	1	14	1	73	12.4	18	11
8	1	43	0	60	13.6	7	9
9	2	16	0	73	13.8	8	9
10	1	52	0	73	11.7	5	9
11	1	59	0	77	12.0	7	10
12	2	55	0	74	14.3	7	10
13	2	68	0	71	14.5	19	9
14	2	51	0	65	14.4	10	9
15	1	2	0	76	10.7	8	9
16	1	67	0	70	14.7	7	9
17	2	66	0	70	16.0	8	9
18	2	66	0	70	14.5	15	11
19	2	28	0	75	13.7	19	10
20	2	50	1	68	12.0	20	11
21	1	69	1	60	16.1	26	9
22	1	67	0	71	15.6	8	8
23	2	65	0	51	11.8	2	6
24	1	24	0	71	13.7	10	9
25	2	45	0	72	11.0	4	8
26	2	64	0	74	14.2	4	6
27	1	61	0	75	13.7	10	12
28	1	26	1	72	15.3	37	11
29	1	42	1	57	13.9	24	12
30	2	57	0	72	14.6	8	10
31	2	70	0	72	13.8	3	9
32	2	5	0	74	15.1	3	9
33	2	54	0	51	15.8	7	8
34	1	36	1	72	16.4	4	9
35	2	70	0	71	13.6	2	10
36	2	67	0	73	13.8	7	8
37	1	23	0	68	12.5	2	8
38	1	62	0	63	13.2	3	8

be similar over the patients in each of the two treatment groups. However, it would not be wise to rely on this assumption. For example, it could turn out that patients in the placebo group had larger tumours on average than those in the group treated with DES. If patients with large tumours have a poorer prognosis than those with small tumours, the size of the treatment effect would be overestimated unless proper account was taken of the size of the tumour in the analysis. Consequently, it will first be necessary to determine if any of the covariates are related to survival time. If so, the effect of these variables will need to be allowed for when comparing the survival experiences of the patients in the two treatment groups.

### 1.3 Survivor function and hazard function

In summarising survival data, there are two functions of central interest, namely the *survivor function* and the *hazard function*. These functions are therefore defined in this first chapter.

The actual survival time of an individual,  $t$ , can be regarded as the value of a variable,  $T$ , which can take any non-negative value. The different values that  $T$  can take have a *probability distribution*, and we call  $T$  the *random variable* associated with the survival time. Now suppose that the random variable  $T$  has a probability distribution with underlying *probability density function*  $f(t)$ . The *distribution function* of  $T$  is then given by

$$F(t) = P(T < t) = \int_0^t f(u) du,$$

and represents the probability that the survival time is less than some value  $t$ .

The survivor function,  $S(t)$ , is defined to be the probability that the survival time is greater than or equal to  $t$ , and so

$$S(t) = P(T \geq t) = 1 - F(t). \quad (1.1)$$

The survivor function can therefore be used to represent the probability that an individual survives from the time origin to some time beyond  $t$ .

The *hazard function* is widely used to express the risk or hazard of death at some time  $t$ , and is obtained from the probability that an individual dies at time  $t$ , conditional on he or she having survived to that time. For a formal definition of the hazard function, consider the probability that the random variable associated with an individual's survival time,  $T$ , lies between  $t$  and  $t + \delta t$ , conditional on  $T$  being greater than or equal to  $t$ , written  $P(t \leq T < t + \delta t | T \geq t)$ . This conditional probability is then expressed as a probability per unit time by dividing by the time interval,  $\delta t$ , to give a *rate*. The hazard function,  $h(t)$ , is then the limiting value of this quantity, as  $\delta t$  tends to zero, so that

$$h(t) = \lim_{\delta t \rightarrow 0} \left\{ \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t} \right\}. \quad (1.2)$$

The function  $h(t)$  is also referred to as the *hazard rate*, the *instantaneous death rate*, the *intensity rate*, or the *force of mortality*.

From equation (1.2),  $h(t)\delta t$  is the approximate probability that an individual dies in the interval  $(t, t + \delta t)$ , conditional on that person having survived to time  $t$ . For example, if the survival time is measured in days,  $h(t)$  is the approximate probability that an individual, who is alive on day  $t$ , dies in the following day. For this reason, the hazard function is often simply interpreted as the risk of death at time  $t$ .

From the definition of the hazard function in equation (1.2), we can obtain some useful relationships between the survivor and hazard functions. According to a standard result from probability theory, the probability of an event  $A$ , conditional on the occurrence of an event  $B$ , is given by  $P(A|B) = P(AB)/P(B)$ , where  $P(AB)$  is the probability of the joint occurrence of  $A$  and  $B$ . Using this result, the conditional probability in the definition of the hazard function in equation (1.2) is

$$\frac{P(t \leq T < t + \delta t)}{P(T \geq t)},$$

which is equal to

$$\frac{F(t + \delta t) - F(t)}{S(t)},$$

where  $F(t)$  is the distribution function of  $T$ . Then,

$$h(t) = \lim_{\delta t \rightarrow 0} \left\{ \frac{F(t + \delta t) - F(t)}{\delta t} \right\} \frac{1}{S(t)}.$$

Now,

$$\lim_{\delta t \rightarrow 0} \left\{ \frac{F(t + \delta t) - F(t)}{\delta t} \right\}$$

is the definition of the derivative of  $F(t)$  with respect to  $t$ , which is  $f(t)$ , and so

$$h(t) = \frac{f(t)}{S(t)}. \quad (1.3)$$

It then follows that

$$h(t) = -\frac{d}{dt} \{\log S(t)\}, \quad (1.4)$$

and so

$$S(t) = \exp \{-H(t)\}, \quad (1.5)$$

where

$$H(t) = \int_0^t h(u) du. \quad (1.6)$$

The function  $H(t)$  features widely in survival analysis, and is called the *integrated* or *cumulative hazard*. From equation (1.5), the cumulative hazard can be obtained from the survivor function, since

$$H(t) = -\log S(t). \quad (1.7)$$

In the analysis of survival data, the survivor function and hazard function are estimated from the observed survival times. Methods of estimation that

do not require the form of the probability density function of  $T$  to be specified are described in Chapters 2 and 3, while methods based on the assumption of a particular survival time distribution are presented in Chapters 5 and 6.

#### 1.4 Further reading

An introduction to the techniques used in the analysis of survival data is included in a number of general books on statistics in medical research, such as those of Altman (1991) and Armitage *et al.* (2001). Parmar and Machin (1995) provide a practical guide to the analysis of survival data from clinical trials, using non-technical language.

There are a number of textbooks that provide an introduction to the methods of survival analysis, illustrated with practical examples. Lee and Wang (2003) provides a broad coverage of topics with illustrations drawn from biology and medicine. Kleinbaum (1996) provides a self-learning text in two column format, which, like the texts of Harris and Albert (1991) and Miller (1998), emphasises non-parametric methods. Marubini and Valsecchi (1995) describe the analysis of survival data from clinical trials and observational studies. Hosmer and Lemeshow (1999) give a balanced account of survival analysis, with excellent chapters on model development and the interpretation of the parameter estimates in a fitted model. Klein and Moeschberger (1997) include many example data sets and exercises in their comprehensive textbook. Applications of survival analysis in the analysis of epidemiological data are described by Breslow and Day (1987) and Woodward (1999). Introductory texts that describe the application of survival analysis in other areas include those of Elandt-Johnson and Johnson (1999), who focus on actuarial applications, and Crowder *et al.* (1991) who provide a good introduction to the analysis of reliability data.

Comprehensive accounts of the subject are given by Kalbfleisch and Prentice (2002), Le (1987) and Lawless (2002). These books have been written for the postgraduate statistician or research worker, and are usually regarded as reference books rather than introductory texts. A concise review of survival analysis is given in the research monograph of Cox and Oakes (1984), and in the chapter devoted to this subject in Hinkley, Reid and Snell (1991).

The book by Hougaard (2000) on multivariate survival data incorporates more advanced topics, after introductory chapters that covers the basic features of survival analysis. Therneau and Grambsch (2000) base their presentation of survival analysis on the counting process approach, leading to a more mathematical development of the material. Smith (2002) describes how a generalisation of least squares allows linear regression models to be used in modelling censored data. Harrell (2001) gives details on many issues that arise in the development of a statistical model not found in other texts, and includes an extensive discussion of two case studies.

## Some non-parametric procedures

---

An initial step in the analysis of a set of survival data is to present numerical or graphical summaries of the survival times for individuals in a particular group. Such summaries may be of interest in their own right, or as a precursor to a more detailed analysis of the data. Survival data are conveniently summarised through estimates of the survivor function and hazard function. Methods for estimating these functions from a single sample of survival data are described in Sections 2.1 and 2.3. These methods are said to be *non-parametric* or *distribution-free*, since they do not require specific assumptions to be made about the underlying distribution of the survival times.

Once the estimated survivor function has been found, the median and other percentiles of the distribution of survival times can be estimated, as shown in Section 2.4. Numerical summaries of the data, derived on the basis of assumptions about the probability distribution from which the data have been obtained, will be considered later in Chapters 5 and 6.

When the survival times of two groups of patients are being compared, an informal comparison of the survival experience of each group of individuals can be made using the estimated survivor functions. However, there are more formal procedures that enable two groups of survival data to be compared. Two non-parametric procedures for comparing two or more groups of survival times, namely the *log-rank test* and the *Wilcoxon test*, are described in Section 2.6.

### 2.1 Estimating the survivor function

Suppose first that we have a single sample of survival times, where none of the observations are censored. The survivor function  $S(t)$ , defined in equation (1.1), is the probability that an individual survives for a time greater than or equal to  $t$ . This function can be estimated by the *empirical survivor function*, given by

$$\hat{S}(t) = \frac{\text{Number of individuals with survival times } \geq t}{\text{Number of individuals in the data set}}. \quad (2.1)$$

Equivalently,  $\hat{S}(t) = 1 - \hat{F}(t)$ , where  $\hat{F}(t)$  is the *empirical distribution function*, that is, the ratio of the total number of individuals alive at time  $t$  to the total number of individuals in the study. Notice that the empirical survivor function is equal to unity for values of  $t$  before the first death time, and zero after the final death time.

The estimated survivor function  $\hat{S}(t)$  is assumed to be constant between

two adjacent death times, and so a plot of  $\hat{S}(t)$  against  $t$  is a step-function. The function decreases immediately after each observed survival time.

#### Example 2.1 Pulmonary metastasis

One complication in the management of patients with a malignant bone tumour, or osteosarcoma, is that the tumour often spreads to the lungs. This pulmonary metastasis is life-threatening. In a study concerned with the treatment of pulmonary metastasis arising from osteosarcoma, Burdette and Gehan (1970) give the following survival times, in months, of eleven male patients.

11 13 13 13 13 13 14 14 15 15 17

Using equation (2.1), the estimated values of the survivor function at times 11, 13, 14, 15 and 17 months are 1.000, 0.909, 0.455, 0.273, and 0.091. The estimated value of the survivor function is unity from the time origin until 11 months, and zero after 17 months. A graph of the estimated survivor function is given in Figure 2.1.

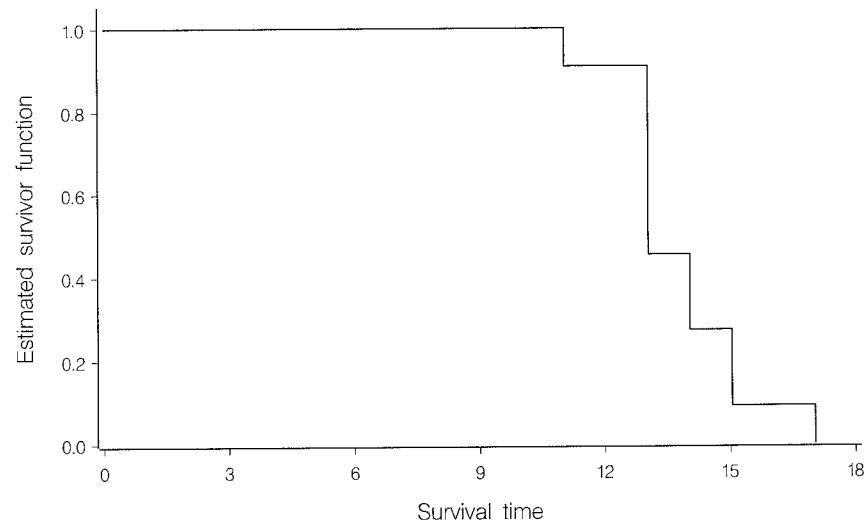


Figure 2.1 Estimated survivor function for the data from Example 2.1.

The method of estimating the survivor function illustrated in the above example cannot be used when there are censored observations. The reason for this is that the method does not allow information provided by an individual whose survival time is censored before time  $t$  to be used in computing the estimated survivor function at  $t$ . Non-parametric methods for estimating  $S(t)$ , which can be used in the presence of censored survival times, are described in the following sections.

#### 2.1.1 Life-table estimate of the survivor function

The *life-table estimate* of the survivor function, also known as the *Actuarial estimate of survivor function*, is obtained by first dividing the period of observation into a series of time intervals. These intervals need not necessarily be of equal length, although they usually are. The number of intervals used will depend on the number of individuals in the study, but would usually be somewhere between 5 and 15.

Suppose that the  $j$ th of  $m$  such intervals,  $j = 1, 2, \dots, m$ , extends from time  $t'_j$  to  $t'_{j+1}$ , and let  $d_j$  and  $c_j$  denote the number of deaths and the number of censored survival times, respectively, in this time interval. Also let  $n_j$  be the number of individuals who are alive, and therefore at risk of death, at the start of the  $j$ th interval. We now make the assumption that the censoring process is such that the censored survival times occur uniformly throughout the  $j$ th interval, so that the average number of individuals who are at risk during this interval is

$$n'_j = n_j - c_j/2. \quad (2.2)$$

This assumption is sometimes known as the *actuarial assumption*.

In the  $j$ th interval, the probability of death can be estimated by  $d_j/n'_j$ , so that the corresponding survival probability is  $(n'_j - d_j)/n'_j$ . Now consider the probability that an individual survives beyond time  $t'_k$ ,  $k = 1, 2, \dots, m$ , that is, until some time after the start of the  $k$ th interval. This will be the product of the probabilities that an individual survives beyond the start of the  $k$ th interval and through each of the  $k-1$  preceding intervals, and so the life-table estimate of the survivor function is given by

$$S^*(t) = \prod_{j=1}^k \left( \frac{n'_j - d_j}{n'_j} \right), \quad (2.3)$$

for  $t'_k \leq t < t'_{k+1}$ ,  $k = 1, 2, \dots, m$ . The estimated probability of surviving until the start of the first interval,  $t'_1$ , is of course unity, while the estimated probability of surviving beyond  $t'_{m+1}$  is zero. A graphical estimate of the survivor function will then be a step-function with constant values of the function in each time interval.

#### Example 2.2 Survival of multiple myeloma patients

To illustrate the computation of the life-table estimate, consider the data on the survival times of the 48 multiple myeloma patients given in Table 1.3. In this illustration, the information collected on other explanatory variables for each individual will be ignored.

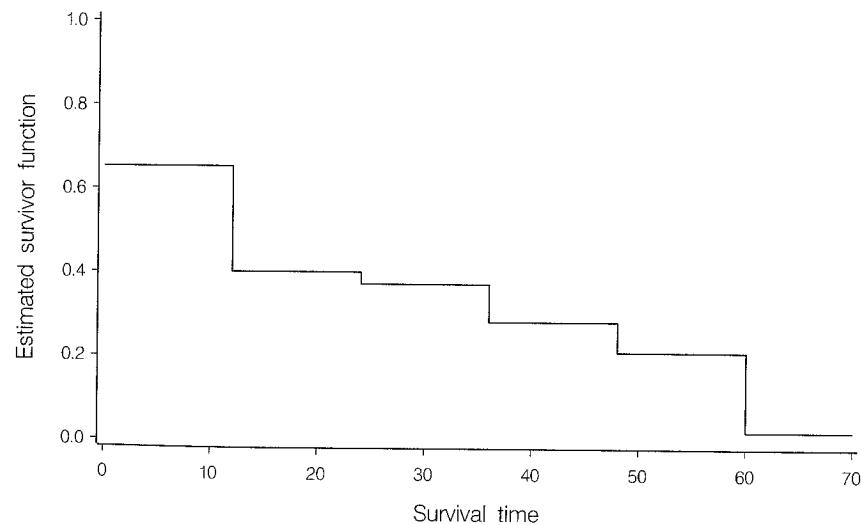
The survival times are first grouped to give the number of patients who die,  $d_j$ , and the number who are censored,  $c_j$ , in each of the first five years of the study, and in the subsequent three-year period. The number at risk of death at the start of each of these intervals,  $n_j$ , is then computed, together with the adjusted number at risk,  $n'_j$ . Finally, the probability of survival through each interval is estimated, from which the estimated survivor function is obtained using equation (2.3). The calculations are shown in Table 2.1, in which the

time period is given in months, and the interval that begins at time  $t'_k$  and ends just before time  $t'_{k+1}$ , for  $k = 1, 2, \dots, m$ , is denoted  $t'_k$ .

**Table 2.1** Life-table estimate of the survivor function for the data from Example 1.3.

Interval	Time period	$d_j$	$c_j$	$n_j$	$n'_j$	$(n'_j - d_j)/n'_j$	$S^*(t)$
1	0–	16	4	48	46.0	0.6522	0.6522
2	12–	10	4	28	26.0	0.6154	0.4013
3	24–	1	0	14	14.0	0.9286	0.3727
4	36–	3	1	13	12.5	0.7600	0.2832
5	48–	2	2	9	8.0	0.7500	0.2124
6	60–	4	1	5	4.5	0.1111	0.0236

A graph of the life-table estimate of the survivor function is shown in Figure 2.2.



**Figure 2.2** Life-table estimate of the survivor function.

The form of the estimated survivor function obtained using this method is sensitive to the choice of the intervals used in its construction, just as the shape of a histogram depends on the choice of the class intervals. On the other hand, the life-table estimate is particularly well suited to situations in which the actual death times are unknown, and the only available information is the number of deaths and the number of censored observations that occur in a series of consecutive time intervals. In practice, such interval-censored survival data occur quite frequently.

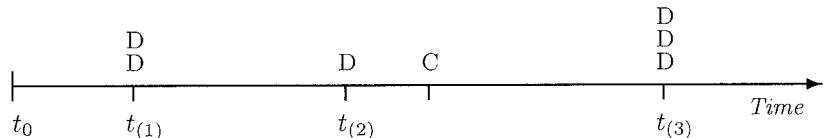
When the actual survival times are known, the life-table estimate can still be

used, as in Example 2.2, but the grouping of the survival times does result in some loss of information. This is particularly so when the number of patients is small, less than about 30, say.

### 2.1.2 Kaplan-Meier estimate of the survivor function

The first step in the analysis of ungrouped censored survival data is normally to obtain the *Kaplan-Meier estimate* of the survivor function. This estimate is therefore considered in some detail. To obtain the Kaplan-Meier estimate, a series of time intervals is constructed, as for the life-table estimate. However, each of these intervals is designed to be such that one death time is contained in the interval, and this death time is taken to occur at the start of the interval.

As an illustration, suppose that  $t_{(1)}$ ,  $t_{(2)}$  and  $t_{(3)}$  are three observed survival times arranged in rank order, so that  $t_{(1)} < t_{(2)} < t_{(3)}$ , and that  $c$  is a censored survival time that falls between  $t_{(2)}$  and  $t_{(3)}$ . The constructed intervals then begin at times  $t_{(1)}$ ,  $t_{(2)}$  and  $t_{(3)}$ , and each interval includes the one death time, although there could be more than one individual who dies at any particular death time. Notice that no interval begins at the censored time of  $c$ . The situation is illustrated diagrammatically in Figure 2.3, in which D represents a death and C a censored survival time. Notice that two individuals die at  $t_{(1)}$ , one dies at  $t_{(2)}$ , and three die at  $t_{(3)}$ .



**Figure 2.3** Construction of intervals used in the derivation of the Kaplan-Meier estimate.

The time origin is denoted by  $t_0$ , and so there is an initial period commencing at  $t_0$ , which ends just before  $t_{(1)}$ , the time of the first death. This means that the interval from  $t_0$  to  $t_{(1)}$  will not include a death time. The first constructed interval extends from  $t_{(1)}$  to just before  $t_{(2)}$ , and since the second death time is at  $t_{(2)}$ , this interval includes the single death time at  $t_{(1)}$ . The second interval begins at time  $t_{(2)}$  and ends just before  $t_{(3)}$ , and includes the death time at  $t_{(2)}$  and the censored time  $c$ . There is also a third interval beginning at  $t_{(3)}$ , which contains the longest survival time,  $t_{(3)}$ .

In general, suppose that there are  $n$  individuals with observed survival times  $t_1, t_2, \dots, t_n$ . Some of these observations may be right-censored, and there may also be more than one individual with the same observed survival time. We therefore suppose that there are  $r$  death times amongst the individuals, where  $r \leq n$ . After arranging these death times in ascending order, the  $j$ th is denoted  $t_{(j)}$ , for  $j = 1, 2, \dots, r$ , and so the  $r$  ordered death times are  $t_{(1)} < t_{(2)} < \dots < t_{(r)}$ . The number of individuals who are alive just before time  $t_{(j)}$ , including those who are about to die at this time, will be denoted  $n_j$ , for  $j = 1, 2, \dots, r$ , and  $d_j$  will denote the number who die at this time. The time interval from

$t_{(j)} - \delta$  to  $t_{(j)}$ , where  $\delta$  is an infinitesimal time interval, then includes one death time. Since there are  $n_j$  individuals who are alive just before  $t_{(j)}$  and  $d_j$  deaths at  $t_{(j)}$ , the probability that an individual dies during the interval from  $t_{(j)} - \delta$  to  $t_{(j)}$  is estimated by  $d_j/n_j$ . The corresponding estimated probability of survival through that interval is then  $(n_j - d_j)/n_j$ .

It sometimes happens that there are censored survival times that occur at the same time as one or more deaths, so that a death time and a censored survival time appear to occur simultaneously. In this event, the censored survival time is taken to occur immediately after the death time when computing the values of the  $n_j$ .

From the manner in which the time intervals are constructed, the interval from  $t_{(j)}$  to  $t_{(j+1)} - \delta$ , the time immediately before the next death time, contains no deaths. The probability of surviving from  $t_{(j)}$  to  $t_{(j+1)} - \delta$  is therefore unity, and the joint probability of surviving from  $t_{(j)} - \delta$  to  $t_{(j)}$  and from  $t_{(j)}$  to  $t_{(j+1)} - \delta$  can be estimated by  $(n_j - d_j)/n_j$ . In the limit, as  $\delta$  tends to zero,  $(n_j - d_j)/n_j$  becomes an estimate of the probability of surviving the interval from  $t_{(j)}$  to  $t_{(j+1)}$ .

We now make the assumption that the deaths of the individuals in the sample occur independently of one another. Then, the estimated survivor function at any time,  $t$ , in the  $k$ th constructed time interval from  $t_{(k)}$  to  $t_{(k+1)}$ ,  $k = 1, 2, \dots, r$ , where  $t_{(r+1)}$  is defined to be  $\infty$ , will be the estimated probability of surviving beyond  $t_{(k)}$ . This is actually the probability of surviving through the interval from  $t_{(k)}$  to  $t_{(k+1)}$ , and all preceding intervals, and leads to the Kaplan-Meier estimate of the survivor function, which is given by

$$\hat{S}(t) = \prod_{j=1}^k \left( \frac{n_j - d_j}{n_j} \right), \quad (2.4)$$

for  $t_{(k)} \leq t < t_{(k+1)}$ ,  $k = 1, 2, \dots, r$ , with  $\hat{S}(t) = 1$  for  $t < t_{(1)}$ , and where  $t_{(r+1)}$  is taken to be  $\infty$ . Strictly speaking, if the largest observation is a censored survival time,  $t^*$ , say,  $\hat{S}(t)$  is undefined for  $t > t^*$ . On the other hand, if the largest observed survival time,  $t_{(r)}$ , is an uncensored observation,  $n_r = d_r$ , and so  $\hat{S}(t)$  is zero for  $t \geq t_{(r)}$ . A plot of the Kaplan-Meier estimate of the survivor function is a step-function, in which the estimated survival probabilities are constant between adjacent death times and decrease at each death time.

Equation (2.4) shows that, as for the life-table estimate of the survivor function in equation (2.3), the Kaplan-Meier estimate is formed as a product of a series of estimated probabilities. In fact, the Kaplan-Meier estimate is the limiting value of the life-table estimate in equation (2.3) as the number of intervals tends to infinity and their width tends to zero. For this reason, the Kaplan-Meier estimate is also known as the *product-limit estimate* of the survivor function.

Note that if there are no censored survival times in the data set,  $n_j - d_j = n_{j+1}$ ,  $j = 1, 2, \dots, k$ , in equation (2.4), and on expanding the product we get

$$\hat{S}(t) = \frac{n_2}{n_1} \times \frac{n_3}{n_2} \times \cdots \times \frac{n_{k+1}}{n_k}. \quad (2.5)$$

This reduces to  $n_{k+1}/n_1$ , for  $k = 1, 2, \dots, r-1$ , with  $\hat{S}(t) = 1$  for  $t < t_{(1)}$  and  $\hat{S}(t) = 0$  for  $t \geq t_{(r)}$ . Now,  $n_1$  is the number of individuals at risk just before the first death time, which is the number of individuals in the sample, and  $n_{k+1}$  is the number of individuals with survival times greater than or equal to  $t_{(k+1)}$ . Consequently, in the absence of censoring,  $\hat{S}(t)$  is simply the empirical survivor function defined in equation (2.1). The Kaplan-Meier estimate is therefore a generalisation of the empirical survivor function that accommodates censored observations.

### Example 2.3 Time to discontinuation of the use of an IUD

Data from 18 women on the time to discontinuation of the use of an IUD were given in Table 1.1. For these data, the survivor function,  $S(t)$ , represents the probability that a woman discontinues the use of the contraceptive device after any time  $t$ . The Kaplan-Meier estimate of the survivor function is readily obtained using equation (2.4), and the required calculations are set out in Table 2.2. The estimated survivor function,  $\hat{S}(t)$ , is plotted in Figure 2.4.

**Table 2.2** Kaplan-Meier estimate of the survivor function for the data from Example 1.1.

Time interval	$n_j$	$d_j$	$(n_j - d_j)/n_j$	$\hat{S}(t)$
0–	18	0	1.0000	1.0000
10–	18	1	0.9444	0.9444
19–	15	1	0.9333	0.8815
30–	13	1	0.9231	0.8137
36–	12	1	0.9167	0.7459
59–	8	1	0.8750	0.6526
75–	7	1	0.8571	0.5594
93–	6	1	0.8333	0.4662
97–	5	1	0.8000	0.3729
107	3	1	0.6667	0.2486

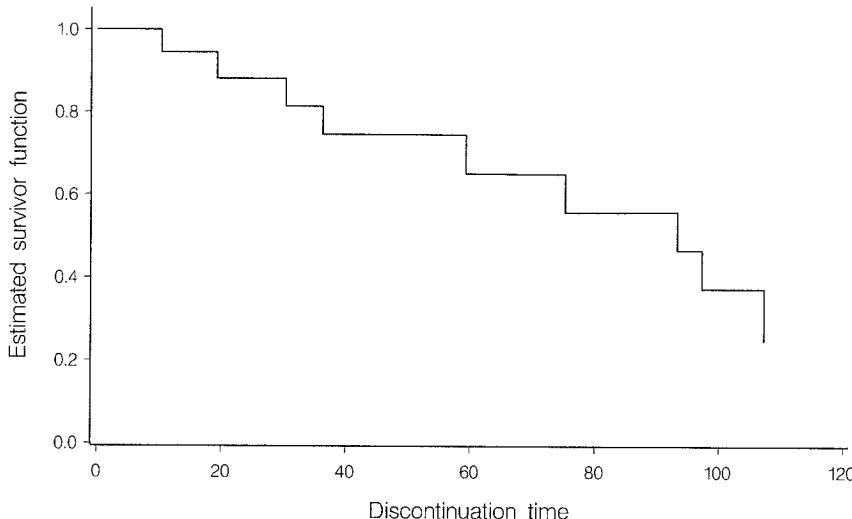
Note that since the largest discontinuation time of 107 days is censored,  $\hat{S}(t)$  is not defined beyond  $t = 107$ .

### 2.1.3 Nelson-Aalen estimate of the survivor function

An alternative estimate of the survivor function, which is based on the individual event times, is the *Nelson-Aalen estimate*, given by

$$\tilde{S}(t) = \prod_{j=1}^k \exp(-d_j/n_j). \quad (2.6)$$

This estimate can be obtained from an estimate of the cumulative hazard function, as shown in Section 2.3.3. Moreover, the Kaplan-Meier estimate of the survivor function can be regarded as an approximation to the Nelson-



**Figure 2.4** Kaplan-Meier estimate of the survivor function for the data from Example 1.1.

Aalen estimate. To show this, we use the result that

$$e^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \dots,$$

which is approximately equal to  $1 - x$  when  $x$  is small. It then follows that  $\exp(-d_j/n_j) \approx 1 - (d_j/n_j) = (n_j - d_j)/n_j$ , so long as  $d_j$  is small relative to  $n_j$ , which it will be except at the latest survival times. Consequently, the Kaplan-Meier estimate,  $\hat{S}(t)$ , in equation (2.4), approximates the Nelson-Aalen estimate,  $\tilde{S}(t)$ , in equation (2.6).

The Nelson-Aalen estimate of the survivor function, also known as *Altshuler's estimate*, will always be greater than the Kaplan-Meier estimate at any given time, since  $e^{-x} \geq 1 - x$ , for all values of  $x$ . Although the Nelson-Aalen estimate has been shown to perform better than the Kaplan-Meier estimate in small samples, in many circumstances, the estimates will be very similar, particularly at the earlier survival times. Since the Kaplan-Meier estimate is a generalisation of the empirical survivor function, the latter estimate has much to commend it.

#### Example 2.4 Time to discontinuation of the use of an IUD

The values shown in Table 2.2, which gives the Kaplan-Meier estimate of the survivor function for the data on the time to discontinuation of the use of an IUD, can be used to calculate the Nelson-Aalen estimate. This estimate is shown in Table 2.3.

From this table we see that the Kaplan-Meier and Nelson-Aalen estimates of the survivor function differ by less than 0.04. However, when we consider

**Table 2.3** Nelson-Aalen estimate of the survivor function for the data from Example 1.1.

Time interval	$\exp(-d_j/n_j)$	$\tilde{S}(t)$
0-	1.0000	1.0000
10-	0.9460	0.9460
19-	0.9355	0.8850
30-	0.9260	0.8194
36-	0.9200	0.7539
59-	0.8825	0.6653
75-	0.8669	0.5768
93-	0.8465	0.4882
97-	0.8187	0.3997
107	0.7165	0.2864

the precision of these estimates, which we do in Section 2.2, we see that a difference of 0.04 is of no practical importance.

## 2.2 Standard error of the estimated survivor function

An essential aid to the interpretation of an estimate of any quantity is the precision of the estimate, which is reflected in the *standard error* of the estimate. This is defined to be the square root of the estimated variance of the estimate, and is used in the construction of an interval estimate for a quantity of interest. In this section, the standard error of estimates of the survivor function are given.

Because the Kaplan-Meier estimate is the most important and widely used estimate of the survivor function, the derivation of the standard error of  $\hat{S}(t)$  will be presented in detail in this section. The details of this derivation can be omitted on a first reading.

### 2.2.1\* Standard error of the Kaplan-Meier estimate

The Kaplan-Meier estimate of the survivor function for any value of  $t$  in the interval from  $t_{(k)}$  to  $t_{(k+1)}$  can be written as

$$\hat{S}(t) = \prod_{j=1}^k \hat{p}_j,$$

for  $k = 1, 2, \dots, r$ , where  $\hat{p}_j = (n_j - d_j)/n_j$  is the estimated probability that an individual survives through the time interval that begins at  $t_{(j)}$ ,  $j = 1, 2, \dots, r$ . Taking logarithms,

$$\log \hat{S}(t) = \sum_{j=1}^k \log \hat{p}_j,$$

and so the variance of  $\log \hat{S}(t)$  is given by

$$\text{var} \left\{ \log \hat{S}(t) \right\} = \sum_{j=1}^k \text{var} \left\{ \log \hat{p}_j \right\}. \quad (2.7)$$

Now, the number of individuals who survive through the interval beginning at  $t_{(j)}$  can be assumed to have a *binomial distribution* with parameters  $n_j$  and  $p_j$ , where  $p_j$  is the true probability of survival through that interval. The observed number who survive is  $n_j - d_j$ , and using the result that the variance of a binomial random variable with parameters  $n, p$  is  $np(1-p)$ , the variance of  $n_j - d_j$  is given by

$$\text{var} (n_j - d_j) = n_j p_j (1 - p_j).$$

Since  $\hat{p}_j = (n_j - d_j)/n_j$ , the variance of  $\hat{p}_j$  is  $\text{var} (n_j - d_j)/n_j^2$ , that is,  $p_j(1 - p_j)/n_j$ . The variance of  $\hat{p}_j$  may then be estimated by

$$\hat{p}_j(1 - \hat{p}_j)/n_j. \quad (2.8)$$

In order to obtain the variance of  $\log \hat{p}_j$ , we make use of a general result for the approximate variance of a function of a random variable. According to this result, the variance of a function  $g(X)$  of the random variable  $X$  is given by

$$\text{var} \{g(X)\} \approx \left\{ \frac{dg(X)}{dX} \right\}^2 \text{var}(X). \quad (2.9)$$

This is known as the *Taylor series approximation* to the variance of a function of a random variable. Using equation (2.9), the approximate variance of  $\log \hat{p}_j$  is  $\text{var}(\hat{p}_j)/\hat{p}_j^2$ , and using expression (2.8), the approximate estimated variance of  $\log \hat{p}_j$  is  $(1 - \hat{p}_j)/(n_j \hat{p}_j)$ , which on substitution for  $\hat{p}_j$ , reduces to

$$\frac{d_j}{n_j(n_j - d_j)}. \quad (2.10)$$

From equation (2.7),

$$\text{var} \left\{ \log \hat{S}(t) \right\} \approx \sum_{j=1}^k \frac{d_j}{n_j(n_j - d_j)}, \quad (2.11)$$

and a further application of the result in equation (2.9) gives

$$\text{var} \left\{ \log \hat{S}(t) \right\} \approx \frac{1}{[\hat{S}(t)]^2} \text{var} \left\{ \hat{S}(t) \right\},$$

so that

$$\text{var} \left\{ \hat{S}(t) \right\} \approx [\hat{S}(t)]^2 \sum_{j=1}^k \frac{d_j}{n_j(n_j - d_j)}. \quad (2.12)$$

Finally, the standard error of the Kaplan-Meier estimate of the survivor function, defined to be the square root of the estimated variance of the estimate,

is given by

$$\text{se} \left\{ \hat{S}(t) \right\} \approx \hat{S}(t) \left\{ \sum_{j=1}^k \frac{d_j}{n_j(n_j - d_j)} \right\}^{\frac{1}{2}}, \quad (2.13)$$

for  $t_{(k)} \leq t < t_{(k+1)}$ . This result is known as *Greenwood's formula*.

If there are no censored survival times,  $n_j - d_j = n_{j+1}$ , and expression (2.10) becomes  $(n_j - n_{j+1})/n_j n_{j+1}$ . Now,

$$\sum_{j=1}^k \frac{n_j - n_{j+1}}{n_j n_{j+1}} = \sum_{j=1}^k \left( \frac{1}{n_{j+1}} - \frac{1}{n_j} \right) = \frac{n_1 - n_{k+1}}{n_1 n_{k+1}},$$

which can be written as

$$\frac{1 - \hat{S}(t)}{n_1 \hat{S}(t)},$$

since  $\hat{S}(t) = n_{k+1}/n_1$  for  $t_{(k)} \leq t < t_{(k+1)}$ ,  $k = 1, 2, \dots, r - 1$ , in the absence of censoring. Hence, from equation (2.12), the estimated variance of  $\hat{S}(t)$  is  $\hat{S}(t)[1 - \hat{S}(t)]/n_1$ . This is an estimate of the variance of the empirical survivor function, given in equation (2.1), on the assumption that the number of individuals at risk at time  $t$  has a binomial distribution with parameters  $n_1, S(t)$ .

### 2.2.2\* Standard error of the life-table and Nelson-Aalen estimates

The life-table estimate of the survivor function is similar in form to the Kaplan-Meier estimate, and so the standard error of this estimator is obtained in a similar manner. The standard error of the life-table estimate is given by

$$\text{se} \{S^*(t)\} \approx S^*(t) \left\{ \sum_{j=1}^k \frac{d_j}{n'_j(n'_j - d_j)} \right\}^{\frac{1}{2}}, \quad (2.14)$$

in the notation of Section 2.1.1.

The standard error of the Nelson-Aalen estimator is

$$\text{se} \left\{ \tilde{S}(t) \right\} \approx \tilde{S}(t) \left\{ \sum_{j=1}^k \frac{d_j}{n_j^2} \right\}^{\frac{1}{2}}, \quad (2.15)$$

although other expressions have been proposed.

### 2.2.3 Confidence intervals for values of the survivor function

Once the standard error of an estimate of the survivor function has been calculated, a *confidence interval* for the corresponding value of the survivor function, at a given time  $t$ , can be found. A confidence interval is an interval estimate of the survivor function, and is the interval which is such that there is a prescribed probability that the value of the true survivor function is included

within it. The intervals constructed in this manner are sometimes referred to as *pointwise confidence intervals*, since they apply to a specific survival time.

A confidence interval for the true value of the survivor function at a given time  $t$  is obtained by assuming that the estimated value of the survivor function at  $t$  is normally distributed with mean  $\hat{S}(t)$  and estimated variance given by equation (2.12). The interval is computed from *percentage points* of the standard normal distribution. Thus, if  $Z$  is a random variable that has a standard normal distribution, the upper (one-sided)  $\alpha/2$ -point, or the two-sided  $\alpha$ -point, of this distribution is that value  $z_{\alpha/2}$  which is such that  $P(Z > z_{\alpha/2}) = \alpha/2$ . This probability is the area under the standard normal curve to the right of  $z_{\alpha/2}$ , as illustrated in Figure 2.5. For example, the two-sided 5% and 1% points of the standard normal distribution,  $z_{0.025}$  and  $z_{0.005}$ , are 1.96 and 2.58, respectively.

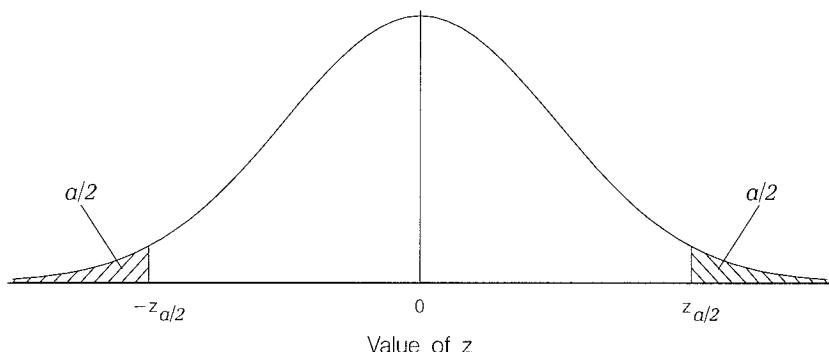


Figure 2.5 Upper and lower  $\alpha/2$ -points of the standard normal distribution.

A  $100(1 - \alpha)\%$  confidence interval for  $S(t)$ , for a given value of  $t$ , is the interval from  $\hat{S}(t) - z_{\alpha/2} \text{se}\{\hat{S}(t)\}$  to  $\hat{S}(t) + z_{\alpha/2} \text{se}\{\hat{S}(t)\}$ , where  $\text{se}\{\hat{S}(t)\}$  is found from equation (2.13). These intervals for  $S(t)$  can be superimposed on a graph of the estimated survivor function, as shown in Example 2.5.

One difficulty with this procedure arises from the fact that the confidence intervals are symmetric. When the estimated survivor function is close to zero or unity, symmetric intervals are inappropriate, since they can lead to confidence limits for the survivor function that lie outside the interval  $(0,1)$ . A pragmatic solution to this problem is to replace any limit that is greater than unity by 1.0, and any limit that is less than zero by 0.0.

An alternative procedure is to transform  $\hat{S}(t)$  to a value in the range  $(-\infty, \infty)$ , and obtain a confidence interval for the transformed value. The resulting confidence limits are then back-transformed to give a confidence interval for  $S(t)$  itself. Possible transformations are the logistic transformation,  $\log[\hat{S}(t)]/\{1 - \hat{S}(t)\}$ , and the complementary log-log transformation,  $\log\{-\log \hat{S}(t)\}$ . Note that from equation (1.7), the latter quantity is the logarithm of the cumulative hazard function. In either case, the standard

error of the transformed value of  $\hat{S}(t)$  can be found using the approximation in equation (2.9).

For example, the variance of  $\log\{-\log \hat{S}(t)\}$  is obtained from the expression for  $\text{var}\{\log \hat{S}(t)\}$  in equation (2.11). Using the general result in equation (2.9),

$$\text{var}\{\log(-X)\} \approx \frac{1}{X^2} \text{var}(X),$$

and setting  $X = \log \hat{S}(t)$  gives

$$\text{var}\left[\log\{-\log \hat{S}(t)\}\right] \approx \frac{1}{\{\log \hat{S}(t)\}^2} \sum_{j=1}^k \frac{d_j}{n_j(n_j - d_j)}.$$

The standard error of  $\log\{-\log \hat{S}(t)\}$  is the square root of this quantity. This leads to  $100(1 - \alpha)\%$  limits of the form

$$\hat{S}(t)^{\exp[\pm z_{\alpha/2} \text{se}\{\log\{-\log \hat{S}(t)\}\}]},$$

where  $z_{\alpha/2}$  is the upper  $\alpha/2$ -point of the standard normal distribution.

A further problem is that in the tails of the distribution of the survival times, that is, when  $\hat{S}(t)$  is close to zero or unity, the variance of  $\hat{S}(t)$  obtained using Greenwood's formula can underestimate the actual variance. In these circumstances, an alternative expression for the standard error of  $\hat{S}(t)$  may be used. Peto *et al.* (1977) propose that the standard error of  $\hat{S}(t)$  should be obtained from the equation

$$\text{se}\{\hat{S}(t)\} = \frac{\hat{S}(t)\sqrt{1 - \hat{S}(t)}}{\sqrt{(n_k)}},$$

for  $t_{(k)} \leq t < t_{(k+1)}$ ,  $k = 1, 2, \dots, r$ , where  $\hat{S}(t)$  is the Kaplan-Meier estimate of  $S(t)$  and  $n_k$  is the number of individuals at risk at  $t_{(k)}$ , the start of the  $k$ th constructed time interval.

This expression for the standard error of  $\hat{S}(t)$  is conservative, in the sense that the standard errors obtained will tend to be larger than they ought to be. For this reason, the Greenwood estimate is recommended for general use.

#### Example 2.5 Time to discontinuation of the use of an IUD

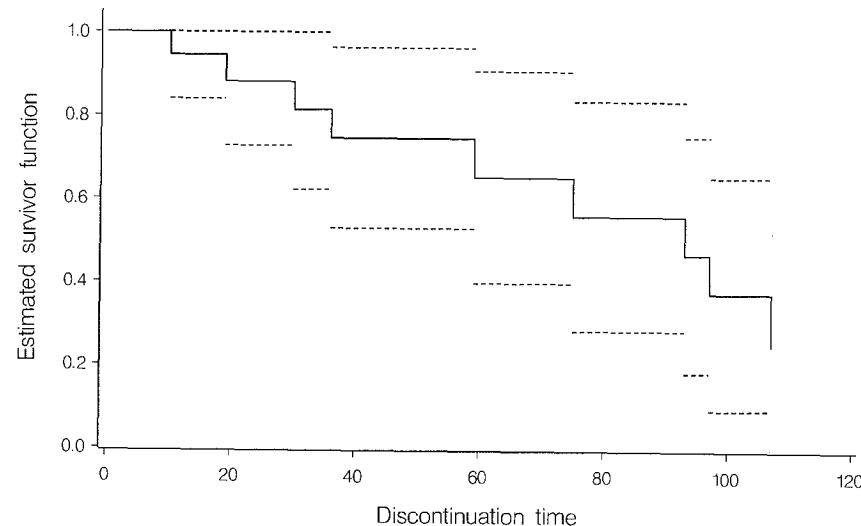
The standard error of the estimated survivor function, and 95% confidence limits for the corresponding true value of the function, for the data from Example 1.1 on the times to discontinuation of use of an IUD, are given in Table 2.4. In this table, confidence limits outside the range  $(0, 1)$  have been replaced by zero or unity.

From this table we see that in general the standard error of the estimated survivor function increases with the discontinuation time. The reason for this is that estimates of the survivor function at later times are based on fewer individuals. A graph of the estimated survivor function, with the 95% confidence limits shown as dashed lines, is given in Figure 2.6.

It is important to observe that the confidence limits plotted on such a graph are only valid for any given time. Different methods are needed to

**Table 2.4** Standard error of  $\hat{S}(t)$  and confidence intervals for  $S(t)$  for the data from Example 1.1.

Time interval	$\hat{S}(t)$	se $\{\hat{S}(t)\}$	95% confidence interval
0–	1.0000	0.0000	
10–	0.9444	0.0540	(0.839, 1.000)
19–	0.8815	0.0790	(0.727, 1.000)
30–	0.8137	0.0978	(0.622, 1.000)
36–	0.7459	0.1107	(0.529, 0.963)
59–	0.6526	0.1303	(0.397, 0.908)
75–	0.5594	0.1412	(0.283, 0.836)
93–	0.4662	0.1452	(0.182, 0.751)
97–	0.3729	0.1430	(0.093, 0.653)
107	0.2486	0.1392	(0.000, 0.522)



**Figure 2.6** Estimated survivor function and 95% confidence limits for  $S(t)$ .

produce confidence bands that are such that there is a given probability, such as 0.95, that the survivor function is contained in the band for all values of  $t$ . These bands will tend to be wider than the band formed from the pointwise confidence limits. Details will not be included, but references to these methods are given in the final section of this chapter. Notice also that the width of these intervals is very much greater than the difference between the Kaplan-Meier and Nelson-Aalen estimates of the survivor function, shown in Tables 2.2 and 2.3. Similar calculations lead to confidence limits based on life-table and Nelson-Aalen estimates of the survivor function.

### 2.3 Estimating the hazard function

A single sample of survival data may also be summarised through the hazard function, which shows the dependence of the instantaneous risk of death on time. There are a number of ways of estimating this function, two of which are described in this section.

#### 2.3.1 Life-table estimate of the hazard function

Suppose that the observed survival times have been grouped into a series of  $m$  intervals, as in the construction of the life-table estimate of the survivor function. An appropriate estimate of the average hazard of death per unit time over each interval is the observed number of deaths in that interval, divided by the average time survived in that interval. This latter quantity is the average number of persons at risk in the interval, multiplied by the length of the interval. Let the number of deaths in the  $j$ th time interval be  $d_j$ ,  $j = 1, 2, \dots, m$ , and suppose that  $n'_j$  is the average number of individuals at risk of death in that interval, where  $n'_j$  is given by equation (2.2). Assuming that the death rate is constant during the  $j$ th interval, the average time survived in that interval is  $(n'_j - d_j/2)\tau_j$ , where  $\tau_j$  is the length of the  $j$ th time interval. The life-table estimate of the hazard function in the  $j$ th time interval is then given by

$$h^*(t) = \frac{d_j}{(n'_j - d_j/2)\tau_j},$$

for  $t'_j \leq t < t'_{j+1}$ ,  $j = 1, 2, \dots, m$ , so that  $h^*(t)$  is a step-function.

The asymptotic standard error of this estimate has been shown by Gehan (1969) to be given by

$$\text{se } \{h^*(t)\} = \frac{h^*(t)\sqrt{1 - [h^*(t)\tau_j/2]^2}}{\sqrt{(d_j)}},$$

and confidence intervals for the corresponding true hazard over each of the  $m$  time intervals can be obtained in the manner described in Section 2.2.3.

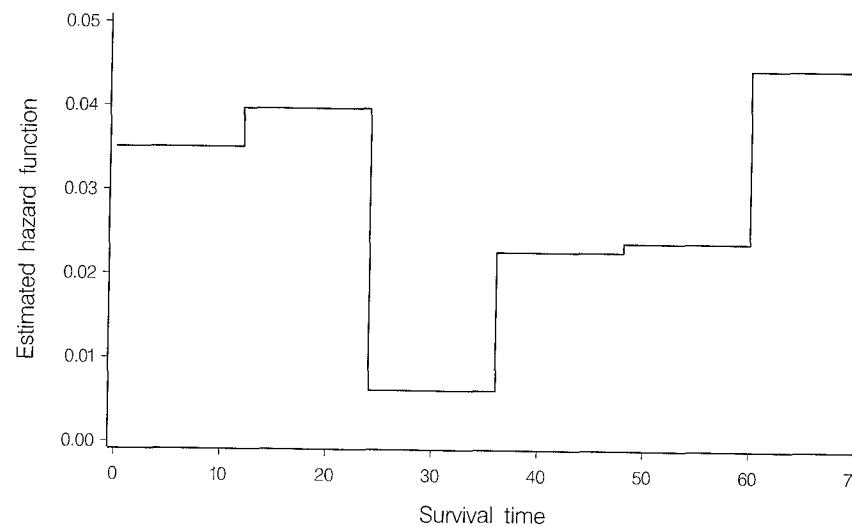
#### Example 2.6 Survival of multiple myeloma patients

The life-table estimate of the survivor function for the data from Example 1.3 on the survival times of 48 multiple myeloma patients was given in Table 2.1. Using the same time intervals as were used in Example 2.2, calculations leading to the life-table estimate of the hazard function are given in Table 2.5.

The estimated hazard function is plotted as a step-function in Figure 2.7. The general pattern is for the hazard to remain roughly constant over the first two years from diagnosis, after which time it declines and then increases gradually. However, some caution is needed in interpreting this estimate, as there are few deaths two years after diagnosis.

**Table 2.5** Life-table estimate of the hazard function for the data from Example 1.3.

Time period	$\tau_j$	$d_j$	$n'_j$	$h^*(t)$
0–	12	16	46.0	0.0351
12–	12	10	26.0	0.0397
24–	12	1	14.0	0.0062
36–	12	3	12.5	0.0227
48–	12	2	8.0	0.0238
60–	36	4	4.5	0.0444



**Figure 2.7** Life-table estimate of the hazard function for the data from Example 1.3.

### 2.3.2 Kaplan-Meier type estimate

A natural way of estimating the hazard function for unground survival data is to take the ratio of the number of deaths at a given death time to the number of individuals at risk at that time. If the hazard function is assumed to be constant between successive death times, the hazard per unit time can be found by further dividing by the time interval. Thus if there are  $d_j$  deaths at the  $j$ th death time,  $t_{(j)}$ ,  $j = 1, 2, \dots, r$ , and  $n_j$  at risk at time  $t_{(j)}$ , the hazard function in the interval from  $t_{(j)}$  to  $t_{(j+1)}$  can be estimated by

$$\hat{h}(t) = \frac{d_j}{n_j \tau_j}, \quad (2.16)$$

for  $t_{(j)} \leq t < t_{(j+1)}$ , where  $\tau_j = t_{(j+1)} - t_{(j)}$ . Notice that it is not possible to use equation (2.16) to estimate the hazard in the interval that begins at the final death time, since this interval is open-ended.

The estimate in equation (2.16) is referred to as a *Kaplan-Meier type estimate*, because the estimated survivor function derived from it is the Kaplan-Meier estimate. To show this, note that since  $\hat{h}(t)$ ,  $t_{(j)} \leq t < t_{(j+1)}$ , is an estimate of the risk of death per unit time in the  $j$ th interval, the probability of death in that interval is  $\hat{h}(t)\tau_j$ , that is,  $d_j/n_j$ . Hence an estimate of the corresponding survival probability in that interval is  $1 - (d_j/n_j)$ , and the estimated survivor function is as given by equation (2.4).

The approximate standard error of  $\hat{h}(t)$  can be found from the variance of  $d_j$ , which, following Section 2.2.1, may be assumed to have a binomial distribution with parameters  $n_j$  and  $p_j$ , where  $p_j$  is the probability of death in the interval of length  $\tau_j$ . Consequently,  $\text{var}(d_j) = n_j p_j (1 - p_j)$ , and estimating  $p_j$  by  $d_j/n_j$  gives

$$\text{se}\{\hat{h}(t)\} = \hat{h}(t) \sqrt{\left(\frac{n_j - d_j}{n_j d_j}\right)}.$$

However, when  $d_j$  is small, confidence intervals constructed using this standard error will be too wide to be of practical use.

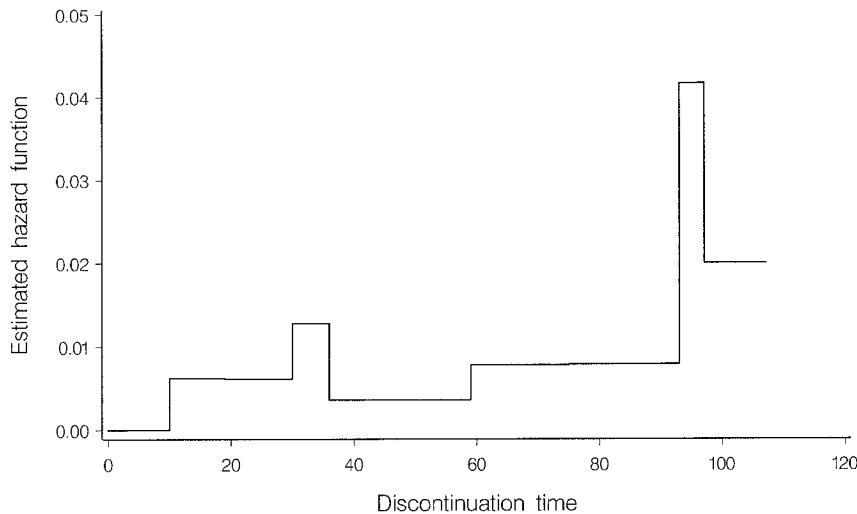
### Example 2.7 Time to discontinuation of the use of an IUD

Consider again the data on the time to discontinuation of the use of an IUD for 18 women, given in Example 1.1. The Kaplan-Meier estimate of the survivor function for these data was given in Table 2.2, and Table 2.6 gives the corresponding Kaplan-Meier type estimate of the hazard function, computed from equation (2.16). The approximate standard errors of  $\hat{h}(t)$  are also given.

**Table 2.6** Kaplan-Meier type estimate of the hazard function for the data from Example 1.1.

Time interval	$\tau_j$	$n_j$	$d_j$	$\hat{h}(t)$	$\text{se}\{\hat{h}(t)\}$
0–	10	18	0	0.0000	–
10–	9	18	1	0.0062	0.0060
19–	11	15	1	0.0061	0.0059
30–	6	13	1	0.0128	0.0123
36–	23	12	1	0.0036	0.0035
59–	16	8	1	0.0078	0.0073
75–	18	7	1	0.0079	0.0073
93–	4	6	1	0.0417	0.0380
97–	10	5	1	0.0200	0.0179

Figure 2.8 shows a plot of the estimated hazard function. From this figure, there is some evidence that the longer the IUD is used, the greater is the risk of discontinuation, but the picture is not very clear. The approximate standard



**Figure 2.8** Kaplan-Meier type estimate of the hazard function for the data from Example 1.1.

errors of the estimated hazard function at different times are of little help in interpreting this plot.

In practice, estimates of the hazard function obtained in this way will often tend to be rather irregular. For this reason, plots of the hazard function may be “smoothed”, so that any pattern can be seen more clearly. There are a number of ways of smoothing the hazard function, that lead to a weighted average of values of the estimated hazard  $\hat{h}(t)$  at death times in the neighbourhood of  $t$ . For example, a *kernel smoothed* estimate of the hazard function, based on the  $r$  ordered death times,  $t_{(1)}, t_{(2)}, \dots, t_{(r)}$ , with  $d_j$  deaths and  $n_j$  at risk at time  $t_{(j)}$ , can be found from

$$h^\dagger(t) = b^{-1} \sum_{j=1}^r 0.75 \left\{ 1 - \left( \frac{t - t_{(j)}}{b} \right)^2 \right\} \frac{d_j}{n_j},$$

where the value of  $b$  needs to be chosen. The function  $h^\dagger(t)$  is defined for all values of  $t$  in the interval from  $b$  to  $t_{(r)} - b$ , where  $t_{(r)}$  is the greatest death time. For any value of  $t$  in this interval, the death times in the interval  $(t - b, t + b)$  will contribute to the weighted average. The parameter  $b$  is known as the *bandwidth* and its value controls the shape of the plot; the larger the value of  $b$ , the greater the degree of smoothing. There are formulae that lead to “optimal” values of  $b$ , but these tend to be rather cumbersome. Fuller details can be found in the references provided in the final section of this chapter. In this book, the use of a modelling approach to the analysis of survival data is advocated, and so model-based estimates of the hazard function will be considered in subsequent chapters.

### 2.3.3 Estimating the cumulative hazard function

The cumulative hazard function is important in the identification of models for survival data, as will be seen later in Sections 4.4 and 5.2. In addition, since the derivative of the cumulative hazard function is the hazard function itself, the slope of the cumulative hazard function provides information about the shape of the underlying hazard function. In particular, a linear cumulative hazard function over some time interval suggests that the hazard is constant over this interval. Accordingly, methods that can be used to estimate this function will now be described.

The cumulative hazard at time  $t$ ,  $H(t)$ , was defined in equation (1.6) to be the integral of the hazard function, but is more conveniently found using equation (1.7). According to this result,  $H(t) = -\log S(t)$ , and so if  $\hat{S}(t)$  is the Kaplan-Meier estimate of the survivor function,  $\hat{H}(t) = -\log \hat{S}(t)$  is an appropriate estimate of the cumulative hazard to time  $t$ .

Now, using equation (2.4),

$$\hat{H}(t) = - \sum_{j=1}^k \log \left( \frac{n_j - d_j}{n_j} \right),$$

for  $t_{(k)} \leq t < t_{(k+1)}$ ,  $k = 1, 2, \dots, r$ , and  $t_{(1)}, t_{(2)}, \dots, t_{(r)}$  are the  $r$  ordered death times, with  $t_{(r+1)} = \infty$ .

If the Nelson-Aalen estimate of the survivor function is used, the estimated cumulative hazard function,  $\tilde{H}(t) = -\log \tilde{S}(t)$ , is given by

$$\tilde{H}(t) = \sum_{j=1}^k \frac{d_j}{n_j}.$$

This is the cumulative sum of the estimated probabilities of death from the first to the  $k$ th time interval,  $k = 1, 2, \dots, r$ . This quantity therefore has immediate intuitive appeal as an estimate of the cumulative hazard.

An estimate of the cumulative hazard function also leads to an estimate of the corresponding hazard function, since the differences between adjacent values of the estimated cumulative hazard function provide estimates of the underlying hazard, after dividing by the time interval. In particular, differences in adjacent values of the Nelson-Aalen estimate of the cumulative hazard lead directly to the hazard function estimate in Section 2.3.2.

## 2.4 Estimating the median and percentiles of survival times

Since the distribution of survival times tends to be positively skew, the median is the preferred summary measure of the location of the distribution. Once the survivor function has been estimated, it is straightforward to obtain an estimate of the *median survival time*. This is the time beyond which 50% of the individuals in the population under study are expected to survive, and is given by that value  $t(50)$  which is such that  $S\{t(50)\} = 0.5$ .

Because the non-parametric estimates of  $S(t)$  are step-functions, it will

not usually be possible to realise an estimated survival time that makes the survivor function exactly equal to 0.5. Instead, the estimated median survival time,  $\hat{t}(50)$ , is defined to be the smallest observed survival time for which the value of the estimated survivor function is less than 0.5.

In mathematical terms,

$$\hat{t}(50) = \min\{t_i \mid \hat{S}(t_i) < 0.5\},$$

where  $t_i$  is the observed survival time for the  $i$ th individual,  $i = 1, 2, \dots, n$ . Since the estimated survivor function only changes at a death time, this is equivalent to the definition

$$\hat{t}(50) = \min\{t_{(j)} \mid \hat{S}(t_{(j)}) < 0.5\},$$

where  $t_{(j)}$  is the  $j$ th ordered death time,  $j = 1, 2, \dots, r$ .

In the particular case where the estimated survivor function is exactly equal to 0.5 for values of  $t$  in the interval from  $t_{(j)}$  to  $t_{(j+1)}$ , the median is taken to be the half-way point in this interval, that is  $(t_{(j)} + t_{(j+1)})/2$ .

In the situation where there are no censored survival times, the estimated median survival time will be the smallest time beyond which 50% of the individuals in the sample survive.

#### *Example 2.8 Time to discontinuation of the use of an IUD*

The Kaplan-Meier estimate of the survivor function for the data from Example 1.1 on the time to discontinuation of the use of an IUD was given in Table 2.2. The estimated survivor function,  $\hat{S}(t)$ , for these data was shown in Figure 2.4. From the estimated survivor function, the smallest discontinuation time beyond which the estimated probability of discontinuation is less than 0.5 is 93 weeks. This is therefore the estimated median time to discontinuation of the IUD for this group of women.

A similar procedure to that described above can be used to estimate other *percentiles* of the distribution of survival times. The  $p$ th percentile of the distribution of survival times is defined to be the value  $t(p)$  which is such that  $F\{t(p)\} = p/100$ . In terms of the survivor function,  $t(p)$  is such that  $S\{t(p)\} = 1 - (p/100)$ , so that for example the 10th and 90th percentiles are given by

$$S\{t(10)\} = 0.9, \quad S\{t(90)\} = 0.1,$$

respectively. Using the estimated survivor function, the estimated  $p$ th percentile is the smallest observed survival time,  $\hat{t}(p)$ , for which  $\hat{S}\{\hat{t}(p)\} < 1 - (p/100)$ .

It sometimes happens that the estimated survivor function is greater than 0.5 for all values of  $t$ . In such cases, the median survival time cannot be estimated. It would then be natural to summarise the data in terms of other percentiles of the distribution of survival times, or the estimated survival probabilities at particular time points.

Estimates of the dispersion of a sample of survival data are not widely used, but should such an estimate be required, the *semi-interquartile range* (SIQR)

can be calculated. This is defined to be half the difference between the 75th and 25th percentiles of the distribution of survival times. Hence,

$$\text{SIQR} = \frac{1}{2} \{t(75) - t(25)\},$$

where  $t(25)$  and  $t(75)$  are the 25th and 75th percentiles of the survival time distribution. These two percentiles are also known as the *first* and *third quartiles*, respectively. The corresponding sample-based estimate of the SIQR is  $\{\hat{t}(75) - \hat{t}(25)\}/2$ . Like the variance, the larger the value of the SIQR, the more dispersed is the survival time distribution.

#### *Example 2.9 Time to discontinuation of the use of an IUD*

From the Kaplan-Meier estimate of the survivor function for the data from Example 1.1, given in Table 2.2, the 25th and 75th percentiles of the distribution of discontinuation times are 36 and 107 weeks, respectively. Hence, the SIQR of the distribution is estimated to be 35.5 weeks.

#### 2.5\* Confidence intervals for the median and percentiles

Approximate confidence intervals for the median and other percentiles of a distribution of survival times can be found once the variance of the estimated percentile has been obtained. An expression for the approximate variance of a percentile can be derived from a direct application of the general result for the variance of a function of a random variable in equation (2.9). Using this result,

$$\text{var}[\hat{S}\{t(p)\}] = \left( \frac{d\hat{S}\{t(p)\}}{dt(p)} \right)^2 \text{var}\{t(p)\}, \quad (2.17)$$

where  $t(p)$  is the  $p$ th percentile of the distribution and  $\hat{S}\{t(p)\}$  is the Kaplan-Meier estimate of the survivor function at  $t(p)$ . Now,

$$-\frac{d\hat{S}\{t(p)\}}{dt(p)} = \hat{f}\{t(p)\},$$

an estimate of the probability density function of the survival times at  $t(p)$ , and on rearranging equation (2.17), we get

$$\text{var}\{t(p)\} = \left( \frac{1}{\hat{f}\{t(p)\}} \right)^2 \text{var}[\hat{S}\{t(p)\}].$$

The standard error of  $\hat{t}(p)$ , the estimated  $p$ th percentile, is therefore given by

$$\text{se}\{\hat{t}(p)\} = \frac{1}{\hat{f}\{\hat{t}(p)\}} \text{se}[\hat{S}\{\hat{t}(p)\}]. \quad (2.18)$$

The standard error of  $\hat{S}\{\hat{t}(p)\}$  is found using Greenwood's formula for the standard error of the Kaplan-Meier estimate of the survivor function, given in equation (2.13), while an estimate of the probability density function at  $\hat{t}(p)$

is

$$\hat{f}\{\hat{t}(p)\} = \frac{\hat{S}\{\hat{u}(p)\} - \hat{S}\{\hat{l}(p)\}}{\hat{l}(p) - \hat{u}(p)},$$

where

$$\hat{u}(p) = \max \left\{ t_{(j)} \mid \hat{S}(t_{(j)}) \geq 1 - \frac{p}{100} + \epsilon \right\},$$

and

$$\hat{l}(p) = \min \left\{ t_{(j)} \mid \hat{S}(t_{(j)}) \leq 1 - \frac{p}{100} - \epsilon \right\},$$

for  $j = 1, 2, \dots, r$ , and small values of  $\epsilon$ . In many cases, taking  $\epsilon = 0.05$  will be satisfactory, but a larger value of  $\epsilon$  will be needed if  $\hat{u}(p)$  and  $\hat{l}(p)$  turn out to be equal. In particular, from equation (2.18), the standard error of the median survival time is given by

$$\text{se}\{\hat{t}(50)\} = \frac{1}{\hat{f}\{\hat{t}(50)\}} \text{se}[\hat{S}\{\hat{t}(50)\}], \quad (2.19)$$

where  $\hat{f}\{\hat{t}(50)\}$  can be found from

$$\hat{f}\{\hat{t}(50)\} = \frac{\hat{S}\{\hat{u}(50)\} - \hat{S}\{\hat{l}(50)\}}{\hat{l}(50) - \hat{u}(50)}. \quad (2.20)$$

In this expression,  $\hat{u}(50)$  is the largest survival time for which the Kaplan-Meier estimate of the survivor function exceeds 0.55, and  $\hat{l}(50)$  is the smallest survival time for which the survivor function is less than or equal to 0.45.

Once the standard error of the estimated  $p$ th percentile has been found, a  $100(1 - \alpha)\%$  confidence interval for  $t(p)$  has limits of

$$\hat{t}(p) \pm z_{\alpha/2} \text{se}\{\hat{t}(p)\},$$

where  $z_{\alpha/2}$  is the upper (one-sided)  $\alpha/2$ -point of the standard normal distribution.

This interval estimate is only approximate, in the sense that the probability that the interval includes the true percentile will not be exactly  $1 - \alpha$ . A number of methods have been proposed for constructing confidence intervals for the median with superior properties, although these alternatives are more difficult to compute than the interval estimate derived in this section.

#### *Example 2.10 Time to discontinuation of the use of an IUD*

The data on the discontinuation times for users of an IUD, given in Example 1.1, is now used to illustrate the calculation of a confidence interval for the median discontinuation time. From Example 2.8, the estimated median discontinuation time for this group of women is given by  $\hat{t}(50) = 93$  weeks. Also, from Table 2.4, the standard error of the Kaplan-Meier estimate of the survivor function at this time is given by  $\text{se}[\hat{S}\{\hat{t}(50)\}] = 0.1452$ .

To obtain the standard error of  $\hat{t}(50)$  using equation (2.19), we need an estimate of the density function at the estimated median discontinuation time. This is obtained from equation (2.20). The quantities  $\hat{u}(50)$  and  $\hat{l}(50)$  needed

in this equation are such that

$$\hat{u}(50) = \max\{t_{(j)} \mid \hat{S}(t_{(j)}) \geq 0.55\},$$

and

$$\hat{l}(50) = \min\{t_{(j)} \mid \hat{S}(t_{(j)}) \leq 0.45\},$$

where  $t_{(j)}$  is the  $j$ th ordered discontinuation time,  $j = 1, 2, \dots, 9$ . Using Table 2.4,  $\hat{u}(50) = 75$  and  $\hat{l}(50) = 97$ , and so

$$\hat{f}\{\hat{t}(50)\} = \frac{\hat{S}(75) - \hat{S}(97)}{97 - 75} = \frac{0.5594 - 0.3729}{22} = 0.0085.$$

Then, the standard error of the median is given by

$$\text{se}\{\hat{t}(50)\} = \frac{1}{0.0085} \times 0.1452 = 17.13.$$

A 95% confidence interval for the median discontinuation time has limits of  $93 \pm 1.96 \times 17.13$ ,

and so the required interval estimate for the median ranges from 59 to 127 days.

#### 2.6 Comparison of two groups of survival data

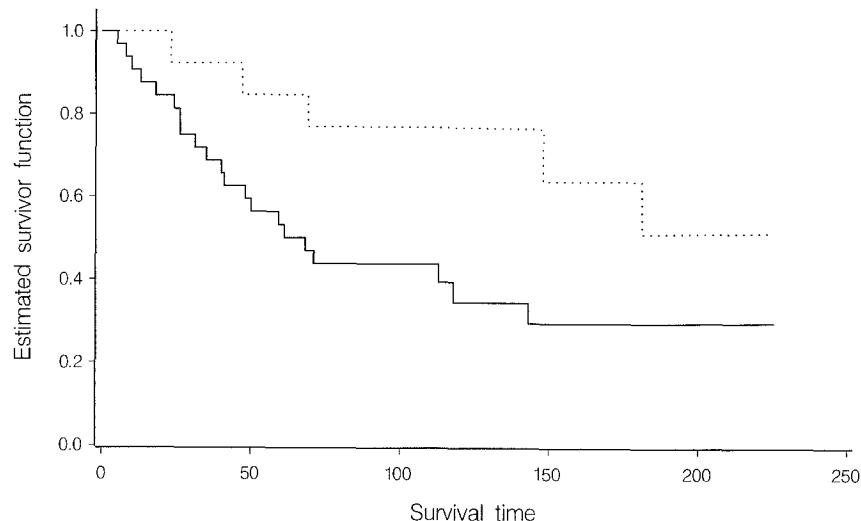
The simplest way of comparing the survival times obtained from two groups of individuals is to plot the corresponding estimates of the two survivor functions on the same axes. The resulting plot can be quite informative, as the following example illustrates.

##### *Example 2.11 Prognosis for women with breast cancer*

Data on the survival times of women with breast cancer, grouped according to whether or not sections of a tumour were positively stained with HPA, were given in Example 1.2. The Kaplan-Meier estimate of the survivor function, for each of the two groups of survival times, is plotted in Figure 2.9. Notice that in this figure, the Kaplan-Meier estimates extend to the time of the largest censored observation in each group.

This figure shows that the estimated survivor function for those women with negatively stained tumours is always greater than that for women with positively stained tumours. This means that at any time  $t$ , the estimated probability of survival beyond  $t$  is greater for women with negative staining, suggesting that the result of the HPA staining procedure might be a useful prognostic indicator. In particular, those women whose tumours are positively stained appear to have a poorer prognosis than those with negatively stained tumours.

There are two possible explanations for an observed difference between two estimated survivor functions, such as those in Example 2.11. One explanation is that there is a real difference between the survival times of the two groups of individuals, so that those in one group have a different survival experience



**Figure 2.9** Kaplan-Meier estimate of the survivor functions for women with tumours that were positively stained (—) and negatively stained (···).

from those in the other. An alternative explanation is that there are no real differences between the survival times in each group, and that the difference that has been observed is merely the result of chance variation. To help distinguish between these two possible explanations, we use a procedure known as the *hypothesis test*. Because the concept of the hypothesis test has a central role in the analysis of survival data, the underlying basis for this procedure is described in detail in the following section.

### 2.6.1 Hypothesis testing

The hypothesis test is a procedure that enables us to assess the extent to which an observed set of data are consistent with a particular hypothesis, known as the *working* or *null hypothesis*. A null hypothesis generally represents a simplified view of the data-generating process, and is typified by hypotheses that specify that there is no difference between two groups of survival data, or that there is no relationship between survival time and explanatory variables such as age or serum cholesterol level. The null hypothesis is then the hypothesis that will be adopted, and subsequently acted upon, unless the data indicate that it is untenable.

The next step is to formulate a *test statistic* that measures the extent to which the observed data depart from the null hypothesis. In general, the test statistic is so constructed that the larger the value of the statistic, the greater the departure from the null hypothesis. Hence, if the null hypothesis is that there is no difference between two groups, relatively large values of the test statistic will be interpreted as evidence against this null hypothesis.

Once the value of the test statistic has been obtained from the observed data, we calculate the probability of obtaining a value as extreme or more extreme than the observed value, when the null hypothesis is true. This quantity summarises the strength of the evidence in the sample data against the null hypothesis, and is known as the *probability value*, or *P-value* for short. If the *P*-value is large, we would conclude that it is quite likely that the observed data would have been obtained when the null hypothesis was true, and that there is no evidence to reject the null hypothesis. On the other hand, if the *P*-value is small, this would be interpreted as evidence against the null hypothesis; the smaller the *P*-value, the stronger the evidence.

In order to obtain the *P*-value for a hypothesis test, the test statistic must have a probability distribution that is known, or at least approximately known, when the null hypothesis is true. This probability distribution is referred to as the *null distribution* of the test statistic. More specifically, consider a test statistic,  $W$ , which is such that the larger the observed value of the test statistic,  $w$ , the greater the deviation of the observed data from that expected under the null hypothesis. If  $W$  has a continuous probability distribution, the *P*-value is then  $P(W \geq w) = 1 - F(w)$ , where  $F(w)$  is the distribution function of  $W$ , under the null hypothesis, evaluated at  $w$ .

In some applications, the most natural test statistic is one for which large positive values correspond to departures from the null hypothesis in one direction, while large negative values correspond to departures in the opposite direction. For example, suppose that patients suffering from a particular illness have been randomised to receive either a standard treatment or a new treatment, and their survival times are recorded. In this situation, a null hypothesis of interest will be that there is no difference in the survival experience of the patients in the two treatment groups. The extent to which the data are consistent with this null hypothesis might then be summarised by a test statistic for which positive values indicate that the new treatment is superior to the standard, while negative values indicate that the standard treatment is superior. When departures from the null hypothesis in either direction are equally important, the null hypothesis is said to have a *two-sided alternative*, and the hypothesis test itself is referred to as a *two-sided test*.

If  $W$  is a test statistic for which large positive or large negative observed values lead to rejection of the null hypothesis, a new test statistic, such as  $|W|$  or  $W^2$ , can be defined, so that only large positive values of the new statistic indicate that there is evidence against the null hypothesis. For example, suppose that  $W$  is a test statistic that under the null hypothesis has a standard normal distribution. If  $w$  is the observed value of  $W$ , the appropriate *P*-value is  $P(W \leq -|w|) + P(W \geq |w|)$ , which in view of the symmetry of the standard normal distribution, is  $2P(W \geq |w|)$ . Alternatively, we can make use of the result that if  $W$  has a standard normal distribution,  $W^2$  has a chi-squared distribution on one degree of freedom, written  $\chi_1^2$ . Thus a *P*-value for the two-sided hypothesis test based on the statistic  $W$  is the probability that a  $\chi_1^2$  random variable exceeds  $w^2$ . The required *P*-value can therefore be found using tables of the standard normal or chi-squared distribution functions.

When interest centres on departures in a particular direction, the hypothesis test is said to be *one-sided*. For example, in comparing the survival times of two groups of patients where one group receives a standard treatment and the other group a new treatment, it might be argued that the new treatment cannot possibly be inferior to the standard. Then, the only relevant alternative to the null hypothesis of no treatment difference is that the new treatment is superior. If positive values of the test statistic  $W$  reflect the superiority of the new treatment, the  $P$ -value is then  $P(W \geq w)$ . If  $W$  has a standard normal distribution, this  $P$ -value is half of that which would have been obtained for the corresponding two-sided alternative hypothesis.

A one-sided hypothesis test can only be appropriate when there is no interest whatsoever in departures from the null hypothesis in the opposite direction to that specified in the one-sided alternative. For example, consider again the comparison of a new treatment with a standard treatment, and suppose that the observed value of the test statistic is either positive or negative, depending on whether the new treatment is superior or inferior to the standard. If the alternative to the null hypothesis of no treatment difference is that the new treatment is superior, a large negative value of the test statistic would not be regarded as evidence against the null hypothesis. Instead, it would be assumed that this large negative value is simply the result of chance variation. Generally speaking, the use of one-sided tests can rarely be justified in medical research, and so two-sided tests will be used throughout this book.

If a  $P$ -value is smaller than some value  $\alpha$ , we say that the hypothesis is rejected at the  $100\alpha\%$  level of significance. The observed value of the test statistic is then said to be significant at this level. But how do we decide on the basis of the  $P$ -value whether or not a null hypothesis should actually be rejected? Traditionally,  $P$ -values of 0.05 or 0.01 have been used in reaching a decision about whether or not a null hypothesis should be rejected, so that if  $P < 0.05$ , for example, the null hypothesis is rejected at the 5% significance level. Guidelines such as these are not hard-and-fast rules and should not be interpreted rigidly. For example, there is no practical difference between a  $P$ -value of 0.046 and 0.056, even though only the former indicates that the observed value of the test statistic is significant at the 5% level.

Instead of reporting that a null hypothesis is rejected or not rejected at some specified significance level, a more satisfactory policy is to report the actual  $P$ -value. This  $P$ -value can then be interpreted as a measure of the strength of evidence against the null hypothesis, using a vocabulary that depends on the range within which the  $P$ -value lies. Thus, if  $P > 0.1$ , there is said to be no evidence to reject the null hypothesis; if  $0.05 < P \leq 0.1$ , there is slight evidence against the null hypothesis; if  $0.01 < P \leq 0.05$ , there is moderate evidence against the null hypothesis; if  $0.001 < P \leq 0.01$ , there is strong evidence against the null hypothesis, and if  $P \leq 0.001$ , the evidence against the null hypothesis is overwhelming.

An alternative to quoting the exact  $P$ -value associated with a hypothesis test, is to compare the observed value of the test statistic with those values that would correspond to particular  $P$ -values, when the null hypothesis is

true. Values of the test statistic that lead to rejection of the null hypothesis at particular levels of significance can be found from tables of the percentage points of the null distribution of that statistic. In particular, if  $W$  is a test statistic that has a standard normal distribution, for a two-sided test, the upper  $\alpha/2$ -point of the distribution, depicted in Figure 2.5, is the value of the test statistic for which the  $P$ -value is  $\alpha$ . For example, values of the test statistic of 1.96, 2.58 and 3.29 correspond to  $P$ -values of 0.05, 0.01 and 0.001. Thus, if the observed value of  $W$  were between 1.96 and 2.58, we would declare that  $0.01 < P < 0.05$ . On the other hand, if the null distribution of  $W$  is chi-squared on one degree of freedom, the upper  $\alpha$ -point of the distribution is the value of the test statistic which would give a  $P$ -value of  $\alpha$ . Then, values of the test statistic of 3.84, 6.64 and 10.83 correspond to  $P$ -values of 0.05, 0.01 and 0.001, respectively. Notice that these values are simply the squares of those for the standard normal distribution, which they must be in view of the fact that the square of a standard normal random variable has a chi-squared distribution on one degree of freedom.

For commonly encountered probability distributions, such as the normal and chi-squared, percentage points are tabulated in many introductory text books on statistics, or in statistical tables such as those of Lindley and Scott (1984). Statistical software packages used in computer-based statistical analyses of survival data usually provide the exact  $P$ -values associated with hypothesis tests as a matter of course. Note that when these are rounded off to, say, three decimal places, a  $P$ -value of 0.000 should be interpreted as  $P < 0.001$ .

In deciding on a course of action, such as whether or not to reject the hypothesis that there is no difference between two treatments, the statistical evidence summarised in the  $P$ -value for the hypothesis test will be just one ingredient of the decision-making process. In addition to the statistical evidence, there will also be scientific evidence to consider. This may, for example, concern whether the size of the treatment effect is clinically important. In particular, in a large trial, a difference between two treatments that is significant at, say, the 5% level may be found when the magnitude of the treatment effect is so small that it does not indicate a major scientific breakthrough. On the other hand, a new formulation of a treatment may prolong life by a factor of two, and yet, because of small sample sizes used in the study, may not appear to be significantly different from the standard.

Rather than report findings in terms of the results of a hypothesis testing procedure, it is more informative to provide an estimate of the size of any treatment difference, supported by a confidence interval for this difference. Unfortunately, the non-parametric approaches to the analysis of survival data being considered in this chapter do not lend themselves to this approach. We will therefore return to this theme in subsequent chapters when we consider models for survival data.

In the comparison of two groups of survival data, there are a number of methods that can be used to quantify the extent of between-group differences. Two non-parametric procedures will now be considered, namely the log-rank test and the Wilcoxon test.

### 2.6.2 The log-rank test

In order to construct the log-rank test, we begin by considering separately each death time in two groups of survival data. These groups will be labelled Group I and Group II. Suppose that there are  $r$  distinct death times,  $t_{(1)} < t_{(2)} < \dots < t_{(r)}$ , across the two groups, and that at time  $t_{(j)}$ ,  $d_{1j}$  individuals in Group I and  $d_{2j}$  individuals in Group II die, for  $j = 1, 2, \dots, r$ . Unless two or more individuals in a group have the same recorded death time, the values of  $d_{1j}$  and  $d_{2j}$  will either be zero or unity. Suppose further that there are  $n_{1j}$  individuals at risk of death in the first group just before time  $t_{(j)}$ , and that there are  $n_{2j}$  at risk in the second group. Consequently, at time  $t_{(j)}$ , there are  $d_j = d_{1j} + d_{2j}$  deaths in total out of  $n_j = n_{1j} + n_{2j}$  individuals at risk. The situation is summarised in Table 2.7.

**Table 2.7** Number of deaths at the  $j$ th death time in each of two groups of individuals.

Group	Number of deaths at $t_{(j)}$	Number surviving beyond $t_{(j)}$	Number at risk just before $t_{(j)}$
I	$d_{1j}$	$n_{1j} - d_{1j}$	$n_{1j}$
II	$d_{2j}$	$n_{2j} - d_{2j}$	$n_{2j}$
Total	$d_j$	$n_j - d_j$	$n_j$

Now consider the null hypothesis that there is no difference in the survival experiences of the individuals in the two groups. One way of assessing the validity of this hypothesis is to consider the extent of the difference between the observed number of individuals in the two groups who die at each of the death times, and the numbers expected under the null hypothesis. Information about the extent of these differences can then be combined over each of the death times.

If the marginal totals in Table 2.7 are regarded as fixed, and the null hypothesis that survival is independent of group is true, the four entries in this table are solely determined by the value of  $d_{1j}$ , the number of deaths at  $t_{(j)}$  in Group I. We can therefore regard  $d_{1j}$  as a random variable, which can take any value in the range from 0 to the minimum of  $d_j$  and  $n_{1j}$ . In fact,  $d_{1j}$  has a distribution known as the *hypergeometric distribution*, according to which the probability that the random variable associated with the number of deaths in the first group takes the value  $d_{1j}$  is

$$\frac{\binom{d_j}{d_{1j}} \binom{n_j - d_j}{n_{1j} - d_{1j}}}{\binom{n_j}{n_{1j}}}. \quad (2.21)$$

In this formula, the expression

$$\binom{d_j}{d_{1j}}$$

represents the number of different ways in which  $d_{1j}$  times can be chosen from  $d_j$  times and is read as “ $d_{1j}$  C  $d_j$ ”. It is given by

$$\binom{d_j}{d_{1j}} = \frac{d_j!}{d_{1j}!(d_j - d_{1j})!},$$

where  $d_j!$ , read as “ $d_j$  factorial”, is such that

$$d_j! = d_j \times (d_j - 1) \times \dots \times 2 \times 1.$$

The other two terms in expression (2.21) are interpreted in a similar manner.

The mean of the hypergeometric random variable  $d_{1j}$  is given by

$$e_{1j} = n_{1j}d_j/n_j, \quad (2.22)$$

so that  $e_{1j}$  is the expected number of individuals who die at time  $t_{(j)}$  in Group I. This value is intuitively appealing, since under the null hypothesis that the probability of death at time  $t_{(j)}$  does not depend on the group that an individual is in, the probability of death at  $t_{(j)}$  is  $d_j/n_j$ . Multiplying this by  $n_{1j}$ , gives  $e_{1j}$  as the expected number of deaths in Group I at  $t_{(j)}$ .

The next step is to combine the information from the individual  $2 \times 2$  tables for each death time to give an overall measure of the deviation of the observed values of  $d_{1j}$  from their expected values. The most straightforward way of doing this is to sum the differences  $d_{1j} - e_{1j}$  over the total number of death times,  $r$ , in the two groups. The resulting statistic is given by

$$U_L = \sum_{j=1}^r (d_{1j} - e_{1j}). \quad (2.23)$$

Notice that this is  $\sum d_{1j} - \sum e_{1j}$ , which is the difference between the total observed and expected numbers of deaths in Group I. This statistic will have zero mean, since  $E(d_{1j}) = e_{1j}$ . Moreover, since the death times are independent of one another, the variance of  $U_L$  is simply the sum of the variances of the  $d_{1j}$ . Now, since  $d_{1j}$  has a hypergeometric distribution, the variance of  $d_{1j}$  is given by

$$v_{1j} = \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)}, \quad (2.24)$$

so that the variance of  $U_L$  is

$$\text{var}(U_L) = \sum_{j=1}^r v_{1j} = V_L, \quad (2.25)$$

say. Furthermore, it can be shown that  $U_L$  has an approximate normal distribution, when the number of death times is not too small. It then follows that  $U_L/\sqrt{V_L}$  has a normal distribution with zero mean and unit variance,

denoted  $N(0, 1)$ . We therefore write

$$\frac{U_L}{\sqrt{V_L}} \sim N(0, 1),$$

where the symbol “ $\sim$ ” is read as “is distributed as”. The square of a standard normal random variable has a chi-squared distribution on one degree of freedom, denoted  $\chi_1^2$ , and so we have that

$$\frac{U_L^2}{V_L} \sim \chi_1^2. \quad (2.26)$$

This method of combining information over a number of  $2 \times 2$  tables was proposed by Mantel and Haenszel (1959), and is known as the Mantel-Haenszel procedure. In fact, the test based on this statistic has various names, including *Mantel-Cox* and *Peto-Mantel-Haenszel*, but it is probably best known as the log-rank test. The reason for this name is that the test statistic can be derived from the ranks of the survival times in the two groups, and the resulting rank test statistic is based on the logarithm of the Nelson-Aalen estimate of the survivor function.

The statistic  $W_L = U_L^2/V_L$  summarises the extent to which the observed survival times in the two groups of data deviate from those expected under the null hypothesis of no group differences. The larger the value of this statistic, the greater the evidence against the null hypothesis. Because the null distribution of  $W$  is approximately chi-squared with one degree of freedom, the  $P$ -value associated with the test statistic can be obtained from the distribution function of a chi-squared random variable. Alternatively, percentage points of the chi-squared distribution can be used to identify a range within which the  $P$ -value lies. An illustration of the log-rank test is presented below in Example 2.12.

#### *Example 2.12 Prognosis for women with breast cancer*

In this example, we return to the data on the survival times of women with breast cancer, grouped according to whether a section of the tumour was positively or negatively stained. In particular the null hypothesis that there is no difference in the survival experience of the two groups will be examined using the log-rank test. The required calculations are laid out in Table 2.8.

We begin by ordering the observed death times across the two groups of women; these times are given in column 1 of Table 2.8. The numbers of women in each group who die at each death time and the numbers who are at risk at each time are then calculated. These values are  $d_{1j}$ ,  $n_{1j}$ ,  $d_{2j}$  and  $n_{2j}$  given in columns 2 to 5 of the table. Columns 6 and 7 contain the total numbers of deaths and the total numbers of women at risk over the two groups, at each death time. The final two columns give the values of  $e_{1j}$  and  $v_{1j}$ , computed from equations (2.22) and (2.24) respectively. Summing the entries in columns 2 and 8 gives  $\sum d_{1j}$  and  $\sum e_{1j}$ , from which the log-rank statistic can be calculated from  $U_L = \sum d_{1j} - \sum e_{1j}$ . The value of  $V_L = \sum v_{1j}$  can be obtained by summing the entries in the final column. We find that

**Table 2.8** Calculation of the log-rank statistic for the data from Example 1.2.

Death time	$d_{1j}$	$n_{1j}$	$d_{2j}$	$n_{2j}$	$d_j$	$n_j$	$e_{1j}$	$v_{1j}$
5	0	13	1	32	1	45	0.2889	0.2054
8	0	13	1	31	1	44	0.2955	0.2082
10	0	13	1	30	1	43	0.3023	0.2109
13	0	13	1	29	1	42	0.3095	0.2137
18	0	13	1	28	1	41	0.3171	0.2165
23	1	13	0	27	1	40	0.3250	0.2194
24	0	12	1	27	1	39	0.3077	0.2130
26	0	12	2	26	2	38	0.6316	0.4205
31	0	12	1	24	1	36	0.3333	0.2222
35	0	12	1	23	1	35	0.3429	0.2253
40	0	12	1	22	1	34	0.3529	0.2284
41	0	12	1	21	1	33	0.3636	0.2314
47	1	12	0	20	1	32	0.3750	0.2344
48	0	11	1	20	1	31	0.3548	0.2289
50	0	11	1	19	1	30	0.3667	0.2322
59	0	11	1	18	1	29	0.3793	0.2354
61	0	11	1	17	1	28	0.3929	0.2385
68	0	11	1	16	1	27	0.4074	0.2414
69	1	11	0	15	1	26	0.4231	0.2441
71	0	9	1	15	1	24	0.3750	0.2344
113	0	6	1	10	1	16	0.3750	0.2344
118	0	6	1	8	1	14	0.4286	0.2449
143	0	6	1	7	1	13	0.4615	0.2485
148	1	6	0	6	1	12	0.5000	0.2500
181	1	5	0	4	1	9	0.5556	0.2469
Total	5						9.5652	5.9289

$U_L = 5 - 9.565 = -4.565$  and  $V_L = 5.929$ , and so the value of the log-rank test statistic is  $W_L = (-4.565)^2/5.929 = 3.515$ .

The corresponding  $P$ -value is calculated from the probability that a chi-squared variate on one degree of freedom is greater than or equal to 3.515, and is 0.061, written  $P = 0.061$ . This  $P$ -value is sufficiently small to cast doubt on the null hypothesis that there is no difference between the survivor functions for the two groups of women. In fact, the evidence against the null hypothesis is nearly significant at the 6% level. We therefore conclude that the data do provide some evidence that the prognosis of a breast cancer patient is dependent on the result of the staining procedure.

#### *2.6.3 The Wilcoxon test*

The Wilcoxon test, sometimes known as the *Breslow test*, is also used to test the null hypothesis that there is no difference in the survivor functions for two

groups of survival data. The Wilcoxon test is based on the statistic

$$U_W = \sum_{j=1}^r n_j(d_{1j} - e_{1j}),$$

where, as in the previous section,  $d_{1j}$  is the number of deaths at time  $t_{(j)}$  in the first group and  $e_{1j}$  is as defined in equation (2.22). The difference between  $U_W$  and  $U_L$  is that in the Wilcoxon test, each difference  $d_{1j} - e_{1j}$  is weighted by  $n_j$ , the total number of individuals at risk at time  $t_{(j)}$ . The effect of this is to give less weight to differences between  $d_{1j}$  and  $e_{1j}$  at those times when the total number of individuals who are still alive is small, that is, at the longest survival times. This statistic is therefore less sensitive than the log-rank statistic to deviations of  $d_{1j}$  from  $e_{1j}$  in the tail of the distribution of survival times.

The variance of the Wilcoxon statistic  $U_W$  is given by

$$V_W = \sum_{j=1}^r n_j^2 v_{1j},$$

where  $v_{1j}$  is given in equation (2.24), and so the Wilcoxon test statistic is

$$W_W = U_W^2/V_W, \quad (2.27)$$

which has a chi-squared distribution on one degree of freedom when the null hypothesis is true. The Wilcoxon test is therefore conducted in the same manner as the log-rank test.

#### *Example 2.13 Prognosis for women with breast cancer*

For the data on the survival times of women with tumours that were positively or negatively stained, the value of the Wilcoxon statistic is  $U_W = -159$ , and the variance of the statistic is  $V_W = 6048.136$ . The value of the chi-squared statistic,  $U_W^2/V_W$ , is 4.180, and the corresponding  $P$ -value is 0.041. This is slightly smaller than the  $P$ -value for the log-rank test, and on the basis of this result, we would declare that the difference between the two groups is significant at the 5% level.

#### *2.6.4 Comparison of the log-rank and Wilcoxon tests*

Of the two tests, the log-rank test is the more suitable when the alternative to the null hypothesis of no difference between two groups of survival times is that the hazard of death at any given time for an individual in one group is proportional to the hazard at that time for a similar individual in the other group. This is the assumption of *proportional hazards*, which underlies a number of methods for analysing survival data. For other types of departure from the null hypothesis, the Wilcoxon test is more appropriate than the log-rank test for comparing the two survivor functions.

In order to help decide which test is the more suitable in any given situation, we make use of the result that if the hazard functions are proportional,

the survivor functions for the two groups of survival data do not cross one another. To show this, suppose that  $h_1(t)$  is the hazard of death at time  $t$  for an individual in Group I, and  $h_2(t)$  is the hazard at that same time for an individual in Group II. If these two hazards are proportional, then we can write  $h_1(t) = \psi h_2(t)$ , where  $\psi$  is a constant that does not depend on the time  $t$ . Integrating both sides of this expression, multiplying by  $-1$  and exponentiating gives

$$\exp \left\{ - \int_0^t h_1(u) du \right\} = \exp \left\{ - \int_0^t \psi h_2(u) du \right\}. \quad (2.28)$$

Now, from equation (1.5),

$$S(t) = \exp \left\{ - \int_0^t h(u) du \right\},$$

and so if  $S_1(t)$  and  $S_2(t)$  are the survivor functions for the two groups of survival data, from equation (2.28),

$$S_1(t) = \{S_2(t)\}^\psi.$$

Since the survivor function takes values between zero and unity, this result shows that  $S_1(t)$  is greater than or less than  $S_2(t)$ , according to whether  $\psi$  is less than or greater than unity, at any time  $t$ . This means that if two hazard functions are proportional, the true survivor functions do not cross. This is a necessary, but not a sufficient condition for proportional hazards.

An informal assessment of the likely validity of the proportional hazards assumption can be made from a plot of the estimated survivor functions for two groups of survival data, such as that shown in Figure 2.9. If the two estimated survivor functions do not cross, the assumption of proportional hazards may be justified, and the log-rank test is appropriate. Of course, sample-based estimates of survivor functions may cross even though the corresponding true hazard functions are proportional, and so some care is needed in the interpretation of such graphs. A more satisfactory graphical method for assessing the validity of the proportional hazards assumption is described in Section 4.4.1 of Chapter 4.

In summary, unless a plot of the estimated survival functions, or previous data, indicate that there is good reason to doubt the proportional hazards assumption, the log-rank test should be used to test the hypothesis of equality of two survivor functions.

#### *Example 2.14 Prognosis for women with breast cancer*

From the graph of the two estimated survivor functions in Figure 2.9, we see that the survivor function for the negatively stained women always lies above that for the positively stained women. This suggests that the proportional hazards assumption is appropriate, and that the log-rank test is more appropriate than the Wilcoxon test. However, in this example, there is very little difference between the results of the two hypothesis tests.

## 2.7\* Comparison of three or more groups of survival data

Both the log-rank and the Wilcoxon tests can be extended to enable three or more groups of survival data to be compared. Suppose that the survival distributions of  $g$  groups of survival data are to be compared, for  $g \geq 2$ . We then define analogues of the  $U$ -statistics for comparing the observed numbers of deaths in groups  $1, 2, \dots, g - 1$  with their expected values. In an obvious extension of the notation used in Section 2.6, we obtain

$$U_{Lk} = \sum_{j=1}^r \left( d_{kj} - \frac{n_{kj}d_j}{n_j} \right),$$

$$U_{Wk} = \sum_{j=1}^r n_j \left( d_{kj} - \frac{n_{kj}d_j}{n_j} \right),$$

for  $k = 1, 2, \dots, g - 1$ . These quantities are then expressed in the form of a vector with  $(g - 1)$  components, which we denote by  $\mathbf{U}_L$  and  $\mathbf{U}_W$ .

We also need expressions for the variances of the  $U_{Lk}$  and  $U_{Wk}$ , and for the covariance between pairs of values. In particular, the covariance between  $U_{Lk}$  and  $U_{Lk'}$  is given by

$$V_{Lkk'} = \sum_{j=1}^r \frac{n_{kj}d_j(n_j - d_j)}{n_j(n_j - 1)} \left( \delta_{kk'} - \frac{n_{k'j}}{n_j} \right),$$

for  $k, k' = 1, 2, \dots, g - 1$ , where  $\delta_{kk'}$  is such that

$$\delta_{kk'} = \begin{cases} 1 & \text{if } k = k', \\ 0 & \text{otherwise.} \end{cases}$$

These terms are then assembled in the form of a *variance-covariance matrix*,  $\mathbf{V}_L$ , which is a symmetric matrix that has the variances of the  $U_{Lk}$  down the diagonal, and covariance terms in the off-diagonals. For example, in the comparison of three groups of survival data, this matrix would be given by

$$\mathbf{V}_L = \begin{pmatrix} V_{L11} & V_{L12} \\ V_{L12} & V_{L22} \end{pmatrix},$$

where  $V_{L11}$  and  $V_{L22}$  are the variances of  $U_{L1}$  and  $U_{L2}$ , respectively, and  $V_{L12}$  is their covariance.

Similarly, the variance-covariance matrix for the Wilcoxon statistic is the matrix  $\mathbf{V}_W$ , whose  $(k, k')$ th element is

$$V_{Wkk'} = \sum_{j=1}^r n_j^2 \frac{n_{kj}d_j(n_j - d_j)}{n_j(n_j - 1)} \left( \delta_{kk'} - \frac{n_{k'j}}{n_j} \right),$$

for  $k, k' = 1, 2, \dots, g - 1$ .

Finally, in order to test the null hypothesis of no group differences, we make use of the result that the test statistic  $\mathbf{U}'_L \mathbf{V}_L^{-1} \mathbf{U}_L$ , or  $\mathbf{U}'_W \mathbf{V}_W^{-1} \mathbf{U}_W$ , has a chi-squared distribution on  $(g - 1)$  degrees of freedom, when the null hypothesis is true.

A number of well-known statistical packages for the analysis of survival data incorporate this methodology. Furthermore, because the interpretation of the resulting chi-squared statistic is straightforward, an example will not be given here.

## 2.8 Stratified tests

In many circumstances, there is a need to compare two or more sets of survival data, after taking account of additional variables recorded on each individual. As an illustration, consider a multicentred clinical trial in which two forms of chemotherapy are to be compared in terms of their effect on the survival times of lung cancer patients. Information on the survival times of patients in each treatment group will be available from each centre. The resulting data are then said to be *stratified* by centre.

Individual log-rank or Wilcoxon tests based on the data from each centre will be informative, but a test that combines information about the treatment difference in each centre would provide a more precise summary of the treatment effect. A similar situation would arise in attempting to test for treatment differences when patients are stratified according to variables such as age group, sex, performance status and other potential risk factors for the disease under study.

In situations such as those described above, a stratified version of the log-rank or Wilcoxon test may be employed. Essentially, this involves calculating the values of the  $U$ - and  $V$ -statistics for each *stratum*, and then combining these values over the strata. In this section, the *stratified log-rank test* will be described, but a stratified version of the Wilcoxon test can be obtained in a similar manner. An equivalent analysis, based on a model for the survival times, is described in Section 11.1.1 of Chapter 11.

Let  $U_{Lk}$  be the value of the log-rank statistic for comparing two treatment groups, computed from the  $k$ th of  $s$  strata using equation (2.23). Also, denote the variance of the statistic for the  $k$ th stratum by  $V_{Lk}$ , where  $V_{Lk}$  would be computed for each stratum using equation (2.24). The stratified log-rank test is then based on the statistic

$$W_S = \frac{\left( \sum_{k=1}^s U_{Lk} \right)^2}{\sum_{k=1}^s V_{Lk}}, \quad (2.29)$$

which has a chi-squared distribution on one degree of freedom (1 d.f.) under the null hypothesis that there is no treatment difference. Comparing the observed value of this statistic with percentage points of the chi-squared distribution enables the hypothesis of no overall treatment difference to be tested.

### Example 2.15 Survival times of melanoma patients

The aim of a study carried out by the University of Oklahoma Health Sciences Center was to compare two immunotherapy treatments for their ability to prolong the life of patients suffering from melanoma, a highly malignant tumour occurring in the skin. For each patient, the tumour was surgically

removed before allocation to *Bacillus Calmette-Guérin* (BCG) vaccine or to a vaccine based on the bacterium *Corynebacterium parvum* (*C. parvum*).

The survival times of the patients in each treatment group were further classified according to the age group of the patient. The data, which were given in Lee and Wang (2003), are shown in Table 2.9. An asterisk against a survival time indicates that the observation is censored.

**Table 2.9** Survival times of melanoma patients in two treatment groups, stratified by age group.

21–40		41–60		61–	
BCG	<i>C. parvum</i>	BCG	<i>C. parvum</i>	BCG	<i>C. parvum</i>
19	27*	34*	8	10	25*
24*	21*	4	11*	5	8
8	18*	17*	23*		11*
17*	16*		12*		
17*	7		15*		
34*	12*		8*		
	24		8*		
	8				
	8*				

These data are analysed by first computing the log-rank statistics for comparing the survival times of patients in the two treatment groups, separately for each age group. The resulting values of the  $U$ -,  $V$ - and  $W$ -statistics, found using equations (2.23), (2.25) and (2.26), are summarised in Table 2.10.

**Table 2.10** Values of the log-rank statistic for each age group.

Age group	$U_L$	$V_L$	$W_L$
21–40	−0.2571	1.1921	0.055
41–60	0.4778	0.3828	0.596
61–	1.0167	0.6497	1.591
Total	1.2374	2.2246	

The values of the  $W_L$ -statistic are quite similar for the three age groups, suggesting that the treatment effect is consistent over these groups. Moreover, none of them are significantly large at the 10% level.

To carry out a stratified log-rank test on these data, we calculate the  $W_S$ -statistic defined in equation (2.29). Using the results in Table 2.10,

$$W_S = \frac{1.2374^2}{2.2246} = 0.688.$$

The observed value of  $W_S$  is not significant when compared with percentage

points of the chi-squared distribution on 1 d.f. We therefore conclude that after allowing for the different age groups, there is no significant difference between the survival times of patients treated with the BCG vaccine and those treated with *C. parvum*.

For comparison, when the division of the patients into the different age groups is ignored, the log-rank test for comparing the two groups of patients leads to  $W_L = 0.756$ . The fact that this is so similar to the value that allows for age group differences suggests that it is not necessary to stratify the patients by age.

The stratified log-rank test can be extended to compare more than two treatment groups. The resulting formulae render it unsuitable for hand calculation, but the methodology can be implemented using computer software for survival analysis. However, this method of taking account of additional variables is not as flexible as that based on a modelling approach, introduced in the next chapter.

## 2.9 Log-rank test for trend

In many applications where three or more groups of survival data are to be compared, these groups are ordered in some way. For example, the groups may correspond to increasing doses of a treatment, the stage of a disease, or the age group of an individual. In comparing these groups using the log-rank test described in previous sections, it can happen that the analysis does not lead to a significant difference between the groups, even though the hazard of death increases or decreases across the groups. Indeed, a test that uses information about the ordering of the groups is more likely to lead to a trend being identified as significant than a standard log-rank test.

The log-rank test for trend across  $g$  ordered groups is based on the statistic

$$U_T = \sum_{k=1}^g w_k(d_{k\cdot} - e_{k\cdot}), \quad (2.30)$$

where  $w_k$  is a code assigned to the  $k$ th group,  $k = 1, 2, \dots, g$ , and

$$d_{k\cdot} = \sum_{j=1}^{r_k} d_{kj}, \quad e_{k\cdot} = \sum_{j=1}^{r_k} e_{kj},$$

are the observed and expected numbers of deaths in the  $k$ th group, where the summation is over the  $r_k$  death times in that group. Note that the dot subscript in the notation  $d_{k\cdot}$  and  $e_{k\cdot}$  stands for summation over the subscript that the dot replaces. The codes are often taken to be equally spaced to correspond to a linear trend across the groups. For example, if there are three groups, the codes might be taken to be 1, 2 and 3, although the equivalent choice of −1, 0 and 1 does simplify the calculations somewhat. The variance

of  $U_T$  is given by

$$V_T = \sum_{k=1}^g (w_k - \bar{w})^2 e_{k.}, \quad (2.31)$$

where  $\bar{w}$  is a weighted sum of the quantities  $w_k$ , in which the expected numbers of deaths,  $e_{k.}$ , are the weights, that is,

$$\bar{w} = \frac{\sum_{k=1}^g w_k e_{k.}}{\sum_{k=1}^g e_{k.}}$$

The statistic  $W_T = U_T^2/V_T$  then has a chi-squared distribution on 1 d.f. under the hypothesis of no trend across the  $g$  groups.

#### *Example 2.16 Survival times of melanoma patients*

The log-rank test for trend will be illustrated using the data from Example 2.15 on the survival times of patients suffering from melanoma. For the purpose of this illustration, only the data from those patients allocated to the BCG vaccine will be used. The log-rank statistic for comparing the survival times of the patients in the three age groups turns out to be 3.739. When compared to percentage points of the chi-squared distribution on 2 d.f., this is not significant ( $P = 0.154$ ).

We now use the log-rank test for trend to examine whether there is a linear trend over age. For this, we will take the codes,  $w_k$ , to be equally spaced, with values  $-1, 0$  and  $1$ . Some of the calculations required for the log-rank test for trend are summarised in Table 2.11.

**Table 2.11** Values of  $w_k$  and the observed and expected numbers of deaths in the three age groups.

Age group	$w_k$	$d_{k.}$	$e_{k.}$
21–40	-1	2	3.1871
41–60	0	1	1.1949
61+	1	2	0.6179

The log-rank test for trend is based on the statistic in equation (2.30), the value of which is

$$U_T = (d_{3.} - e_{3.}) - (d_{1.} - e_{1.}) = 2.5692.$$

Using the values of the expected numbers of deaths in each group, given in Table 2.11, the weighted mean of the  $w_k$ 's is given by

$$\bar{w} = \frac{e_{3.} - e_{1.}}{e_{1.} + e_{3.}} = 0.5138.$$

The three values of  $(w_k - \bar{w})^2$  are 0.2364, 0.2640 and 2.2917, and, from equa-

tion (2.31),  $V_T = 2.4849$ . Finally, the test statistic is

$$W_T = \frac{U_T^2}{V_T} = 2.656,$$

which is just about significant at the 10% level ( $P = 0.103$ ) when judged against a chi-squared distribution on 1 d.f. We therefore conclude that there is slight evidence of a linear trend across the age groups.

An alternative method of examining whether there is a trend across the levels of an ordered categorical variable, based on a modelling approach to the analysis of survival data, is described and illustrated in Section 3.6.2 of the next chapter.

## 2.10 Further reading

The life-table, which underpins the calculation of the life-table estimate of the survivor function, is widely used in the analysis of data from epidemiological studies. Fuller details of this application can be found in Armitage *et al.* (2001), and books on statistical methods in demography and epidemiology, such as Pollard *et al.* (1990) and Woodward (1999).

The product-limit estimate of the survivor function has been in use since the early 1900s. Kaplan and Meier (1958) derived the estimate using the method of maximum likelihood, which is why the estimate now bears their name. The properties of the Kaplan-Meier estimate of the survivor function have been further explored by Breslow and Crowley (1974) and Meier (1975). The Nelson-Aalen estimate is due to Altshuler (1970), Nelson (1972) and Aalen (1978), and the estimator is considered in a counting process framework by Therneau and Grambsch (2000).

The expression for the standard error of the Kaplan-Meier estimate was first given by Greenwood (1926), but an alternative expression is given by Aalen and Johansen (1978). Alternative expressions for the variance of the Nelson-Aalen estimate of the cumulative hazard function are compared by Klein (1991). Although Section 2.2.1 shows how a confidence interval for the value of the survivor function at particular times can be found using Greenwood's formula, alternative procedures are needed for the construction of confidence bands for the complete survivor function. Hall and Wellner (1980) and Efron (1981) have shown how such bands can be computed, and these procedures are also described by Harris and Albert (1991).

Methods for constructing confidence intervals for the median survival time are described by Brookmeyer and Crowley (1982), Emerson (1982), Nair (1984), Simon and Lee (1982) and Slud *et al.* (1984). Simon (1986) emphasises the importance of confidence intervals in reporting the results of clinical trials, and includes an illustration of a method described in Slud *et al.* (1984). Klein and Moeschberger (1997) include a comprehensive review of kernel-smoothed estimates of the hazard function.

The formulation of the hypothesis testing procedure in the frequentist ap-

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

$$h_N(t) = \psi h_S(t), \quad (3.1)$$

for any non-negative value of  $t$ , where  $\psi$  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  $\psi$  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  $\psi$  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  $i$ th of these by  $h_i(t)$ ,  $i = 1, 2, \dots, n$ . 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  $\psi$  cannot be negative, and so it is convenient to set  $\psi = \exp(\beta)$ . The parameter  $\beta$  is then the logarithm of the hazard ratio, that is,  $\beta = \log \psi$ , and any value of  $\beta$  in the range  $(-\infty, \infty)$  will lead to a positive value of  $\psi$ . Note that positive values of  $\beta$  are obtained when the hazard ratio,  $\psi$ , 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  $i$ th individual in the study,  $i = 1, 2, \dots, n$ , the hazard function for this individual can be written as

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

where  $x_i = 1$  if the  $i$ th 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  $x_1, x_2, \dots, x_p$  of  $p$  explanatory variables,  $X_1, X_2, \dots, X_p$ . 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  $\mathbf{x}$ , so that  $\mathbf{x} = (x_1, x_2, \dots, x_p)'$ . 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  $\mathbf{x}$  are zero. The function  $h_0(t)$  is called the *baseline hazard function*. The hazard function for the  $i$ th individual can then be written as

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

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

Again, since the relative hazard,  $\psi(\mathbf{x}_i)$ , cannot be negative, it is convenient to write this as  $\exp(\eta_i)$ , where  $\eta_i$  is a linear combination of the  $p$  explanatory variables in  $\mathbf{x}_i$ . 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,  $\eta_i = \boldsymbol{\beta}' \mathbf{x}_i$ , where  $\boldsymbol{\beta}$  is the vector of coefficients of the explanatory variables  $x_1, x_2, \dots, x_p$  in the model. The quantity  $\eta_i$  is called the *linear component* of the model, but it is also known as the *risk score* or *prognostic index* for the  $i$ th individual. The general proportional hazards model then becomes

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

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  $\psi(\mathbf{x}_i)$ , but the choice  $\psi(\mathbf{x}_i) = \exp(\boldsymbol{\beta}' \mathbf{x}_i)$  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  $\beta_0$ , 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  $\beta$ -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

$\beta$ -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  $i$ th individual will be  $x_{1i}$  and  $x_{2i}$ , respectively, and the proportional hazards model for the  $i$ th 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  $\alpha_j$  which represents the effect due to the  $j$ th level of the factor. The terms  $\alpha_1, \alpha_2, \dots, \alpha_a$  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  $\alpha_j$  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  $\alpha_j$  can be included in the model by defining  $a - 1$  indicator variables,  $X_2, X_3, \dots, X_a$ , 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  $\alpha_j$  can then be incorporated in the linear part of the proportional hazards model by including the  $a - 1$  explanatory variables  $X_2, X_3, \dots, X_a$  with coefficients  $\alpha_2, \alpha_3, \dots, \alpha_a$ . In other words, the term  $\alpha_j$  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  $j$ ,  $j = 2, 3, \dots, a$ . 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  $j = 1, 2, \dots, a$  and  $k = 1, 2, \dots, b$ .

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  $\alpha_j$  or  $\beta_k$  are excluded from the model, the term  $(\alpha\beta)_{jk}$  represents the effect of one factor *nested* within the other. For example, if  $\alpha_j$  is included in the model, but not  $\beta_k$ , then  $(\alpha\beta)_{jk}$  is the effect of  $B$  nested within  $A$ . If both  $\alpha_j$  and  $\beta_k$  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  $\alpha_j$  and  $\beta_k$ , 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 $A$	$U_2$	Level of $B$	$V_2$	$V_3$
1	0	1	0	0
2	1	2	1	0

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

$$(\alpha\beta)_{22} u_2 v_2 + (\alpha\beta)_{23} u_2 v_3.$$

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  $j$ th level,  $j = 1, 2, \dots, a$ . 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  $U_j X$  are  $u_j x$ , and the coefficient of  $u_j x$  in the model is  $\alpha_j$ , where  $j$  indexes the level of the factor  $A$ .

If the same definition of indicator variables in the previous discussion were used,  $\alpha_1$ , 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  $\alpha'_2$  and  $\alpha'_3$ , respectively, and let the coefficient of the value of the variate  $X$  in the model be  $\beta$ . Then, the model contains the terms  $\beta x + \alpha'_2(u_2x) + \alpha'_3(u_3x)$ . 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  $\beta$ . 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  $\beta x$  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  $\beta x$ ,  $a$  indicator variables  $Z_1, Z_2, \dots, Z_a$  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,  $z_1x, z_2x, \dots, z_ax$ , are included in the model with coefficients  $\alpha_1, \alpha_2, \dots, \alpha_a$ . 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  $Z_j X$ , 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  $\alpha$ -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,  $X_1, X_2, \dots, X_p$ , in the linear component of the model,  $\beta_1, \beta_2, \dots, \beta_p$ . 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  $\beta$ '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,  $X_1, X_2, \dots, X_p$ , 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  $\beta$ -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  $\beta$ -parameters in the linear component of the model. Estimates of the  $\beta$ '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  $t_{(1)} < t_{(2)} < \dots < t_{(r)}$ , so that  $t_{(j)}$  is the  $j$ th 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

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

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

death times have been recorded. Individuals for whom the survival times are censored do not contribute to the numerator of the log-likelihood function, but they do enter into the summation over the risk sets at death times that occur before a censored time. Moreover, the likelihood function depends only on the ranking of the death times, since this determines the risk set at each death time. Consequently, inferences about the effect of explanatory variables on the hazard function depend only on the rank order of the survival times.

Now suppose that the data consist of  $n$  observed survival times, denoted by  $t_1, t_2, \dots, t_n$ , and that  $\delta_i$  is an event indicator, which is zero if the  $i$ th survival time  $t_i$ ,  $i = 1, 2, \dots, n$ , is right-censored, and unity otherwise. The likelihood function in equation (3.4) can then be expressed in the form

$$\prod_{i=1}^n \left\{ \frac{\exp(\beta' x_i)}{\sum_{l \in R(t_i)} \exp(\beta' x_l)} \right\}^{\delta_i},$$

where  $R(t_i)$  is the risk set at time  $t_i$ . The corresponding log-likelihood function is given by

$$\log L(\beta) = \sum_{i=1}^n \delta_i \left\{ \beta' x_i - \log \sum_{l \in R(t_i)} \exp(\beta' x_l) \right\}. \quad (3.5)$$

The maximum likelihood estimates of the  $\beta$ -parameters in the proportional hazards model can be found by maximising this log-likelihood function using numerical methods. This maximisation is generally accomplished using the *Newton-Raphson procedure* described below in Section 3.3.3.

Fortunately, most of the major statistical packages have facilities which enable the proportional hazards model to be fitted. Such software also gives the standard errors of the parameter estimates in the fitted model. The calculations in this book have been carried out using the package SAS, with the SAS procedure `proc phreg` being used to fit the proportional hazards model of equation (3.3).

The justification for using equation (3.4) as a likelihood function, and further details on the structure of the likelihood function, are given in Section 3.3.1. The treatment of tied survival times is then discussed in Section 3.3.2 and the Newton-Raphson procedure is outlined in Section 3.3.3. These three sections can be omitted without loss of continuity.

### 3.3.1\* Likelihood function for the model

The basis of the argument used in the construction of a likelihood function for the proportional hazards model is that intervals between successive death times convey no information about the effect of explanatory variables on the hazard of death. This is because the baseline hazard function has an arbitrary form, and so it is conceivable that  $h_0(t)$ , and hence  $h(t)$ , is zero in those time intervals in which there are no deaths. This in turn means that these intervals give no information about the values of the  $\beta$ -parameters. We therefore con-

sider the probability that the  $i$ th individual dies at some time  $t_{(j)}$ , conditional on  $t_{(j)}$  being one of the observed set of  $r$  death times  $t_{(1)}, t_{(2)}, \dots, t_{(r)}$ . If the vector of explanatory variables for the individual who dies at  $t_{(j)}$  is denoted by  $x_{(j)}$ , this probability is

$$P(\text{individual with variables } x_{(j)} \text{ dies at } t_{(j)} \mid \text{one death at } t_{(j)}). \quad (3.6)$$

Next, from the result that the probability of an event  $A$ , given that an event  $B$  has occurred, is given by

$$P(A|B) = P(A \text{ and } B)/P(B),$$

the probability in expression (3.6) becomes

$$\frac{P(\text{individual with variables } x_{(j)} \text{ dies at } t_{(j)})}{P(\text{one death at } t_{(j)})}. \quad (3.7)$$

Since the death times are assumed to be independent of one another, the denominator of this expression is the sum of the probabilities of death at time  $t_{(j)}$  over all individuals who are at risk of death at that time. If these individuals are indexed by  $l$ , with  $R(t_{(j)})$  denoting the set of individuals who are at risk at time  $t_{(j)}$ , expression (3.7) becomes

$$\frac{P(\text{individual with variables } x_{(j)} \text{ dies at } t_{(j)})}{\sum_{l \in R(t_{(j)})} P(\text{individual } l \text{ dies at } t_{(j)})}. \quad (3.8)$$

The probabilities of death at time  $t_{(j)}$ , in expression (3.8), are now replaced by probabilities of death in the interval  $(t_{(j)}, t_{(j)} + \delta t)$ , and dividing both the numerator and denominator of expression (3.8) by  $\delta t$ , we get

$$\frac{P\{\text{individual with variables } x_{(j)} \text{ dies in } (t_{(j)}, t_{(j)} + \delta t)\}/\delta t}{\sum_{l \in R(t_{(j)})} P\{\text{individual } l \text{ dies in } (t_{(j)}, t_{(j)} + \delta t)\}/\delta t}.$$

The limiting value of this expression as  $\delta t \rightarrow 0$  is then the ratio of the probabilities in expression (3.8). But from equation (1.2), this limit is also the ratio of the corresponding hazards of death at time  $t_{(j)}$ , that is,

$$\frac{\text{Hazard of death at time } t_{(j)} \text{ for individual with variables } x_{(j)}}{\sum_{l \in R(t_{(j)})} \{\text{Hazard of death at time } t_{(j)} \text{ for individual } l\}}.$$

If it is the  $i$ th individual who dies at  $t_{(j)}$ , the hazard function in the numerator of this expression can be written  $h_i(t_{(j)})$ . Similarly, the denominator is the sum of the hazards of death at time  $t_{(j)}$  over all individuals who are at risk of death at this time. This is the sum of the values  $h_l(t_{(j)})$  over those individuals in the risk set at time  $t_{(j)}$ ,  $R(t_{(j)})$ . Consequently, the conditional probability in expression (3.6) becomes

$$\frac{h_i(t_{(j)})}{\sum_{l \in R(t_{(j)})} h_l(t_{(j)})}.$$

On using equation (3.3), the baseline hazard function in the numerator and

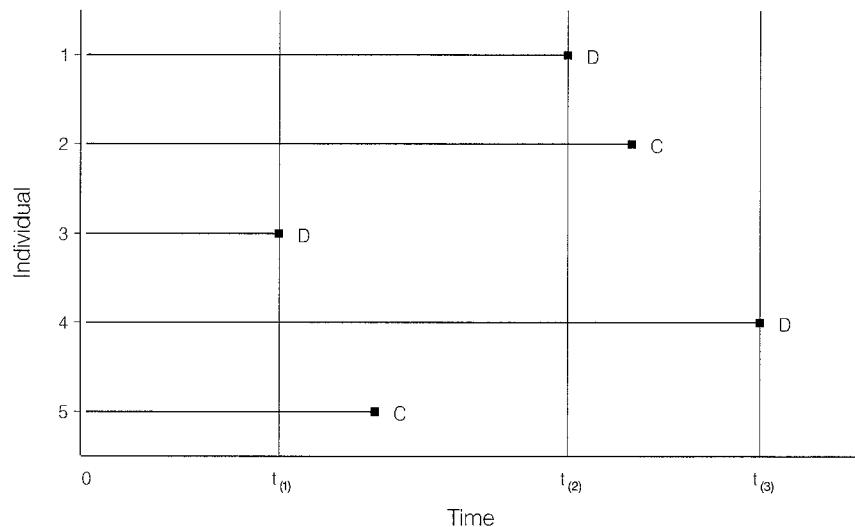
denominator cancels out, and we are left with

$$\frac{\exp(\beta' \mathbf{x}_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\beta' \mathbf{x}_l)}.$$

Finally, taking the product of these conditional probabilities over the  $r$  death times gives the likelihood function in equation (3.4).

The likelihood function that has been obtained is not a true likelihood, since it does not make direct use of the actual censored and uncensored survival times. For this reason it is referred to as a *partial likelihood function*.

In order to throw more light on the structure of the partial likelihood, consider a sample of survival data from five individuals, numbered from 1 to 5. The survival data are illustrated in Figure 3.1.



**Figure 3.1** Survival times of five individuals.

The observed survival times of individuals 2 and 5 will be taken to be right-censored, and the three ordered death times are denoted  $t_{(1)} < t_{(2)} < t_{(3)}$ . Then,  $t_{(1)}$  is the death time of individual 3,  $t_{(2)}$  is that of individual 1, and  $t_{(3)}$  that of individual 4.

The risk set at each of the three ordered death times consists of the individuals who are alive and uncensored just prior to each death time. Hence, the risk set  $R(t_{(1)})$  consists of all five individuals, risk set  $R(t_{(2)})$  consists of individuals 1, 2 and 4, while risk set  $R(t_{(3)})$  only includes individual 4. Now write  $\psi(i) = \exp(\beta' \mathbf{x}_i)$ ,  $i = 1, 2, \dots, 5$ , for the risk score for the  $i$ th individual, where  $\mathbf{x}_i$  is the vector of explanatory variables for that individual. The numerators of the partial likelihood function for times  $t_{(1)}$ ,  $t_{(2)}$  and  $t_{(3)}$ , respectively, are  $\psi(3)$ ,  $\psi(1)$  and  $\psi(4)$ , since individuals 3, 1 and 4, respectively, die at the three ordered death times. The partial likelihood function over the

three death times is then

$$\frac{\psi(3)}{\psi(1) + \psi(2) + \psi(3) + \psi(4) + \psi(5)} \times \frac{\psi(1)}{\psi(1) + \psi(2) + \psi(4)} \times \frac{\psi(4)}{\psi(4)}.$$

It turns out that standard results used in maximum likelihood estimation carry over without modification to maximum partial likelihood estimation. In particular, the results given in Appendix A for the variance-covariance matrix of the estimates of the  $\beta$ 's can be used, as can distributional results associated with likelihood ratio testing, to be discussed in Section 3.4.

### 3.3.2\* Treatment of ties

The proportional hazards model for survival data assumes that the hazard function is continuous, and under this assumption, tied survival times are not possible. Of course, survival times are usually recorded to the nearest day, month or year, and so tied survival times can arise as a result of this rounding process. Indeed, Examples 1.2, 1.3 and 1.4 in Chapter 1 all contain tied observations.

In addition to the possibility of more than one death at a given time, there might also be one or more censored observations at a death time. When there are both censored survival times and deaths at a given time, the censoring is assumed to occur after all the deaths. Potential ambiguity concerning which individuals should be included in the risk set at that death time is then resolved and tied censored observations present no further difficulties in the computation of the likelihood function using equation (3.4). Accordingly, we only need consider how tied survival times can be handled in fitting the proportional hazards model.

In order to accommodate tied observations, the likelihood function in equation (3.4) has to be modified in some way. The appropriate likelihood function in the presence of tied observations has been given by Kalbfleisch and Prentice (2002). However, this likelihood has a very complicated form, and will not be reproduced here. In addition, the computation of this likelihood function can be very time consuming, particularly when there are a relatively large number of ties at one or more death times. Fortunately, there are a number of approximations to the likelihood function that have computational advantages over the exact method. But before these are given, some additional notation needs to be developed.

Let  $\mathbf{s}_j$  be the vector of sums of each of the  $p$  covariates for those individuals who die at the  $j$ th death time,  $t_{(j)}$ ,  $j = 1, 2, \dots, r$ . If there are  $d_j$  deaths at  $t_{(j)}$ , the  $h$ th element of  $\mathbf{s}_j$  is  $s_{hj} = \sum_{k=1}^{d_j} x_{hjk}$ , where  $x_{hjk}$  is the value of the  $h$ th explanatory variable,  $h = 1, 2, \dots, p$ , for the  $k$ th of  $d_j$  individuals,  $k = 1, 2, \dots, d_j$ , who die at the  $j$ th death time,  $j = 1, 2, \dots, r$ .

The simplest approximation to the likelihood function is that due to Breslow

(1974), who proposed the approximate likelihood

$$\prod_{j=1}^r \frac{\exp(\beta' s_j)}{\left\{ \sum_{l \in R(t_{(j)})} \exp(\beta' x_l) \right\}^{d_j}}. \quad (3.9)$$

In this approximation, the  $d_j$  deaths at time  $t_{(j)}$  are considered to be distinct and to occur sequentially. The probabilities of all possible sequences of deaths are then summed to give the likelihood in equation (3.9). Apart from a constant of proportionality, this is also the approximation suggested by Peto (1972). This likelihood is quite straightforward to compute, and is an adequate approximation when the number of tied observations at any one death time is not too large. For these reasons, this method is usually the default procedure for handling ties in statistical software for survival analysis, and will be used in the examples given in this book.

Efron (1977) proposed

$$\prod_{j=1}^r \frac{\exp(\beta' s_j)}{\prod_{k=1}^{d_j} \left[ \sum_{l \in R(t_{(j)})} \exp(\beta' x_l) - (k-1)d_j^{-1} \sum_{l \in D(t_{(j)})} \exp(\beta' x_l) \right]} \quad (3.10)$$

as an approximate likelihood function for the proportional hazards model, where  $D(t_{(j)})$  is the set of all individuals who die at time  $t_{(j)}$ . This is a closer approximation to the appropriate likelihood function than that due to Breslow, although in practice, both approximations often give similar results.

Cox (1972) suggested the approximation

$$\prod_{j=1}^r \frac{\exp(\beta' s_j)}{\sum_{l \in R(t_{(j)}; d_j)} \exp(\beta' s_l)}, \quad (3.11)$$

where the notation  $R(t_{(j)}; d_j)$  denotes a set of  $d_j$  individuals drawn from  $R(t_{(j)})$ , the risk set at  $t_{(j)}$ . The summation in the denominator is the sum over all possible sets of  $d_j$  individuals sampled from the risk set without replacement. The approximation in expression (3.11) is based on a model for the situation where the time-scale is discrete, so that under this model, tied observations are permissible. Now, the hazard function for an individual with vector of explanatory variables  $x_i$ ,  $h_i(t)$ , is the probability of death in the unit time interval  $(t, t+1)$ , conditional on survival to time  $t$ . A discrete version of the proportional hazards model of equation (3.3) is the model

$$\frac{h_i(t)}{1 - h_i(t)} = \exp(\beta' x_i) \frac{h_0(t)}{1 - h_0(t)},$$

for which the likelihood function is that given in equation (3.11). In fact, in the limit as the width of the discrete time intervals becomes zero, this model tends to the proportional hazards model of equation (3.3).

When there are no ties, that is, when  $d_j = 1$  for each death time, the approximations in equations (3.9), (3.10), and (3.11) all reduce to the likelihood function in equation (3.4).

### 3.3.3\* The Newton-Raphson procedure

Models for censored survival data are usually fitted by using the Newton-Raphson procedure to maximise the partial likelihood function, and so the procedure is outlined in this section.

Let  $\mathbf{u}(\boldsymbol{\beta})$  be the  $p \times 1$  vector of first derivatives of the log-likelihood function in equation (3.5) with respect to the  $\beta$ -parameters. This quantity is known as the *vector of efficient scores*. Also, let  $\mathbf{I}(\boldsymbol{\beta})$  be the  $p \times p$  matrix of negative second derivatives of the log-likelihood, so that the  $(j, k)$ th element of  $\mathbf{I}(\boldsymbol{\beta})$  is

$$-\frac{\partial^2 \log L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k}.$$

The matrix  $\mathbf{I}(\boldsymbol{\beta})$  is known as the *observed information matrix*.

According to the Newton-Raphson procedure, an estimate of the vector of  $\beta$ -parameters at the  $(s+1)$ th cycle of the iterative procedure,  $\hat{\boldsymbol{\beta}}_{s+1}$ , is

$$\hat{\boldsymbol{\beta}}_{s+1} = \hat{\boldsymbol{\beta}}_s + \mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}_s) \mathbf{u}(\hat{\boldsymbol{\beta}}_s),$$

for  $s = 0, 1, 2, \dots$ , where  $\mathbf{u}(\hat{\boldsymbol{\beta}}_s)$  is the vector of efficient scores and  $\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}_s)$  is the inverse of the information matrix, both evaluated at  $\hat{\boldsymbol{\beta}}_s$ . The procedure can be started by taking  $\hat{\boldsymbol{\beta}}_0 = \mathbf{0}$ . The process is terminated when the change in the log-likelihood function is sufficiently small, or when the largest of the relative changes in the values of the parameter estimates is sufficiently small.

When the iterative procedure has converged, the variance-covariance matrix of the parameter estimates can be approximated by the inverse of the information matrix, evaluated at  $\hat{\boldsymbol{\beta}}$ , that is,  $\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})$ . The square root of the diagonal elements of this matrix are then the standard errors of the estimated values of  $\beta_1, \beta_2, \dots, \beta_p$ .

### 3.4 Confidence intervals and hypothesis tests for the $\beta$ 's

When a statistical package is used to fit a proportional hazards model, the parameter estimates that are provided are usually accompanied by their standard errors. These standard errors can be used to obtain approximate confidence intervals for the unknown  $\beta$ -parameters. In particular, a  $100(1 - \alpha)\%$  confidence interval for a parameter  $\beta$  is the interval with limits  $\hat{\beta} \pm z_{\alpha/2} \text{se}(\hat{\beta})$ , where  $\hat{\beta}$  is the estimate of  $\beta$ , and  $z_{\alpha/2}$  is the upper  $\alpha/2$ -point of the standard normal distribution.

If a  $100(1 - \alpha)\%$  confidence interval for  $\beta$  does not include zero, this is evidence that the value of  $\beta$  is non-zero. More specifically, the null hypothesis that  $\beta = 0$  can be tested by calculating the value of the statistic  $\hat{\beta}/\text{se}(\hat{\beta})$ . The observed value of this statistic is then compared to percentage points of the standard normal distribution in order to obtain the corresponding  $P$ -value. Equivalently, the square of this statistic can be compared with percentage points of a chi-squared distribution on one degree of freedom. This procedure is sometimes called a *Wald test*. Indeed, the  $P$ -values for this test are often

given alongside parameter estimates and their standard errors in computer output.

When attempting to interpret the  $P$ -value for a given parameter,  $\beta_j$ , say, it is important to recognise that the hypothesis that is being tested is that  $\beta_j = 0$  in the presence of all other terms that are in the model. For example, suppose that a model contains the three explanatory variables  $X_1, X_2, X_3$ , and that their coefficients are  $\beta_1, \beta_2, \beta_3$ . The test statistic  $\hat{\beta}_j / \text{se}(\hat{\beta}_j)$  is then used to test the null hypothesis that  $\beta_j = 0$  in the presence of  $\beta_1$  and  $\beta_3$ . If there was no evidence to reject this hypothesis, we would conclude that  $X_2$  was not needed in the model in the presence of  $X_1$  and  $X_3$ .

In general, the individual estimates of the  $\beta$ 's in a proportional hazards model are not all independent of one another. This means that the results of testing separate hypotheses about the  $\beta$ -parameters in a model may not be easy to interpret. For example, consider again the situation where there are three explanatory variables,  $X_1, X_2, X_3$ . If  $\hat{\beta}_1$  and  $\hat{\beta}_2$  were not found to be significantly different from zero, when compared with their standard errors, we could not conclude that only  $X_3$  need be included in the model. This is because the coefficient of  $X_1$ , for example, could well change when  $X_2$  is excluded from the model, and vice versa. This would certainly happen if  $X_1$  and  $X_2$  were correlated.

Because of the difficulty in interpreting the results of tests concerning the coefficients of the explanatory variables in a model, alternative methods for comparing different proportional hazards models are required. It turns out that the methods to be described in Section 3.5 are much more satisfactory than the Wald tests. Little attention should therefore be paid to the results of these tests given in computer-based analyses of survival data.

### 3.4.1 Standard errors and confidence intervals for hazard ratios

We have seen that in situations where there are two groups of survival data, the parameter  $\beta$  is the logarithm of the ratio of the hazard of death at time  $t$  for individuals in one group relative to those in the other. Hence the hazard ratio itself is  $\psi = e^\beta$ . The corresponding estimate of the hazard ratio is  $\hat{\psi} = \exp(\hat{\beta})$ , and the standard error of  $\hat{\psi}$  can be obtained from the standard error of  $\hat{\beta}$  using the result given as equation (2.9) in Chapter 2. From this result, the approximate variance of  $\hat{\psi}$ , a function of  $\hat{\beta}$ , is

$$\left\{ \exp(\hat{\beta}) \right\}^2 \text{var}(\hat{\beta}),$$

that is,  $\hat{\psi}^2 \text{var}(\hat{\beta})$ , and so the standard error of  $\hat{\psi}$  is given by

$$\text{se}(\hat{\psi}) = \hat{\psi} \text{se}(\hat{\beta}). \quad (3.12)$$

Generally speaking, a confidence interval for the true hazard ratio will be more informative than the standard error of the estimated hazard ratio. A  $100(1 - \alpha)\%$  confidence interval for the true hazard ratio,  $\psi$ , can be found simply by exponentiating the confidence limits for  $\beta$ . An interval estimate

obtained in this way is preferable to one found using  $\hat{\psi} \pm z_{\alpha/2} \text{se}(\hat{\psi})$ . This is because the distribution of the logarithm of the estimated hazard ratio will be more closely approximated by a normal distribution than that of the hazard ratio itself.

The construction of a confidence interval for a hazard ratio is illustrated in Example 3.1 below. Fuller details on the interpretation of the parameters in the linear component of a proportional hazards model are given in Section 3.7.

### 3.4.2 Two examples

In this section, the results of fitting a proportional hazards model to data from two of the examples introduced in Chapter 1 are given.

#### *Example 3.1 Prognosis for women with breast cancer*

Data on the survival times of breast cancer patients, classified according to whether or not sections of their tumours were positively stained, were first given in Example 1.2. The variable that indexes the result of the staining process can be regarded as a factor with two levels. From the arguments given in Section 3.2.1, this factor can be fitted by using an indicator variable  $X$  to denote the staining result, where  $X = 0$  corresponds to negative staining and  $X = 1$  to positive staining. Under the proportional hazards model, the hazard of death at time  $t$  for the  $i$ th woman, for whom the value of the indicator variable is  $x_i$ , is

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

where  $x_i$  is zero or unity. The baseline hazard function  $h_0(t)$  is then the hazard function for a women with a negatively stained tumour. This is essentially the model considered in Section 3.1.1, and given in equation (3.2).

In the group of women whose tumours were positively stained, there are two who die at 26 months. To cope with this tie, the Breslow approximation to the likelihood function will be used. This model is fitted by finding that value of  $\beta$ ,  $\hat{\beta}$ , which maximises the likelihood function in equation (3.9). The maximum likelihood estimate of  $\beta$  is  $\hat{\beta} = 0.908$ . The standard error of this estimate is also obtained from statistical packages for fitting the Cox regression model, and turns out to be given by  $\text{se}(\hat{\beta}) = 0.501$ .

The quantity  $e^\beta$  is the ratio of the hazard function for a woman with  $X = 1$  to that for a woman with  $X = 0$ , so that  $\beta$  is the logarithm of the ratio of the hazard of death at time  $t$  for positively stained relative to negatively stained women. The estimated value of this hazard ratio is  $e^{0.908} = 2.48$ . Since this is greater than unity, we conclude that a woman who has a positively stained tumour will have a greater risk of death at any given time than a comparable women whose tumour was negatively stained. Positive staining therefore indicates a poorer prognosis for a breast cancer patient.

The standard error of the hazard ratio can be found from the standard error of  $\hat{\beta}$ , using the result in equation (3.12). Since the estimated relative hazard is  $\hat{\psi} = \exp(\hat{\beta}) = 2.480$ , and the standard error of  $\hat{\beta}$  is 0.501, the standard error

of  $\hat{\psi}$  is given by

$$\text{se}(\hat{\psi}) = 2.480 \times 0.501 = 1.242.$$

We can go further and construct a confidence interval for this hazard ratio. The first step is to obtain a confidence interval for the logarithm of the hazard ratio,  $\beta$ . For example, a 95% confidence interval for  $\beta$  is the interval from  $\hat{\beta} - 1.96 \text{se}(\hat{\beta})$  to  $\hat{\beta} + 1.96 \text{se}(\hat{\beta})$ , that is, the interval from  $-0.074$  to  $1.890$ . Exponentiating these confidence limits gives  $(0.93, 6.62)$  as a 95% confidence interval for the hazard ratio itself. Notice that this interval barely includes unity, suggesting that there is evidence that the two groups of women have a different survival experience.

### *Example 3.2 Survival of multiple myeloma patients*

Data on the survival times of 48 patients suffering from multiple myeloma were given in Example 1.3. The data base also contains the values of seven other variables that were recorded for each patient. For convenience, the values of the variable that describes the sex of a patient have been redefined to be zero and unity for males and females respectively. The variables are then as follows:

- Age*: Age of the patient,
- Sex*: Sex of the patient ( $0 = \text{male}$ ,  $1 = \text{female}$ ),
- Bun*: Blood urea nitrogen,
- Ca*: Serum calcium,
- Hb*: Serum haemoglobin,
- Pcells*: Percentage of plasma cells,
- Protein*: Bence-Jones protein ( $0 = \text{absent}$ ,  $1 = \text{present}$ ).

The sex of the patient and the variable associated with the occurrence of Bence-Jones protein are factors with two levels. These terms are fitted using the indicator variables *Sex* and *Protein*. The proportional hazards model for the  $i$ th individual is then

$$h_i(t) = \exp(\beta_1 \text{Age}_i + \beta_2 \text{Sex}_i + \beta_3 \text{Bun}_i + \beta_4 \text{Ca}_i + \beta_5 \text{Hb}_i + \beta_6 \text{Pcells}_i + \beta_7 \text{Protein}_i) h_0(t),$$

where the subscript  $i$  on an explanatory variable denotes the value of that variable for the  $i$ th individual. The baseline hazard function is the hazard function for an individual for whom the values of all seven of these variables are zero. This function therefore corresponds to a male aged zero, who has zero values of *Bun*, *Ca*, *Hb* and *Pcells*, and no Bence-Jones protein. In view of the obvious difficulty in interpreting this function, it might be more sensible to redefine the variables *Age*, *Bun*, *Ca*, *Hb* and *Pcells* by subtracting values for an average patient. For example, if we took *Age* – 60 in place of *Age*, the baseline hazard would correspond to a male aged 60 years. This procedure also avoids the introduction of a function that describes the hazard of individuals whose ages are rather different from the age range of patients in the study. Although this leads to a baseline hazard function that has a more natural interpretation,

it will not affect inference about the influence of the explanatory variables on the hazard of death. For this reason, the untransformed variables will be used in this example. On fitting the model, the estimates of the coefficients of the explanatory variables and their standard errors are found to be those shown in Table 3.1.

**Table 3.1** Estimated values of the coefficients of the explanatory variables on fitting a proportional hazards model to the data from Example 1.3.

Variable	$\hat{\beta}$	$\text{se}(\hat{\beta})$
<i>Age</i>	-0.019	0.028
<i>Sex</i>	-0.251	0.402
<i>Bun</i>	0.021	0.006
<i>Ca</i>	0.013	0.132
<i>Hb</i>	-0.135	0.069
<i>Pcells</i>	-0.002	0.007
<i>Protein</i>	-0.640	0.427

We see from Table 3.1 that some of the estimates are close to zero. Indeed, if individual 95% confidence intervals are calculated for the coefficients of the seven variables, only those for *Bun* and *Hb* exclude zero. This suggests that the hazard function does not depend on all seven explanatory variables.

However, we cannot deduce from this that *Bun* and *Hb* are the relevant variables, since the estimates of the coefficients of the seven explanatory variables in the fitted model are not independent of one another. This means that if one of the seven explanatory variables were excluded from the model, the coefficients of the remaining six might be different from those in Table 3.1. For example, if *Bun* is omitted, the estimated coefficients of the six remaining explanatory variables, *Age*, *Sex*, *Ca*, *Hb*, *Pcells* and *Protein*, turn out to be -0.009, -0.301, -0.036, -0.140, -0.001, and -0.420, respectively. Comparison with the values shown in Table 3.1 shows that there are differences in the estimated coefficients of each of these six variables, although in this case the differences are not very great.

In general, to determine on which of the seven explanatory variables the hazard function depends, a number of different models will need to be fitted, and the results compared. Methods for comparing the fit of alternative models, and strategies for model building are considered in subsequent sections of this chapter.

### 3.5 Comparing alternative models

In a modelling approach to the analysis of survival data, a model is developed for the dependence of the hazard function on one or more explanatory vari-

ables. In this development process, proportional hazards models with linear components that contain different sets of terms are fitted, and comparisons made between them.

As a specific example, consider the situation where there are two groups of survival times, corresponding to individuals who receive either a new treatment or a standard. The common hazard function under the model for no treatment difference can be taken to be  $h_0(t)$ . This model is a special case of the general proportional hazards model in equation (3.3), in which there are no explanatory variables in the linear component of the model. This model is therefore referred to as the *null model*.

Now let  $X$  be an indicator variable that takes the value zero for individuals receiving the standard treatment and unity otherwise. Under a proportional hazards model, the hazard function for an individual for whom  $X$  takes the value  $x$  is  $e^{\beta x}h_0(t)$ . The hazard functions for individuals on the standard and new treatments are then  $h_0(t)$  and  $e^{\beta}h_0(t)$ , respectively. The difference between this model and the null model is that the linear component of the latter contains the additional term  $\beta x$ . Since  $\beta = 0$  corresponds to no treatment effect, the extent of any treatment difference can be investigated by comparing these two proportional hazards models for the observed survival data.

More generally, suppose that two models are contemplated for a particular data set, Model (1) and Model (2), say, where Model (1) contains a subset of the terms in Model (2). Model (1) is then said to be *parametrically nested* within Model (2). Specifically, suppose that the  $p$  explanatory variables,  $X_1, X_2, \dots, X_p$ , are fitted in Model (1), so that the hazard function under this model can be written as

$$\exp\{\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p\}h_0(t).$$

Also suppose that the  $p+q$  explanatory variables  $X_1, X_2, \dots, X_p, X_{p+1}, \dots, X_{p+q}$  are fitted in Model (2), so that the hazard function under this model is

$$\exp\{\beta_1 x_1 + \dots + \beta_p x_p + \beta_{p+1} x_{p+1} + \dots + \beta_{p+q} x_{p+q}\}h_0(t).$$

Model (2) then contains the  $q$  additional explanatory variables  $X_{p+1}, X_{p+2}, \dots, X_{p+q}$ . Because Model (2) has a larger number of terms than Model (1), Model (2) must be a better fit to the observed data. The statistical problem is then to determine whether the additional  $q$  terms in Model (2) significantly improve the explanatory power of the model. If not, they might be omitted, and Model (1) would be deemed to be adequate.

In the discussion of Example 3.2, we saw that when there are a number of explanatory variables of possible relevance, the effect of each term cannot be studied independently of the others. The effect of any given term therefore depends on the other terms currently included in the model. For example, in Model (1), the effect of any of the  $p$  explanatory variables on the hazard function depends on the  $p-1$  variables that have already been fitted, and so the effect of  $X_p$  is said to be *adjusted* for the remaining  $p-1$  variables. In particular, the effect of  $X_p$  is adjusted for  $X_1, X_2, \dots, X_{p-1}$ , but we also speak of the effect of  $X_p$  *eliminating* or *allowing for*  $X_1, X_2, \dots, X_{p-1}$ . Similarly,

when the  $q$  variables  $X_{p+1}, X_{p+2}, \dots, X_{p+q}$  are added to Model (1), the effect of these variables on the hazard function is said to be adjusted for the  $p$  variables that have already been fitted,  $X_1, X_2, \dots, X_p$ .

### 3.5.1 The statistic $-2 \log \hat{L}$

In order to compare alternative models fitted to an observed set of survival data, a statistic that measures the extent to which the data are fitted by a particular model is required. Since the likelihood function summarises the information that the data contain about the unknown parameters in a given model, a suitable summary statistic is the value of the likelihood function when the parameters are replaced by their maximum likelihood estimates. This is the maximised likelihood under an assumed model, and can be computed from equation (3.4) by replacing the  $\beta$ 's by their maximum likelihood estimates under the model. For a given set of data, the larger the value of the maximised likelihood, the better is the agreement between the model and the observed data.

For reasons given in the sequel, it is more convenient to use minus twice the logarithm of the maximised likelihood in comparing alternative models. If the maximised likelihood for a given model is denoted by  $\hat{L}$ , the summary measure of agreement between the model and the data is  $-2 \log \hat{L}$ . From Section 3.3.1,  $\hat{L}$  is in fact the product of a series of conditional probabilities, and so this statistic will be less than unity. In consequence,  $-2 \log \hat{L}$  will always be positive, and for a given data set, the smaller the value of  $-2 \log \hat{L}$ , the better the model.

The statistic  $-2 \log \hat{L}$  cannot be used on its own as a measure of model adequacy. The reason for this is that the value of  $\hat{L}$ , and hence of  $-2 \log \hat{L}$ , is dependent upon the number of observations in the data set. Thus if, after fitting a model to a set of data, additional data became available to which the fit of the model was the same as that to the original data, the value of  $-2 \log \hat{L}$  for the enlarged data set would be different from that of the original data. Accordingly the value of  $-2 \log \hat{L}$  is only useful when making comparisons between models fitted to the same data.

### 3.5.2 Comparing nested models

Consider again Model (1) and Model (2) defined above, and let the value of the maximised log-likelihood function for each model be denoted by  $\hat{L}(1)$  and  $\hat{L}(2)$ , respectively. The two models can then be compared on the basis of the difference between the values of  $-2 \log \hat{L}$  for each model. In particular, a large difference between  $-2 \log \hat{L}(1)$  and  $-2 \log \hat{L}(2)$  would lead to the conclusion that the  $q$  variates in Model (2), that are additional to those in Model (1), do improve the adequacy of the model. Naturally, the amount by which the value of  $-2 \log \hat{L}$  changes when terms are added to a model will depend on which terms have already been included. In particular, the difference in the values of  $-2 \log \hat{L}(1)$  and  $-2 \log \hat{L}(2)$ , that is,  $-2 \log \hat{L}(1) + 2 \log \hat{L}(2)$ , will reflect the

combined effect of adding the variables  $X_{p+1}, X_{p+2}, \dots, X_{p+q}$  to a model that already contains  $X_1, X_2, \dots, X_p$ . This is said to be the change in the value of  $-2 \log \hat{L}$  due to fitting  $X_{p+1}, X_{p+2}, \dots, X_{p+q}$ , adjusted for  $X_1, X_2, \dots, X_p$ .

The statistic  $-2 \log \hat{L}(1) + 2 \log \hat{L}(2)$ , can be written as

$$-2 \log \{\hat{L}(1)/\hat{L}(2)\},$$

and this is the log-likelihood ratio statistic for testing the null hypothesis that the  $q$  parameters  $\beta_{p+1}, \beta_{p+2}, \dots, \beta_{p+q}$  in Model (2) are all zero. From results associated with the theory of likelihood ratio testing (see Appendix A), this statistic has an asymptotic chi-squared distribution, under the null hypothesis that the coefficients of the additional variables are zero. The number of degrees of freedom of this chi-squared distribution is equal to the difference between the number of independent  $\beta$ -parameters being fitted under the two models. Hence, in order to compare the value of  $-2 \log \hat{L}$  for Model (1) and Model (2), we use the fact that the statistic  $-2 \log \hat{L}(1) + 2 \log \hat{L}(2)$  has a chi-squared distribution on  $q$  degrees of freedom, under the null hypothesis that  $\beta_{p+1}, \beta_{p+2}, \dots, \beta_{p+q}$  are all zero. If the observed value of the statistic is not significantly large, the two models will be adjudged to be equally suitable. Then, other things being equal, the more simple model, that is, the one with fewer terms, would be preferred. On the other hand, if the values of  $-2 \log \hat{L}$  for the two models are significantly different, we would argue that the additional terms are needed and the more complex model would be adopted.

Some texts, and some software packages, ascribe degrees of freedom to the quantity  $-2 \log \hat{L}$ . However, the value of  $-2 \log \hat{L}$  for a particular model does not have a chi-squared distribution, and so the quantity cannot be considered to have an associated number of degrees of freedom. Additionally, the quantity  $-2 \log \hat{L}$  is sometimes referred to as a *deviance*. This is also inappropriate, since unlike the deviance used in the context of generalised linear modelling,  $-2 \log \hat{L}$  does not measure deviation from a model that is a perfect fit to the data.

### *Example 3.3 Prognosis for women with breast cancer*

Consider again the data from Example 1.2 on the survival times of breast cancer patients. On fitting a proportional hazards model that contains no explanatory variables, that is, the null model, the value of  $-2 \log \hat{L}$  is 173.968. As in Example 3.1, the indicator variable  $X$ , will be used to represent the result of the staining procedure, so that  $X$  is zero for women whose tumours are negatively stained and unity otherwise. When the variable  $X$  is included in the linear component of the model, the value of  $-2 \log \hat{L}$  decreases to 170.096. The values of  $-2 \log \hat{L}$  for alternative models are conveniently summarised in tabular form, as illustrated in Table 3.2.

The difference between the values of  $-2 \log \hat{L}$  for the null model and the model that contains  $X$  can be used to assess the significance of the difference between the hazard functions for the two groups of women. Since one model contains one more  $\beta$ -parameter than the other, the difference in the values of  $-2 \log \hat{L}$  has a chi-squared distribution on one degree of freedom. The differ-

**Table 3.2** Values of  $-2 \log \hat{L}$  on fitting proportional hazards models to the data from Example 1.2.

Variables in model	$-2 \log \hat{L}$
none	173.968
$X$	170.096

ence in the two values of  $-2 \log \hat{L}$  is  $173.968 - 170.096 = 3.872$ , which is just significant at the 5% level ( $P = 0.049$ ). We may therefore conclude that there is evidence, significant at the 5% level, that the hazard functions for the two groups of women are different.

In Example 2.12, the extent of the difference between the survival times of the two groups of women was investigated using the log-rank test. The chi-squared value for this test was found to be 3.515 ( $P = 0.061$ ). This value is not very different from the figure of 3.872 ( $P = 0.049$ ) obtained above. The similarity of these two  $P$ -values means that essentially the same conclusions are drawn about the extent to which the data provide evidence against the null hypothesis of no group difference. From the practical viewpoint, the fact that one result is just significant at the 5% level, while the other is not quite significant at that level, is immaterial.

Although the model-based approach used in this example is operationally different from the log-rank test, the two procedures are in fact closely related. This relationship will be explored in greater detail in Section 3.9.

### *Example 3.4 Treatment of hypernephroma*

In a study carried out at the University of Oklahoma Health Sciences Center, data were obtained on the survival times of 36 patients with a malignant tumour in the kidney, or hypernephroma. The patients had all been treated with a combination of chemotherapy and immunotherapy, but additionally a nephrectomy, the surgical removal of the kidney, had been carried out on some of the patients. Of particular interest is whether the survival time of the patients depends on their age at the time of diagnosis and on whether or not they had received a nephrectomy. The data obtained in the study were given in Lee and Wang (2003). In the data set to be used as a basis for this example, the age of a patient has been classified according to whether the patient is less than 60, between 60 and 70 or greater than 70. Table 3.3 gives the survival times of the patients in months, where an asterisk denotes a censored observation.

In this example, there is a factor, age group, with three levels ( $< 60$ ,  $60-70$ ,  $> 70$ ), and a factor associated with whether or not a nephrectomy was performed. There are a number of possible models for these data depending on whether the hazard function is related to neither, one or both of these factors. Suppose that the effect due to the  $j$ th age group is denoted by  $\alpha_j$ ,  $j = 1, 2, 3$ ,

**Table 3.3** Survival times of 36 patients classified according to age group and whether or not they have had a nephrectomy.

No nephrectomy			Nephrectomy		
< 60	60–70	> 70	< 60	60–70	> 70
9	15	12	104*	108*	10
6	8		9	26	9
21	17		56	14	18
			35	115	6
			52	52	
			68	5*	
			77*	18	
			84	36	
			8	9	
			38		
			72		
			36		
			48		
			26		
			108		
			5		

and that due to nephrectomy status is denoted by  $\nu_k$ ,  $k = 1, 2$ . The terms  $\alpha_j$  and  $\nu_k$  may then be included in proportional hazards models for  $h_i(t)$ , the hazard function for the  $i$ th individual in the study. Five possible models are as follows:

$$\text{Model (1): } h_i(t) = h_0(t);$$

$$\text{Model (2): } h_i(t) = \exp\{\alpha_j\}h_0(t);$$

$$\text{Model (3): } h_i(t) = \exp\{\nu_k\}h_0(t);$$

$$\text{Model (4): } h_i(t) = \exp\{\alpha_j + \nu_k\}h_0(t);$$

$$\text{Model (5): } h_i(t) = \exp\{\alpha_j + \nu_k + (\alpha\nu)_{jk}\}h_0(t).$$

Under Model (1), the hazard of death does not depend on either of the two factors and is the same for all 36 individuals in the study. In Models (2) and (3), the hazard depends on either the age group or on whether a nephrectomy was performed, but not on both. In Model (4), the hazard depends on both factors, where the impact of nephrectomy on the hazard is independent of the age group of the patient. Model (5) includes an interaction between age group and nephrectomy, so that under this model the effect of a nephrectomy on the hazard of death depends on the age group of the patient.

To fit the term  $\alpha_j$ , two indicator variables  $A_2$  and  $A_3$  are defined with values shown in the following table.

Age group	$A_2$	$A_3$
< 60	0	0
60–70	1	0
> 70	0	1

The term  $\nu_k$  is fitted by defining a variable  $N$  which takes the value zero when no nephrectomy has been performed and unity when it has. With this choice of indicator variables, the baseline hazard function will correspond to an individual in the youngest age group who has not had a nephrectomy.

Models that contain the term  $\alpha_j$  are then fitted by including the variables  $A_2, A_3$  in the model, while the term  $\nu_k$  is fitted by including  $N$ . The interaction is fitted by including the products  $A_2N = A_2 \times N$  and  $A_3N = A_3 \times N$  in the model. The explanatory variables fitted, and the values of  $-2 \log \hat{L}$  for each of the five models under consideration, are shown in Table 3.4. Some computer software for modelling survival data enables factors to be included in a model without the user having to define appropriate indicator variables. The values of  $-2 \log \hat{L}$  in Table 3.4 could then have been obtained directly using such software.

**Table 3.4** Values of  $-2 \log \hat{L}$  on fitting five models to the data in Table 3.3.

Terms in model	Variables in model	$-2 \log \hat{L}$
null model	none	177.667
$\alpha_j$	$A_2, A_3$	172.172
$\nu_k$	$N$	170.247
$\alpha_j + \nu_k$	$A_2, A_3, N$	165.508
$\alpha_j + \nu_k + (\alpha\nu)_{jk}$	$A_2, A_3, N, A_2N, A_3N$	162.479

The first step in comparing these different models is to determine if there is an interaction between nephrectomy status and age group. To do this, Model (4) is compared with Model (5). The reduction in the value of  $-2 \log \hat{L}$  on including the interaction term in the model that contains the main effects of age group and nephrectomy status is  $165.508 - 162.479 = 3.029$  on 2 d.f. This is not significant ( $P = 0.220$ ) and so we conclude that there is no interaction between age group and whether or not a nephrectomy has been performed.

We now determine whether the hazard function is related to neither, one or both of the factors age group and nephrectomy status. The change in the value of  $-2 \log \hat{L}$  on including the term  $\alpha_j$  in the model that contains  $\nu_k$  is  $170.247 - 165.508 = 4.739$  on 2 d.f. This is significant at the 10% level ( $P = 0.094$ ) and so there is some evidence that  $\alpha_j$  is needed in a model

that contains  $\nu_k$ . The change in  $-2 \log \hat{L}$  when  $\nu_k$  is added to the model that contains  $\alpha_j$  is  $172.172 - 165.508 = 6.664$  on 1 d.f., which is significant at the 1% level ( $P = 0.010$ ). Putting these two results together, the term  $\alpha_j$  may add something to the model that includes  $\nu_k$ , and  $\nu_k$  is certainly needed in the model that contains  $\alpha_j$ . This means that both terms are required, and that the hazard function depends on both the patient's age group and on whether or not a nephrectomy has been carried out.

Before leaving this example, let us consider other possible results from the comparison of the five models, and how they would affect the conclusion as to which model is the most appropriate. If the term corresponding to age group,  $\alpha_j$ , was needed in a model in addition to the term corresponding to nephrectomy status,  $\nu_k$ , and yet  $\nu_k$  was not needed in the presence of  $\alpha_j$ , the model containing just  $\alpha_j$ , Model (2), is probably the most suitable. To make sure that  $\alpha_j$  was needed at all, Model (2) would be further compared with Model (1), the null model. Similarly, if the term corresponding to nephrectomy status,  $\nu_k$ , was needed in addition to the term corresponding to age group,  $\alpha_j$ , but  $\alpha_j$  was not required in the presence of  $\nu_k$ , Model (3) would probably be satisfactory. However, the significance of  $\nu_k$  would be checked by comparing Model (3) with Model (1). If neither of the terms corresponding to age group and nephrectomy status were needed in the presence of the other, a maximum of one variable would be required. To determine which of the two is necessary, Model (2) would be compared with Model (1) and Model (3) with Model (1). If both results were significant, on statistical grounds, the model that leads to the biggest reduction in the value of  $-2 \log \hat{L}$  from that for the null model would be adopted. If neither Model (2) nor Model (3) was superior to Model (1), we would conclude that neither age group nor nephrectomy status had an effect on the hazard function.

There are two further steps in the modelling approach to the analysis of survival data. First, we will need to critically examine the fit of a model to the observed data in order to ensure that the fitted proportional hazards model is indeed appropriate. Second, we will need to interpret the model, in order to quantify the effect that the explanatory variables have on the hazard function. Interpretation of parameters in a fitted model is considered in Section 3.7, while methods for assessing the adequacy of a fitted model will be considered in Chapter 4. But first, some general comments are made on possible strategies for model selection.

### 3.6 Strategy for model selection

An initial step in the model selection process is to identify a set of explanatory variables that have the potential for being included in the linear component of a proportional hazards model. This set will contain those variates and factors that have been recorded for each individual, but additionally terms corresponding to interactions between factors or between variates and factors may also be required.

Once a set of potential explanatory variables has been isolated, the combination of variables that are to be used in modelling the hazard function has to be determined. In practice, a hazard function will not depend on a unique combination of variables. Instead, there are likely to be a number of equally good models, rather than a single "best" model. For this reason, it is desirable to consider a wide range of possible models.

An important principle in statistical modelling is that when a term corresponding to the interaction between two factors is to be included in a model, the corresponding lower-order terms should also be included. This rule is known as the *hierarchic principle*, and means that interactions should not be fitted unless the corresponding main effects are present. Models that are not hierarchic are difficult to interpret.

The model selection strategy depends to some extent on the purpose of the study. In some applications, information on a number of variables will have been obtained, and the aim might be to determine which of them has an effect on the hazard function, as in Example 1.3 on multiple myeloma. In other situations, there may be one or more variables of primary interest, such as terms corresponding to a treatment effect. The aim of the modelling process is then to evaluate the effect of such variables on the hazard function, as in Example 1.4 on prostatic cancer. Since the other variables that have been recorded might also be expected to influence the magnitude of the treatment effect, these variables will need to be taken account of in the modelling process.

#### 3.6.1 Variable selection procedures

We first consider the situation where all explanatory variables are on an equal footing, and the aim is to identify subsets of variables upon which the hazard function depends. When the number of potential explanatory variables, including interactions, non-linear terms and so on, is not too large, it might be feasible to fit all possible combinations of terms, paying due regard to the hierarchic principle. Alternative nested models can be compared by examining the change in the value of  $-2 \log \hat{L}$  on adding terms into a model or deleting terms from a model.

Comparisons between a number of possible models, which need not necessarily be nested, can also be made on the basis of the statistic

$$AIC = -2 \log \hat{L} + \alpha q,$$

in which  $q$  is the number of unknown  $\beta$ -parameters in the model and  $\alpha$  is a predetermined constant. This statistic is known as *Akaike's information criterion*; the smaller the value of this statistic, the better the model. The motivation behind this statistic is that if one includes unnecessary covariates, the values of  $AIC$  for the two models will not be very different. Indeed, the value of  $AIC$  will tend to increase when unnecessary terms are added to the model.

Values of  $\alpha$  between 2 and 6 are generally used in computing the value of the statistic. The choice  $\alpha = 3$  is roughly equivalent to using a 5% significance

level in judging the difference between the values of  $-2 \log \hat{L}$  for two nested models that differ by between one and three parameters. This value of  $\alpha$  is recommended for general use.

Of course, some terms may be identified as alternatives to those in a particular model, leading to subsets that are equally suitable. The decision on which of these subsets is the most appropriate should not then rest on statistical grounds alone. When there are no subject matter grounds for model choice, the model chosen for initial consideration from a set of alternatives might be the one for which the value of  $-2 \log \hat{L}$  or  $AIC$  is a minimum. It will then be important to confirm that the model does fit the data using the methods for model checking described in Chapter 4.

In some applications, information might be recorded on a number of variables, all of which relate to the same general feature. For example, the variables height, weight, body mass index (weight/height<sup>2</sup>), head circumference, arm length, and so on, are all concerned with the size of an individual. In view of inter-relationships between these variables, a model for the survival times of these individuals may not need to include each of them. It would then be appropriate to determine which variables from this group should be included in the model, although it may not matter exactly which variables are chosen.

When the number of variables is relatively large, it can be computationally expensive to fit all possible models. In particular, if there is a pool of  $p$  potential explanatory variables, there are  $2^p$  possible combinations of terms, so that if  $p > 10$ , there are more than a thousand possible combinations of explanatory variables. In this situation, automatic routines for variable selection that are available in many software packages might seem an attractive prospect. These routines are based on *forward selection*, *backward elimination* or a combination of the two known as the *stepwise procedure*.

In forward selection, variables are added to the model one at a time. At each stage in the process, the variable added is the one that gives the largest decrease in the value of  $-2 \log \hat{L}$  on its inclusion. The process ends when the next candidate for inclusion in the model does not reduce the value of  $-2 \log \hat{L}$  by more than a prespecified amount. This is known as the *stopping rule*. This rule is often couched in terms of the significance level of the difference in the values of  $-2 \log \hat{L}$  when a variable is added to a model, so that the selection process ends when the next term for inclusion ceases to be significant at a pre-assigned level.

In backward elimination, a model that contains the largest number of variables under consideration is first fitted. Variables are then excluded one at a time. At each stage, the variable omitted is the one that increases the value of  $-2 \log \hat{L}$  by the smallest amount on its exclusion. The process ends when the next candidate for deletion increases the value of  $-2 \log \hat{L}$  by more than a prespecified amount.

The stepwise procedure operates in the same way as forward selection. However, a variable that has been included in the model can be considered for exclusion at a later stage. Thus after adding a variable to the model, the procedure then checks whether any previously included variable can now be

deleted. These decisions are again made on the basis of prespecified stopping rules.

These automatic routines have a number of disadvantages. Typically, they lead to the identification of one particular subset, rather than a set of equally good ones. The subsets found by these routines often depend on the variable selection process that has been used, that is, whether it is forward selection, backward elimination or the stepwise procedure, and generally tend not to take any account of the hierachic principle. They also depend on the stopping rule that is used to determine whether a term should be included in or excluded from a model. For all these reasons, these automatic routines have a limited role in model selection, and should certainly not be used uncritically.

Instead of using automatic variable selection procedures, the following general strategy for model selection is recommended.

1. The first step is to fit models that contain each of the variables one at a time. The values of  $-2 \log \hat{L}$  for these models are then compared with that for the null model to determine which variables on their own significantly reduce the value of this statistic.
2. The variables that appear to be important from Step 1 are then fitted together. In the presence of certain variables, others may cease to be important. Consequently, those variables that do not significantly increase the value of  $-2 \log \hat{L}$  when they are omitted from the model can now be discarded. We therefore compute the change in the value of  $-2 \log \hat{L}$  when each variable on its own is omitted from the set. Only those that lead to a significant increase in the value of  $-2 \log \hat{L}$  are retained in the model. Once a variable has been dropped, the effect of omitting each of the remaining variables in turn should be examined.
3. Variables that were not important on their own, and so were not under consideration in Step 2, may become important in the presence of others. These variables are therefore added to the model from Step 2, one at a time, and any that reduce  $-2 \log \hat{L}$  significantly are retained in the model. This process may result in terms in the model determined at Step 2 ceasing to be significant.
4. A final check is made to ensure that no term in the model can be omitted without significantly increasing the value of  $-2 \log \hat{L}$ , and that no term not included significantly reduces  $-2 \log \hat{L}$ .

When using this selection procedure, rigid application of a particular significance level should be avoided. In order to guide decisions on whether to include or omit a term, the significance level should not be too small; a level of around 10% is recommended.

In some applications, a small number of interactions and other higher-order terms, such as powers of certain variates, may need to be considered for inclusion in a model. Such terms would be added to the model identified in Step 3 above, after ensuring that any terms necessitated by the hierachic principle have already been included in the model. If any higher-order term leads to a

significant reduction in the value of  $-2 \log \hat{L}$ , that term would be included in the model.

The procedure outlined above is now illustrated in an example.

#### *Example 3.5 Survival of multiple myeloma patients*

The analysis of the data on the survival times of multiple myeloma patients in Example 3.2 suggested that not all of the seven explanatory variables, *Age*, *Sex*, *Bun*, *Ca*, *Hb*, *Pcells*, and *Protein*, are needed in a proportional hazards model. We now determine the most appropriate subsets of these variables. In this example, transformations of the original variables and interactions between them will not be considered. We will further assume that there are no medical grounds for including particular variables in a model. A summary of the values of  $-2 \log \hat{L}$  for all models that are to be considered is given in Table 3.5.

**Table 3.5** Values of  $-2 \log \hat{L}$  for models fitted to the data from Example 1.3.

Variables in model	$-2 \log \hat{L}$
none	215.940
<i>Age</i>	215.817
<i>Sex</i>	215.906
<i>Bun</i>	207.453
<i>Ca</i>	215.494
<i>Hb</i>	211.068
<i>Pcells</i>	215.875
<i>Protein</i>	213.890
<i>Hb</i> + <i>Bun</i>	202.938
<i>Hb</i> + <i>Protein</i>	209.829
<i>Bun</i> + <i>Protein</i>	203.641
<i>Bun</i> + <i>Hb</i> + <i>Protein</i>	200.503
<i>Hb</i> + <i>Bun</i> + <i>Age</i>	202.669
<i>Hb</i> + <i>Bun</i> + <i>Sex</i>	202.553
<i>Hb</i> + <i>Bun</i> + <i>Ca</i>	202.937
<i>Hb</i> + <i>Bun</i> + <i>Pcells</i>	202.773

The first step is to fit the null model and models that contain each of the seven explanatory variables on their own. Of these variables, *Bun* leads to the largest reduction in  $-2 \log \hat{L}$ , reducing the value of the statistic from 215.940 to 207.453. This reduction of 8.487 is significant at the 1% level ( $P = 0.004$ ) when compared with percentage points of the chi-squared distribution on 1 d.f. The reduction in  $-2 \log \hat{L}$  on adding *Hb* to the null model is 4.872, which is also significant at the 5% level ( $P = 0.027$ ). The only other variable that on its own has some explanatory power is *Protein*, which leads to a reduction in  $-2 \log \hat{L}$  that is nearly significant at the 15% level ( $P = 0.152$ ). Although

this *P*-value is relatively high, we will for the moment keep *Protein* under consideration for inclusion in the model.

The next step is to fit the model that contains *Bun*, *Hb* and *Protein*, which leads to a value of  $-2 \log \hat{L}$  of 200.503. The effect of omitting each of the three variables in turn from this model is shown in Table 3.5. In particular, when *Bun* is omitted, the increase in  $-2 \log \hat{L}$  is 9.326, when *Hb* is omitted the increase is 3.138, and when *Protein* is omitted it is 2.435. Each of these changes in the value of  $-2 \log \hat{L}$  can be compared with percentage points of a chi-squared distribution on 1 d.f. Since *Protein* does not appear to be needed in the model, in the presence of *Hb* and *Bun*, this variable will not be further considered for inclusion.

If either *Hb* or *Bun* is excluded from the model that contains both of these variables, the increase in  $-2 \log \hat{L}$  is 4.515 and 8.130, respectively. Both of these increases are significant at the 5% level, and so neither *Hb* nor *Bun* can be excluded from the model without significantly increasing the value of the  $-2 \log \hat{L}$  statistic.

Finally, we look to see if any of variables *Age*, *Sex*, *Ca* and *Pcells* should be included in the model that contains *Bun* and *Hb*. Table 3.5 shows that when any of these four variables is added, the reduction in  $-2 \log \hat{L}$  is less than 0.5, and so none of them need to be included in the model. We therefore conclude that the most satisfactory model is that containing *Bun* and *Hb*.

We now turn to studies where there are variables of primary importance, such as a treatment effect. Here, we proceed in the following manner.

1. The important prognostic variables are first selected, ignoring the treatment effect. Models with all possible combinations of the variables can be fitted when their number is not too large. Alternatively, the variable selection process might follow similar lines to those described previously in Steps 1 to 4.
2. The treatment effect is then included in the model. In this way, any differences between the two groups that arise as a result of differences between the distributions of the prognostic variables in each treatment group, are not attributed to the treatment.
3. If the possibility of interactions between the treatment and other explanatory variables has not been discounted, these must be considered before the treatment effect can be interpreted.

It will often be interesting to fit a model that contains the treatment effect alone. This enables the effect that the prognostic variables have on the magnitude of the treatment effect to be evaluated.

In this discussion on strategies for model selection, the use of statistical criteria to guide the selection process has been emphasised. In addition, due account must be taken of the application area. In particular, on subject area grounds, it may be inappropriate to include particular combinations of variables. On the other hand, there might be some variables that it is not sensible to omit from the model, even if they appear not to be needed in modelling a

particular data set. Indeed, there is always a need for non-statistical considerations in model building.

*Example 3.6 Comparison of two treatments for prostatic cancer*

In the data from Example 1.4 on the survival times of 38 prostatic cancer patients, there are four prognostic variables that might have an effect on the survival times. These are the age of the patient in years (*Age*), serum haemoglobin level (*Shb*), tumour size (*Size*) and Gleason index (*Index*). All possible combinations of these variates are fitted in a proportional hazards model and the values of  $-2 \log \hat{L}$  computed. These values are shown in Table 3.6, together with the values of Akaike's information criterion, computed with  $\alpha = 3$ .

**Table 3.6** Values of  $-2 \log \hat{L}$  and AIC for models fitted to the data from Example 1.4.

Variables in model	$-2 \log \hat{L}$	AIC
none	36.349	36.349
<i>Age</i>	36.269	39.269
<i>Shb</i>	36.196	39.196
<i>Size</i>	29.042	32.042
<i>Index</i>	29.127	32.127
<i>Age + Shb</i>	36.151	42.151
<i>Age + Size</i>	28.854	34.854
<i>Age + Index</i>	28.760	34.760
<i>Shb + Size</i>	29.019	35.019
<i>Shb + Index</i>	27.981	33.981
<i>Size + Index</i>	23.533	29.533
<i>Age + Shb + Size</i>	28.852	37.852
<i>Age + Shb + Index</i>	27.893	36.893
<i>Age + Size + Index</i>	23.269	32.269
<i>Shb + Size + Index</i>	23.508	32.508
<i>Age + Shb + Size + Index</i>	23.231	35.231

The two most important explanatory variables when considered separately are *Size* and *Index*. From the change in the value of  $-2 \log \hat{L}$  on omitting either of them from a model that contains both, we deduce that both variables are needed in a proportional hazards model. The value of  $-2 \log \hat{L}$  is only reduced by a very small amount when *Age* and *Shb* are added to the model that contains *Size* and *Index*. We therefore conclude that only *Size* and *Index* are important prognostic variables.

From the values of Akaike's information criterion in Table 3.6, the model with *Size* and *Index* leads to the smallest value of the statistic, confirming that this is the most suitable model of those tried. Notice also that there are no other combinations of explanatory variables that lead to similar values of

the *AIC*-statistic, which shows that there are no obvious alternatives to using *Size* and *Index* in the model.

We now consider the treatment effect. Let *Treat* be a variable that takes the value zero for individuals allocated to the placebo, and unity for those allocated to DES. When *Treat* is added to the model that contains *Size* and *Index*, the value of  $-2 \log \hat{L}$  is reduced to 22.572. This reduction of 0.961 on 1 d.f. is not significant ( $P = 0.327$ ). This indicates that there is no treatment effect, but first we ought to examine whether the coefficients of the two explanatory variables in the model depend on treatment. To do this, we form the products *Tsize* = *Treat* × *Size* and *Tindex* = *Treat* × *Index*, and add these to the model that contains *Size*, *Index* and *Treat*. When *Tsize* and *Tindex* are added to the model,  $-2 \log \hat{L}$  is reduced to 20.829 and 20.792, respectively. On adding both of these mixed terms,  $-2 \log \hat{L}$  becomes 19.705. The reductions in  $-2 \log \hat{L}$  on adding these terms to the model are not significant, and so there is no evidence that the treatment effect depends on *Size* and *Index*. This means that our original interpretation of the size of the treatment effect is valid, and that on the basis of these data, treatment with DES does not appear to affect the hazard of death. The estimated size of this treatment effect will be considered later in Example 3.10.

Before leaving this example, we note that when either *Tsize* or *Tindex* is added to the model, their estimated coefficient, and that of *Treat*, become large. The standard errors of these estimates are also very large. In particular, in the model that contains *Size*, *Index*, *Treat* and *Tsize*, the estimated coefficient of *Treat* is -11.28 with a standard error of 18.50. For the model that contains *Size*, *Index*, *Treat* and *Tindex*, the coefficients of *Treat* and *Tindex* are -161.52 and 14.66, respectively, while the standard errors of these estimates are 18476 and 1680, respectively! This is evidence of *overfitting*.

In an overfitted model, the estimated values of some of the  $\beta$ -coefficients will be highly dependent on the actual data. A very slight change to the values of one of these variables could then have a large impact on the estimate of the corresponding coefficient. This is the reason for such estimates having large standard errors.

An overfitted model is one that is more complicated than is justified by the data, and does not provide a useful summary of the data. This is another reason for not including the mixed terms in the model for the hazard of death from prostatic cancer.

### 3.6.2 Testing for non-linearity

When the dependence of the hazard function on a variate that takes a wide range of values is to be modelled, we should consider whether the variate should be included as a linear term in the proportional hazards model.

For some variates, transformations of their original values might be used in place of the original variate. For example, if a variate takes a wide range of values, that variate might first be transformed by taking logarithms. This is particularly appropriate for variates that are strictly positive. The logarithm

of a variate may also be used when the distribution of the values of the variate is highly positively skew.

When there are no *a priori* reasons for transforming a variate, the assumption of linearity in the variate should at least be critically examined. One possibility is to add quadratic or even cubic terms to the model, and examine the consequent reduction in the value of  $-2 \log \hat{L}$ . If the inclusion of such terms significantly reduces the value of this statistic, we would conclude that there is non-linearity, and incorporate the polynomial terms in the model.

In many situations, non-linearity in an explanatory variate cannot be adequately represented by including polynomial terms in a model, or by transforming the original variable. For this reason, the following procedure is recommended for general use.

The values of the variate are first grouped into four or five categories containing approximately equal numbers of observations. A factor is then defined whose levels correspond to this grouping. For example, a variate reflecting the size of a tumour could be fitted as a factor whose levels correspond to very small, small, medium and large.

More specifically, let  $A$  be a factor with  $m$  levels formed from a continuous variate, and let  $X$  be a variate that takes the value  $j$  when  $A$  is at level  $j$ , for  $j = 1, 2, \dots, m$ . Linearity in the original variate will then correspond to there being a linear trend across the levels of  $A$ . This linear trend can be modelled by fitting  $X$  alone. Now, fitting the  $m - 1$  terms  $X, X^2, \dots, X^{m-1}$  is equivalent to fitting  $A$  as a factor in the model, using indicator variables as in Section 3.2.1. Accordingly, the difference between the value of  $-2 \log \hat{L}$  for the model that contains  $X$ , and that for the model that contains  $A$ , is a measure of non-linearity across the levels of  $A$ . If this difference is not significant we would conclude that there is no non-linearity and the original variate would be fitted. On the other hand, if there is evidence of non-linearity, the factor which corresponds to the variate is fitted.

The actual form of the non-linearity can be further studied from the coefficients of the indicator variables corresponding to  $A$ . Indeed, a plot of these coefficients may help in establishing the nature of any trend across the levels of the factor  $A$ .

#### *Example 3.7 Survival of multiple myeloma patients*

In Example 3.5, we found that a proportional hazards model that contained the explanatory variables  $Bun$  and  $Hb$  appeared to be appropriate for the data on the survival times of multiple myeloma patients. We now consider whether there is any evidence of non-linearity in the values of serum haemoglobin level, and examine whether a quadratic term is needed in the proportional hazards model that contains  $Bun$  and  $Hb$ . When the term  $Hb^2$  is added to this model, the value of  $-2 \log \hat{L}$  is reduced from 202.938 to 202.917. This reduction of 0.021 on 1 d.f. is clearly not significant, which suggests that a linear term in  $Hb$  is sufficient.

An alternative way of examining the extent of non-linearity is to use a factor to model the effect of serum haemoglobin level on the hazard function.

Suppose that a factor with four levels is defined, where level 1 corresponds to values of  $Hb$  less than or equal to 7, level 2 to values between 7 and 10, level 3 to values between 10 and 13 and level 4 to values greater than 13. This choice of levels corresponds roughly to the quartiles of the distribution of the values of  $Hb$ . This factor can be fitted by defining three indicator variables,  $Hb2$ ,  $Hb3$  and  $Hb4$ , which take the values shown in the following table.

Level of factor ( $X$ )	Value of $Hb$	$Hb2$	$Hb3$	$Hb4$
1	$Hb \leq 7$	0	0	0
2	$7 < Hb \leq 10$	1	0	0
3	$10 < Hb \leq 13$	0	1	0
4	$Hb > 13$	0	0	1

When a model containing  $Bun$ ,  $Hb2$ ,  $Hb3$  and  $Hb4$  is fitted, the value of  $-2 \log \hat{L}$  is 200.417. The change in the value of this statistic on adding the indicator variables  $Hb2$ ,  $Hb3$  and  $Hb4$  to the model that contains  $Bun$  alone is 7.036 on 3 d.f., which is significant at the 10% level ( $P = 0.071$ ). However, it is difficult to identify any pattern across the factor levels.

A linear trend across the levels of the factor corresponding to haemoglobin level can be modelled by fitting the variate  $X$ , which takes values 1, 2, 3, 4, according to the factor level. When the model containing  $Bun$  and  $X$  is fitted,  $-2 \log \hat{L}$  is 203.891, and the change in the value of  $-2 \log \hat{L}$  due to any non-linearity is  $203.891 - 200.417 = 3.474$  on 2 d.f. This is not significant when compared with percentage points of the chi-squared distribution on 2 d.f. ( $P = 0.176$ ). We therefore conclude that the effect of haemoglobin level on the hazard of death in this group of patients is adequately modelled by using the linear term  $Hb$ .

### 3.7 Interpretation of parameter estimates

When the proportional hazards model is used in the analysis of survival data, the coefficients of the explanatory variables in the model can be interpreted as logarithms of the ratio of the hazard of death to the baseline hazard. This means that estimates of this hazard ratio, and corresponding confidence intervals, can easily be found from the fitted model. The interpretation of parameters corresponding to different types of term in the proportional hazards model is described in the following sections.

#### 3.7.1 Models with a variate

Suppose that a proportional hazards model contains a single continuous variable  $X$ , so that the hazard function for the  $i$ th of  $n$  individuals, for whom  $X$  takes the value  $x_i$ , is

$$h_i(t) = e^{\beta x_i} h_0(t).$$

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  $r \text{se}(\hat{\beta})$ , 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,  $m \geq 2$ , 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  $j$ th group,  $j = 1, 2, \dots, m$ , is given by

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

where  $\gamma_j$  is the effect due to the  $j$ th 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  $j$ th group,  $j \geq 2$ , relative to an individual in the first group, is then  $\exp(\gamma_j)$ . Consequently, the parameter  $\gamma_j$  is the logarithm of this relative hazard, that is,

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

A model that contains the terms  $\gamma_j$ ,  $j = 1, 2, \dots, m$ , with  $\gamma_1 = 0$ , can be fitted by defining  $m - 1$  indicator variables,  $X_2, X_3, \dots, X_m$ , as shown in Section 3.2.2. Fitting this model leads to estimates  $\hat{\gamma}_2, \hat{\gamma}_3, \dots, \hat{\gamma}_m$ , 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  $\hat{\gamma}_j - z_{\alpha/2} \text{se}(\hat{\gamma}_j)$  to  $\hat{\gamma}_j + z_{\alpha/2} \text{se}(\hat{\gamma}_j)$ , 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  $j$ th age group is denoted by  $\alpha_j$ ,  $j = 1, 2, 3$ , the proportional hazards model for the hazard at time  $t$  for a patient in the  $j$ th 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  $\alpha_j$  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  $\alpha_2$  and  $\alpha_3$ , 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.
$\alpha_2$	-0.065	0.498
$\alpha_3$	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

$$\begin{matrix} & \begin{matrix} \hat{\alpha}_2 & \hat{\alpha}_3 \end{matrix} \\ \begin{matrix} \hat{\alpha}_2 \\ \hat{\alpha}_3 \end{matrix} & \begin{pmatrix} 0.2484 & 0.0832 \\ 0.0832 & 0.4649 \end{pmatrix} \end{matrix},$$

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  $\hat{\alpha}_3 - \hat{\alpha}_2$  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,  $X_2, X_3, \dots, X_m$ , 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  $j$ th level,  $j = 2, 3, \dots, m$ . 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  $j$ th level of the factor is

$$\exp(\alpha_j),$$

for  $j \geq 2$ . The ratio of the hazard for an individual in group  $j$ ,  $j \geq 2$ , 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

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

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  $i$ th individual can then be expressed in the form

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

for  $i = 1, 2, \dots, 38$ . Estimates of the  $\beta$ -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( $\hat{\beta}$ )
<i>Size</i>	0.083	0.048
<i>Index</i>	0.710	0.338
<i>Treat</i>	-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  $\alpha_j$ ,  $j = 1, 2, 3$ , corresponding to age group, and terms  $\nu_k$ ,  $k = 1, 2$ , 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  $j$ th age group and the  $k$ th level of nephrectomy status is

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

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  $j$ th age group,  $j = 1, 2, 3$ , and the  $k$ th 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  $j$ th age group,  $j = 1, 2, 3$ , and the  $k$ th 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  $\beta$ -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,  $X_1, X_2, \dots, X_p$ , and that the estimated coefficients of these variables are  $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ . The estimated hazard function for the  $i$ th of  $n$  individuals in the study is then given by

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

where  $\mathbf{x}_i$  is the vector of values of the explanatory variables for the  $i$ th individual,  $i = 1, 2, \dots, n$ ,  $\hat{\beta}$  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  $t_{(1)} < t_{(2)} < \dots < t_{(r)}$ , 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

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

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

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

for  $j = 1, 2, \dots, r$ . In equation (3.16),  $D(t_{(j)})$  is the set of all  $d_j$  individuals who die at the  $j$ th 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  $\beta$ 's, which form the vector  $\hat{\beta}$ , 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  $j = 1, 2, \dots, r$ , the left-hand side of equation (3.16) will be a single term. This equation can then be solved to give

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

where  $\mathbf{x}_{(j)}$  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

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

for  $t_{(j)} \leq t < t_{(j+1)}$ ,  $j = 1, 2, \dots, r - 1$ , 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

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

for  $t_{(k)} \leq t < t_{(k+1)}$ ,  $k = 1, 2, \dots, r - 1$ , 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  $t \geq t_{(r)}$ , 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

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

for  $t_{(k)} \leq t < t_{(k+1)}$ ,  $k = 1, 2, \dots, r - 1$ , 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  $\mathbf{x}_i$ . In particular, from equation (3.14), the hazard function is estimated by  $\exp(\hat{\beta}' \mathbf{x}_i) \hat{h}_0(t)$ . Next, integrating both sides of equation (3.14), we get

$$\int_0^t \hat{h}_i(u) du = \exp(\hat{\beta}' \mathbf{x}_i) \int_0^t \hat{h}_0(u) du, \quad (3.20)$$

so that the estimated cumulative hazard function for the  $i$ th individual is

given by

$$\hat{H}_i(t) = \exp(\hat{\beta}' \mathbf{x}_i) \hat{H}_0(t). \quad (3.21)$$

On multiplying each side of equation (3.20) by  $-1$  and exponentiating, and making use of equation (1.5), we find that the estimated survivor function for the  $i$ th individual is

$$\hat{S}_i(t) = \left\{ \hat{S}_0(t) \right\}^{\exp(\hat{\beta}' \mathbf{x}_i)}, \quad (3.22)$$

for  $t_{(k)} \leq t < t_{(k+1)}$ ,  $k = 1, 2, \dots, r - 1$ . Note that once the estimated survivor function,  $\hat{S}_i(t)$ , has been obtained, an estimate of the cumulative hazard function is simply  $-\log \hat{S}_i(t)$ .

### 3.8.1 The special case of no covariates

When there are no covariates, so that we have just a single sample of survival times, equation (3.16) becomes

$$\frac{d_j}{1 - \hat{\xi}_j} = n_j,$$

from which

$$\hat{\xi}_j = \frac{n_j - d_j}{n_j}.$$

Then, the estimated baseline hazard function at time  $t_{(j)}$  is  $1 - \hat{\xi}_j$ , which is  $d_j/n_j$ . The corresponding estimate of the survivor function from equation (3.18) is  $\prod_{j=1}^k \hat{\xi}_j$ , that is,

$$\prod_{j=1}^k \left( \frac{n_j - d_j}{n_j} \right),$$

which is the Kaplan-Meier estimate of the survivor function given earlier in equation (2.4). This shows that the estimate of the survivor function given in equation (3.22) generalises the Kaplan-Meier estimate to the case where the hazard function depends on explanatory variables.

Furthermore, the estimate of the hazard function in equation (3.17) reduces to  $d_j/\{n_j(t_{(j+1)} - t_{(j)})\}$ , which is the estimate of the hazard function given in equation (2.16) of Chapter 2.

### 3.8.2 Some approximations to estimates of the baseline functions

When there are tied survival times, the estimated baseline hazard can only be found by using an iterative method to solve equation (3.16). This iterative process can be avoided by using an approximation to the summation on the left-hand side of equation (3.16).

The term

$$\hat{\xi}_j^{\exp(\hat{\beta}' \mathbf{x}_l)},$$

in the denominator of the left-hand side of equation (3.16), can be written as

$$\exp \left\{ e^{\hat{\beta}' \mathbf{x}_l} \log \hat{\xi}_j \right\},$$

and taking the first two terms in the expansion of the exponent gives

$$\exp \left\{ e^{\hat{\beta}' \mathbf{x}_l} \log \hat{\xi}_j \right\} \approx 1 + e^{\hat{\beta}' \mathbf{x}_l} \log \hat{\xi}_j.$$

Writing  $1 - \tilde{\xi}_j$  for the estimated baseline hazard at time  $t_{(j)}$ , obtained using this approximation, and substituting  $1 + e^{\hat{\beta}' \mathbf{x}_l} \log \tilde{\xi}_j$  for  $\hat{\xi}_j^{\exp(\hat{\beta}' \mathbf{x}_l)}$  in equation (3.16), we find that  $\tilde{\xi}_j$  is such that

$$- \sum_{l \in D(t_{(j)})} \frac{1}{\log \tilde{\xi}_j} = \sum_{l \in R(t_{(j)})} \exp(\hat{\beta}' \mathbf{x}_l).$$

Therefore,

$$\frac{-d_j}{\log \tilde{\xi}_j} = \sum_{l \in R(t_{(j)})} \exp(\hat{\beta}' \mathbf{x}_l),$$

since  $d_j$  is the number of deaths at the  $j$ th ordered death time,  $t_{(j)}$ , and so

$$\tilde{\xi}_j = \exp \left( \frac{-d_j}{\sum_{l \in R(t_{(j)})} \exp(\hat{\beta}' \mathbf{x}_l)} \right). \quad (3.23)$$

From equation (3.18), an estimate of the survivor function, based on the values of  $\tilde{\xi}_j$ , is given by

$$\tilde{S}_0(t) = \prod_{j=1}^k \exp \left( \frac{-d_j}{\sum_{l \in R(t_{(j)})} \exp(\hat{\beta}' \mathbf{x}_l)} \right), \quad (3.24)$$

for  $t_{(k)} \leq t < t_{(k+1)}$ ,  $k = 1, 2, \dots, r - 1$ . From this definition, the estimated survivor function is not necessarily zero at the longest survival time, when that time is uncensored, unlike the estimate in equation (3.18). The estimate of the baseline cumulative hazard function derived from  $\tilde{S}_0(t)$  is

$$\tilde{H}_0(t) = -\log \tilde{S}_0(t) = \sum_{j=1}^k \frac{d_j}{\sum_{l \in R(t_{(j)})} \exp(\hat{\beta}' \mathbf{x}_l)}, \quad (3.25)$$

for  $t_{(k)} \leq t < t_{(k+1)}$ ,  $k = 1, 2, \dots, r - 1$ . This estimate is often referred to as the *Nelson-Aalen estimate* or the *Breslow estimate* of the baseline cumulative hazard function.

When there are no covariates, the estimated baseline survivor function in equation (3.24) becomes

$$\prod_{j=1}^k \exp(-d_j/n_j), \quad (3.26)$$

since  $n_j$  is the number of individuals at risk at time  $t_{(j)}$ . This is the Nelson-Aalen estimate of the survivor function given in equation (2.6) of Chapter 2, and the corresponding estimate of the baseline cumulative hazard function is  $\sum_{j=1}^k d_j/n_j$ , as in Section 2.3.3 of Chapter 2.

A further approximation is found from noting that the expression

$$\frac{-d_j}{\sum_{l \in R(t_{(j)})} \exp(\hat{\beta}' \mathbf{x}_l)},$$

in the exponent of equation (3.23), will tend to be small, unless there are large numbers of ties at particular death times. Taking the first two terms of the expansion of this exponent, and denoting this new approximation to  $\xi_j$  by  $\xi_j^*$  gives

$$\xi_j^* = 1 - \frac{d_j}{\sum_{l \in R(t_{(j)})} \exp(\hat{\beta}' \mathbf{x}_l)}.$$

Adapting equation (3.17), the estimated baseline hazard function in the interval from  $t_{(j)}$  to  $t_{(j+1)}$  is then given by

$$h_0^*(t) = \frac{d_j}{(t_{(j+1)} - t_{(j)}) \sum_{l \in R(t_{(j)})} \exp(\hat{\beta}' \mathbf{x}_l)}, \quad (3.27)$$

for  $t_{(j)} \leq t < t_{(j+1)}$ ,  $j = 1, 2, \dots, r-1$ . Using  $\xi_j^*$  in place of  $\xi_j$  in equation (3.18), the corresponding estimated baseline survivor function is

$$S_0^*(t) = \prod_{j=1}^k \left( 1 - \frac{d_j}{\sum_{l \in R(t_{(j)})} \exp(\hat{\beta}' \mathbf{x}_l)} \right),$$

and a further approximate estimate of the baseline cumulative hazard function is  $H_0^*(t) = -\log S_0^*(t)$ . Notice that the cumulative hazard function in equation (3.25) at time  $t$  can be expressed in the form

$$\tilde{H}_0(t) = \sum_{j=1}^k (t_{(j+1)} - t_{(j)}) h_0^*(t),$$

where  $h_0^*(t)$  is given in equation (3.27). Consequently, differences in successive values of the estimated baseline cumulative hazard function in equation (3.25) provide an approximation to the baseline hazard function, at times  $t_{(1)}, t_{(2)}, \dots, t_{(r)}$ , that can easily be computed.

In the particular case where there are no covariates, the estimates  $h_0^*(t)$ ,  $S_0^*(t)$  and  $H_0^*(t)$  are the same as those given in Section 3.8.1. Equations similar to equations (3.21) and (3.22) can be used to estimate the cumulative hazard and survivor functions for an individual whose vector of explanatory variables is  $\mathbf{x}_i$ .

In practice, it will often be computationally advantageous to use either  $\hat{S}_0(t)$  or  $S_0^*(t)$  in place of  $\tilde{S}_0(t)$ . When the number of tied survival times is small, all three estimates will tend to be very similar. Moreover, since the estimates are

generally used as descriptive summaries of the survival data, small differences between the estimates are unlikely to be of practical importance.

Once an estimate of the survivor function has been obtained, the median and other percentiles of the survival time distribution can be found from tabular or graphical displays of the function for individuals with particular values of explanatory variables. The method used is very similar to that described in Section 2.4, and is illustrated in the following example.

### *Example 3.12 Treatment of hypernephroma*

In Example 3.4, a proportional hazards model was fitted to the data on the survival times of patients with hypernephroma. The hazard function was found to depend on the age group of a patient, and whether or not a nephrectomy had been performed. The estimated hazard function for the  $i$ th patient was found to be

$$\hat{h}_i(t) = \exp\{0.013 A_{2i} + 1.342 A_{3i} - 1.412 N_i\} \hat{h}_0(t),$$

where  $A_{2i}$  is unity if the patient is aged between 60 and 70 and zero otherwise,  $A_{3i}$  is unity if the patient is aged over 70 and zero otherwise, and  $N_i$  is unity if the patient has had a nephrectomy and zero otherwise. The estimated baseline hazard function is therefore the estimated hazard of death at time  $t$ , for an individual whose age is less than 60 and who has not had a nephrectomy.

In Table 3.10, the estimated baseline hazard function,  $\hat{h}_0(t)$ , cumulative hazard function,  $\hat{H}_0(t)$ , and survivor function,  $\hat{S}_0(t)$ , obtained using equations (3.15), (3.19) and (3.18), respectively, are tabulated.

From this table, we see that the general trend is for the estimated baseline hazard function to increase with time. From the manner in which the estimated baseline hazard function has been computed, the estimates only apply at the death times of the patients in the study. However, if the assumption of a constant hazard in each time interval is made, by dividing the estimated hazard by the corresponding time interval, the risk of death per unit time can be found. This leads to the estimate in equation (3.17). A graph of this hazard function is shown in Figure 3.2.

This graph shows that the risk of death per unit time is roughly constant over the duration of the study. Table 3.10, also shows that the values of  $\hat{h}_0(t)$  are very similar to differences in the values of  $\hat{H}_0(t)$  between successive observations, as would be expected.

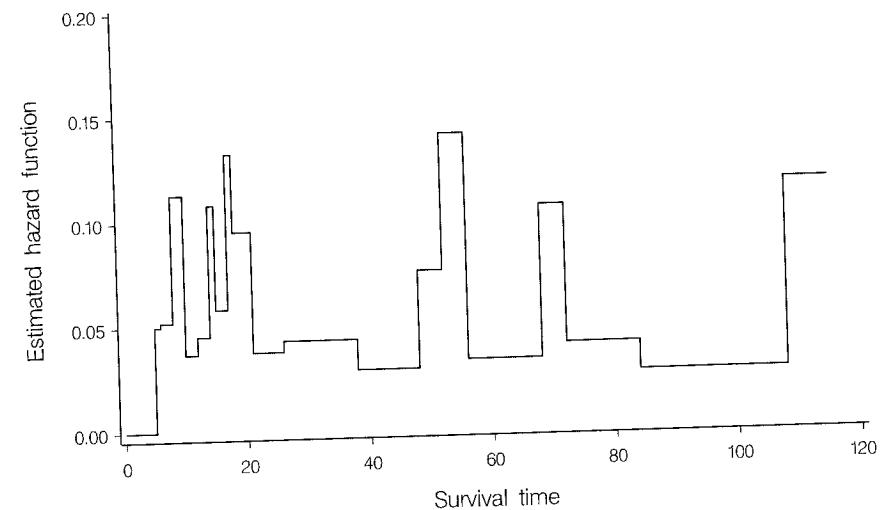
We now consider the estimation of the median survival time, which is the smallest observed survival time for which the estimated survivor function is less than 0.5. From Table 3.10, the estimated median survival time for patients aged less than 60 who have not had a nephrectomy is 12 months.

By raising the estimate of the baseline survivor function to a suitable power, the estimated survivor functions for patients in other age groups, and for patients who have had a nephrectomy, can be obtained through equation (3.22). Thus, the estimated survivor function for the  $i$ th individual is given by

$$\hat{S}_i(t) = \left\{ \hat{S}_0(t) \right\}^{\exp\{0.013A_{2i} + 1.342A_{3i} - 1.412N_i\}}.$$

**Table 3.10** Estimates of the baseline hazard and survivor functions for the data from Example 3.4.

Time	$\hat{h}_0(t)$	$\hat{S}_0(t)$	$\hat{H}_0(t)$
0	0.000	1.000	0.000
5	0.050	0.950	0.051
6	0.104	0.852	0.161
8	0.113	0.755	0.281
9	0.237	0.576	0.552
10	0.073	0.534	0.628
12	0.090	0.486	0.722
14	0.108	0.433	0.836
15	0.116	0.383	0.960
17	0.132	0.333	1.101
18	0.285	0.238	1.436
21	0.185	0.194	1.641
26	0.382	0.120	2.123
35	0.232	0.092	2.387
36	0.443	0.051	2.972
38	0.279	0.037	3.299
48	0.299	0.026	3.655
52	0.560	0.011	4.476
56	0.382	0.007	4.958
68	0.421	0.004	5.504
72	0.467	0.002	6.134
84	0.599	0.001	7.045
108	0.805	0.000	8.692
115	—	0.000	—



**Figure 3.2** Estimated baseline hazard function, per unit time, assuming constant hazard between adjacent death times.

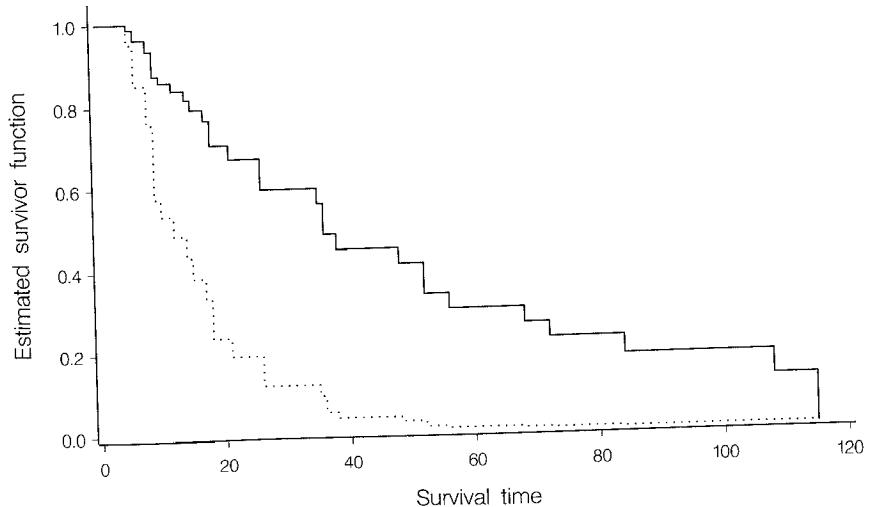
For an individual aged less than 60 who has had a nephrectomy,  $A_2 = 0$ ,  $A_3 = 0$ , and  $N = 1$ , so that the estimated survivor function for this individual is

$$\left\{ \hat{S}_0(t) \right\}^{\exp\{-1.412\}}.$$

This function is plotted in Figure 3.3, together with the estimated baseline survivor function, which is for an individual in the same age group but who has not had a nephrectomy.

This figure shows that the probability of surviving beyond any given time is greater for those who have had a nephrectomy, confirming that a nephrectomy improves the prognosis for patients with hypernephroma.

Note that because of the assumption of proportional hazards, the two estimated survivor functions in Figure 3.3 cannot cross. Moreover, the estimated survivor function, for those who have had a nephrectomy, lies above that of those on whom a nephrectomy has not been performed. This is a direct con-

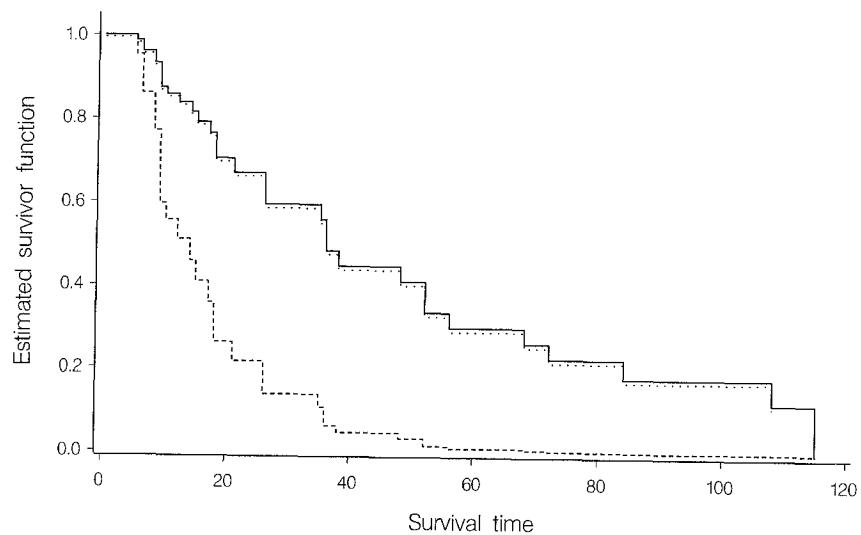


**Figure 3.3** Estimated survivor functions for patients aged less than 60, with (—) and without (···) a nephrectomy.

sequence of the estimated hazard ratio for those who have had the operation, relative to those who have not, being less than unity.

An estimate of the median survival time for this type of patient can be obtained from the tabulated values of the estimated survivor function, or from the graph in Figure 3.3. We find that the estimated median survival time for a patient aged less than 60 who has had a nephrectomy is 36 months. Other percentiles of the distribution of survival times can be estimated using a similar approach.

In a similar manner, the survivor functions for patients in the different age groups can be compared, either for those who have had or not had a nephrectomy. For example, for patients who have had a nephrectomy, the estimated survivor functions for patients in the three age groups are respectively  $\{\hat{S}_0(t)\}^{\exp\{-1.412\}}$ ,  $\{\hat{S}_0(t)\}^{\exp\{-1.412+0.013\}}$  and  $\{\hat{S}_0(t)\}^{\exp\{-1.412+1.342\}}$ . These estimated survivor functions are shown in Figure 3.4.



**Figure 3.4** Estimated survivor functions for patients aged less than 60 (—), between 60 and 70 (···) and greater than 70 (---), who have had a nephrectomy.

This figure clearly shows that patients aged over 70 have a poorer prognosis than those in the other two age groups.

### 3.9\* Proportional hazards modelling and the log-rank test

The proportional hazards model can be used to test the null hypothesis that there is no difference between the hazard functions for two groups of survival times, as illustrated in Example 3.3. This modelling approach therefore provides an alternative to the log-rank test in this situation. However, there is a close connection between the two procedures, which is explored in greater detail in this section.

Following the notation used in Section 2.6.2, and summarised in Table 2.7, the two groups will be labelled Group I and Group II, respectively. The numbers of individuals in the two groups who die at the  $j$ th ordered death time,  $t_{(j)}$ ,  $j = 1, 2, \dots, r$ , will be denoted by  $d_{1j}$  and  $d_{2j}$ , respectively. Similarly, the numbers of individuals at risk in the two groups at time  $t_{(j)}$ , that is, the numbers who are alive and uncensored just prior to this time, will be denoted  $n_{1j}$  and  $n_{2j}$ , respectively.

Now let  $X$  be an indicator variable that is unity when an individual is in Group I and zero when an individual is in Group II. The proportional hazards model for the  $i$ th individual can be written as

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

where  $x_i$  is the value of  $X$  for the  $i$ th individual,  $i = 1, 2, \dots, n$ . When there are no tied observations, that is, when  $d_j = d_{1j} + d_{2j} = 1$ , this model can be fitted by finding that value  $\hat{\beta}$  which maximises the likelihood function in equation (3.4). Denoting the value of  $X$  for the individual who dies at  $t_{(j)}$  by  $x_{(j)}$ , the likelihood function is given by

$$L(\beta) = \prod_{j=1}^r \frac{\exp(\beta x_{(j)})}{\sum_{l=1}^{n_j} \exp(\beta x_l)}, \quad (3.28)$$

since there are  $n_j = n_{1j} + n_{2j}$  individuals in the risk set,  $R(t_{(j)})$ , at time  $t_{(j)}$ , and the corresponding log-likelihood function is

$$\log L(\beta) = \sum_{j=1}^r \beta x_{(j)} - \sum_{j=1}^r \log \left\{ \sum_{l=1}^{n_j} \exp(\beta x_l) \right\}.$$

Since  $x_{(j)}$  is zero for individuals in Group II, the first summation in this expression is over the death times in Group I, and so is simply  $d_1\beta$ , where  $d_1 = \sum_{j=1}^r d_{1j}$  is the total number of deaths in Group I. Also,

$$\sum_{l=1}^{n_j} \exp(\beta x_l) = n_{1j} e^\beta + n_{2j},$$

and so

$$\log L(\beta) = d_1\beta - \sum_{j=1}^r \log \{n_{1j} e^\beta + n_{2j}\}. \quad (3.29)$$

The maximum likelihood estimate of  $\beta$  can be found by maximising this expression with respect to  $\beta$ , for which a non-linear optimisation routine is required. Then, the null hypotheses that  $\beta = 0$  can be tested by comparing the value of  $-2 \log \hat{L}(\beta)$  with  $-2 \log \hat{L}(0)$ . This latter quantity is simply  $2 \sum_{j=1}^r \log n_j$ .

Computation of  $\hat{\beta}$  can be avoided by using a *score test* of the null hypothesis that  $\beta = 0$ . This test procedure, which is outlined in Appendix A, is based on the test statistic

$$\frac{u^2(0)}{i(0)},$$

where

$$u(\beta) = \frac{\partial \log L(\beta)}{\partial \beta}$$

is the efficient score, and

$$i(\beta) = -\frac{\partial^2 \log L(\beta)}{\partial \beta^2}$$

is Fisher's (observed) information function. Under the null hypothesis that  $\beta = 0$ ,  $u^2(0)/i(0)$  has a chi-squared distribution on one degree of freedom.

Now, from equation (3.29),

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{j=1}^r \left( d_{1j} - \frac{n_{1j}e^\beta}{n_{1j}e^\beta + n_{2j}} \right),$$

and

$$\begin{aligned} \frac{\partial^2 \log L(\beta)}{\partial \beta^2} &= - \sum_{j=1}^r \frac{(n_{1j}e^\beta + n_{2j})n_{1j}e^\beta - (n_{1j}e^\beta)^2}{(n_{1j}e^\beta + n_{2j})^2} \\ &= - \sum_{j=1}^r \frac{n_{1j}n_{2j}e^\beta}{(n_{1j}e^\beta + n_{2j})^2}. \end{aligned}$$

The efficient score and information function, evaluated at  $\beta = 0$ , are therefore given by

$$u(0) = \sum_{j=1}^r \left( d_{1j} - \frac{n_{1j}}{n_{1j} + n_{2j}} \right),$$

and

$$i(0) = \sum_{j=1}^r \frac{n_{1j}n_{2j}}{(n_{1j} + n_{2j})^2}.$$

These are simply the expressions for  $U_L$  and  $V_L$  given in equations (2.23) and (2.25) of Chapter 2, for the special case where there are no ties, that is, where  $d_j = 1$  for  $j = 1, 2, \dots, r$ .

When there are tied observations, the likelihood function in equation (3.28) has to be replaced by one that allows for ties. In particular, if the likelihood function in equation (3.11) is used, the efficient score and information function are exactly those given in equations (2.23) and (2.25). Hence, when there are tied survival times, the log-rank test corresponds to using the score test for the discrete proportional hazards model due to Cox (1972). In practice, the  $P$ -value that results from this score test will not usually differ much from that obtained from comparing the values of the statistic  $-2 \log \hat{L}$  for the models with and without a term corresponding to the treatment effect. This was noted in the discussion of Example 3.3. Of course, one advantage of using the Cox regression model in the analysis of such data is that it leads directly to an estimate of the hazard ratio.

### 3.10 Further reading

Comprehensive introductions to statistical modelling in the context of linear regression analysis are given by Draper and Smith (1981) and Montgomery *et al.* (2001). McCullagh and Nelder (1989) include a chapter on models for survival data in their encyclopaedic survey of generalised linear modelling. Aitkin *et al.* (1989) illustrate the theory and practice of linear modelling through the statistical package GLIM, and also include a chapter on the analysis of survival data.

Model formulation and strategies for model selection are discussed in books on linear regression analysis, and also in Chapter 5 of Chatfield (1995), Chapter 4 of Cox and Snell (1981), and Appendix 2 of Cox and Snell (1989). Miller (2002) describes a wide range of procedures for identifying suitable subsets of variables to use in linear regression modelling. What has come to be known as Akaike's information criterion was introduced by Akaike (1974). It is widely used in time series analysis and described in books on this subject, such as Chatfield (1996) and Janacek (2001). The hierarchic principle is fully discussed by Nelder (1977), and in Chapter 3 of McCullagh and Nelder (1989). Harrell (2001) addresses many practical issues in model building and illustrates the process using two extensive case studies involving survival data.

The proportional hazards model for survival data, in which the baseline hazard function remains unspecified, was proposed by Cox (1972). This paper introduced the notion of partial likelihood, which was subsequently considered in greater detail by Cox (1975). See also the contributions to the discussion of Cox (1972) by Kalbfleisch and Prentice (1972) and Breslow (1972). A detailed review of the model, and extensions of it, is contained in Therneau and Grambsch (2000).

Introductions to the proportional hazards model, intended for medical researchers have been given by Christensen (1987), Elashoff (1983) and Tibshirani (1982). More recent accounts are given in the textbooks referenced in Section 1.4 of Chapter 1. In particular, Hosmer and Lemeshow (1999) include a careful discussion on model development and the interpretation of model-based parameter estimates.

A detailed treatment of ties in survival data is given in Kalbfleisch and Prentice (2002) and Lawless (2002); see also Breslow (1972) and Peto (1972). DeLong *et al.* (1994) give an equivalent expression for the exact partial likelihood in the presence of ties that has computational advantages. The estimate of the baseline survivor function, denoted by  $\hat{S}_0(t)$  in Section 3.8, was introduced by Kalbfleisch and Prentice (1973) and is also described in Kalbfleisch and Prentice (2002). The estimate  $S_0^*(t)$  was presented by Breslow (1972, 1974), although it was derived using a different argument from that used in Section 3.8.2.

## 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 lead 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,  $X_1, X_2, \dots, X_p$ . The fitted hazard function for the  $i$ th individual,  $i = 1, 2, \dots, n$ , is therefore

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

where  $\hat{\beta}' \mathbf{x}_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_p x_{pi}$  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  $i$ th individual,  $i = 1, 2, \dots, n$ , is given by

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

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  $i$ th 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

$$f_Y(y) = f_X\{g^{-1}(y)\} / \left| \frac{dy}{dx} \right|,$$

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

$$f_Y(y) = f_T\{S^{-1}(e^{-y})\} / \left| \frac{dy}{dt} \right|, \quad (4.2)$$

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

$$\frac{dy}{dt} = \frac{d\{-\log S(t)\}}{dt} = \frac{f_T(t)}{S(t)},$$

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

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

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  $i$ th 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,  $-\log \hat{S}_i(t_i)$ ,  $i = 1, 2, \dots, n$ , 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  $i$ th 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  $i$ th 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  $i$ th 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  $i$ th individual, the greater the value of the Cox-Snell residual for that individual. It then follows that the residual for the  $i$ th 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  $\Delta$ , 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  $i$ th observation, defined in equation (4.1). It now remains to identify a suitable value for  $\Delta$ . 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  $\lambda^{-1}$ , and consider the probability that  $T$  exceeds  $t_0 + t_1$ ,  $t_1 \geq 0$ , 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

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

The numerator of this expression is simply  $P(T \geq t_0 + t_1)$ , 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

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

which is the survivor function of an exponential random variable at time  $t_1$ , that is  $P(T \geq t_1)$ . 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  $\lambda^{-1}$ . 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,  $\Delta$ , will also have a unit exponential distribution. The expected value of  $\Delta$  is therefore unity, suggesting that  $\Delta$  may be taken to be unity, and this leads to modified Cox-Snell residuals, given by

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

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

Then the modified Cox-Snell residual is given by

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

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

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

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

$$r_{Mi} = \delta_i - r_{Ci}. \quad (4.6)$$

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  $i$ th 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  $\delta_i$ . 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  $i$ th 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

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

where  $r_{Mi}$  is the martingale residual for the  $i$ th 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  $\beta$ -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  $i$ th partial or Schoenfeld residual for  $X_j$ , the  $j$ th explanatory variable in the model, is given by

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

where  $x_{ji}$  is the value of the  $j$ th explanatory variable,  $j = 1, 2, \dots, p$ , for the  $i$ th individual in the study,

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

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  $i$ th Schoenfeld residual, for the explanatory variable  $X_j$ , is an estimate of the  $i$ th component of the first derivative of the logarithm of the partial

likelihood function with respect to  $\beta_j$ , which, from equation (3.5), is given by

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

where

$$a_{ji} = \frac{\sum_l x_{jl} \exp(\beta' x_l)}{\sum_l \exp(\beta' x_l)} \quad (4.11)$$

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

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

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  $i$ th individual be denoted  $\mathbf{r}_{Pi} = (r_{P1i}, r_{P2i}, \dots, r_{Ppi})'$ . The scaled, or weighted, Schoenfeld residuals,  $r_{Pji}^*$ , are then the components of the vector

$$\mathbf{r}_{Pj}^* = r \text{ var}(\hat{\beta}) \mathbf{r}_{Pi},$$

where  $r$  is the number of deaths among the  $n$  individuals, and  $\text{var}(\hat{\beta})$  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  $\beta_j$ ,  $j = 1, 2, \dots, p$ . However, the derivative in equation (4.10) is now expressed in a quite different form, namely

$$\frac{\partial \log L(\beta)}{\partial \beta_j} = \sum_{i=1}^n \left\{ \delta_i (x_{ji} - a_{ji}) + \exp(\beta' x_i) \sum_{t_r \leq t_i} \frac{(a_{jr} - x_{ji}) \delta_r}{\sum_{l \in R(t_r)} \exp(\beta' x_l)} \right\}, \quad (4.12)$$

where  $x_{ji}$  is the  $i$ th value of the  $j$ th explanatory variable,  $\delta_i$  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  $i$ th 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  $i$ th 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  $\beta_j$ , is the efficient score for  $\beta_j$ , and so these residuals are known as *score residuals*.

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

$$r_{Sji} = \delta_i (x_{ji} - \hat{a}_{ji}) + \exp(\hat{\beta}' x_i) \sum_{t_r \leq t_i} \frac{(\hat{a}_{jr} - x_{ji}) \delta_r}{\sum_{l \in R(t_r)} \exp(\hat{\beta}' x_l)}.$$

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

$$r_{Sji} = r_{Pji} + \exp(\hat{\beta}' x_i) \sum_{t_r \leq t_i} \frac{(\hat{a}_{jr} - x_{ji}) \delta_r}{\sum_{l \in R(t_r)} \exp(\hat{\beta}' x_l)}, \quad (4.13)$$

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  $i$ th patient,  $i = 1, 2, \dots, 13$ , is found to be

$$\hat{h}_i(t) = \exp\{0.030 \text{Age}_i - 2.711 \text{Sex}_i\} \hat{h}_0(t), \quad (4.14)$$

where  $\text{Age}_i$  and  $\text{Sex}_i$  refer to the age and sex of the  $i$ th patient.

The variable  $\text{Sex}$  is certainly important, since when  $\text{Sex}$  is added to the model that contains  $\text{Age}$  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  $\text{Age}$  in the model, since the change in the value of the statistic  $-2 \log \hat{L}$  on adding  $\text{Age}$  to the model that contains  $\text{Sex}$  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  $\text{Age}$  and  $\text{Sex}$ , 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  $\text{Sex}$ . The reason for this is that for infection times greater than 30 days, the value of the variable  $\text{Sex}$  is always equal to 2. This means that the value of the term  $\hat{\alpha}_{ji}$  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  $\beta$ '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

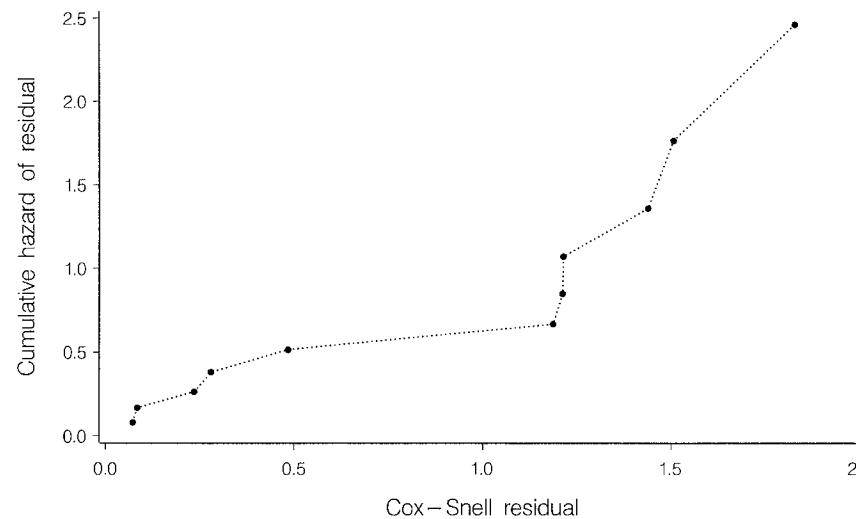


Figure 4.1 Cumulative hazard plot of the Cox-Snell residuals.

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  $i$ th 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  $k$ th uncensored survival time,  $k = 1, 2, \dots, r - 1$ , 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  $k$ th 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  $i$ th 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.

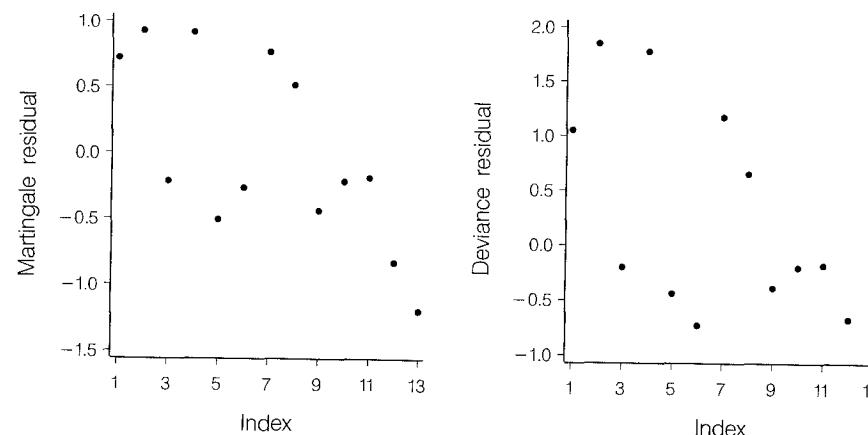
Since the deviance residuals are more symmetrically distributed than the martingale residuals, plots based on these residuals tend to be easier to interpret. Consequently, an index plot of the deviance residuals may also be used to identify individuals whose survival times are out of line.

In a fitted Cox regression model, the hazard of death for the  $i$ th individual at any time depends on the values of explanatory variables for that individual,  $\mathbf{x}_i$ , through the function  $\exp(\hat{\beta}' \mathbf{x}_i)$ . This means that individuals for whom  $\hat{\beta}' \mathbf{x}_i$  has large negative values have a lower than average risk of death, and individuals for whom  $\hat{\beta}' \mathbf{x}_i$  has a large positive value have a higher than average risk. The quantity  $\hat{\beta}' \mathbf{x}_i$  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 particularly helpful diagnostic.

#### Example 4.3 Infection in patients on dialysis

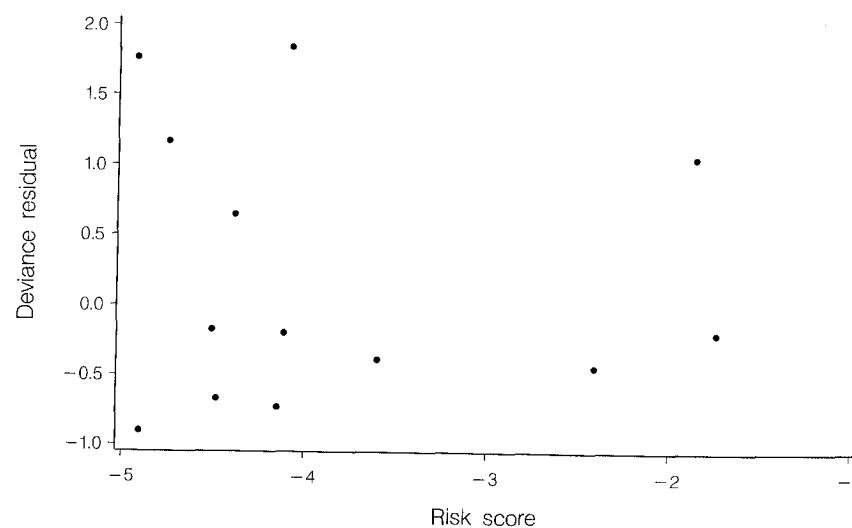
Consider again the data on times to infection in kidney patients. From the values of the martingale and deviance residuals given in Table 4.2, we see that patient 2 has the largest positive residual, suggesting that the time to removal of the catheter is shorter for this patient than might have been 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  $0.030 \text{Age}_i - 2.711 \text{Sex}_i$ , for  $i = 1, 2, \dots, 13$ .



**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 be the case 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.

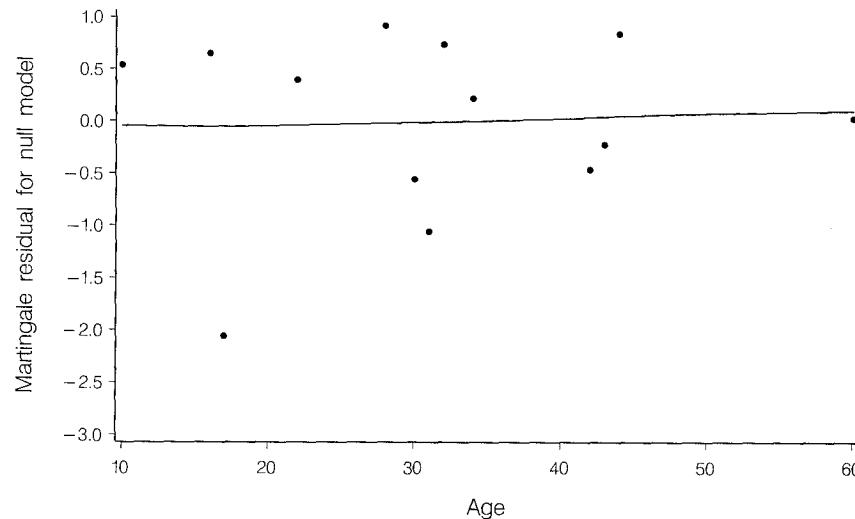
The graphs obtained in this way are usually quite “noisy” and their interpretation is much helped by superimposing a smoothed curve that is fitted to the scatterplot. There are a number of such *smoothers* that can be obtained, including *smoothing splines*, but the one that is most commonly used is 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.

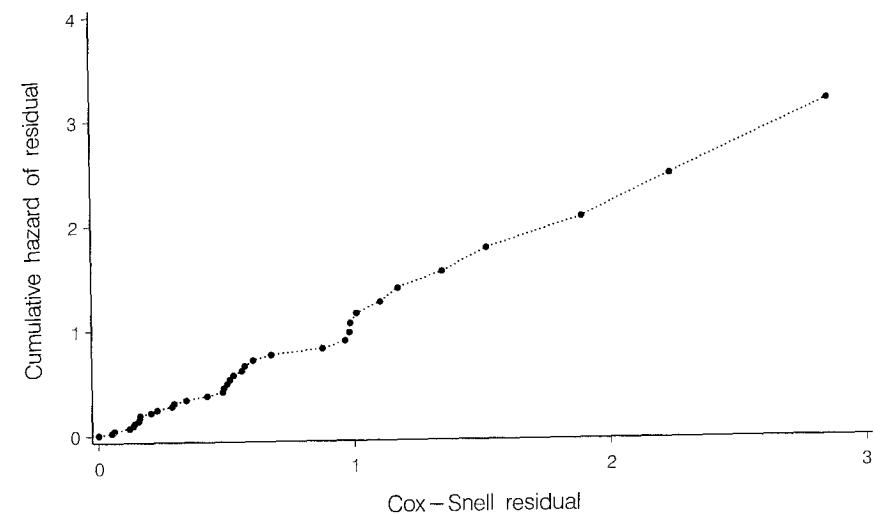
We end this section with a further illustrative example.

#### *Example 4.5 Survival of multiple myeloma patients*

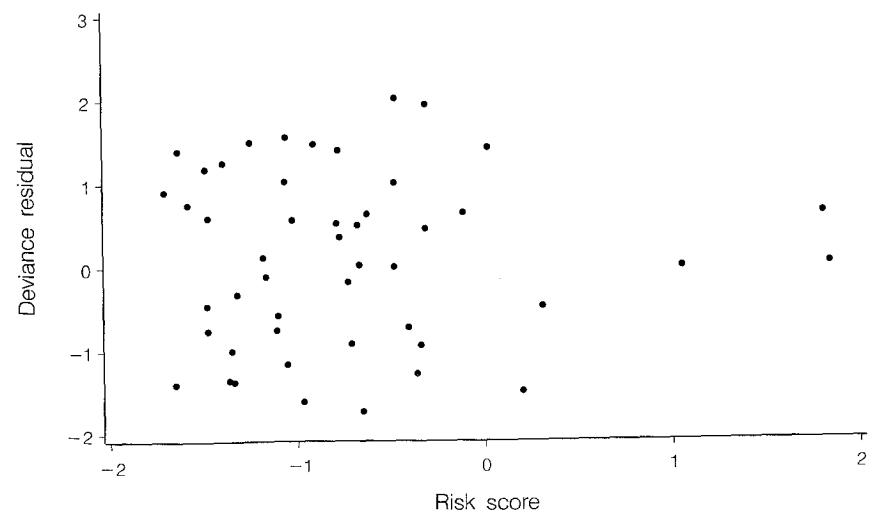
In this example we return to the data on the survival times of 48 patients with multiple myeloma, described in Example 1.3. In Example 3.5, a Cox regression model that contained the explanatory variables *Hb* (serum haemoglobin) and *Bun* (blood urea nitrogen) was found to be a suitable model for the 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.

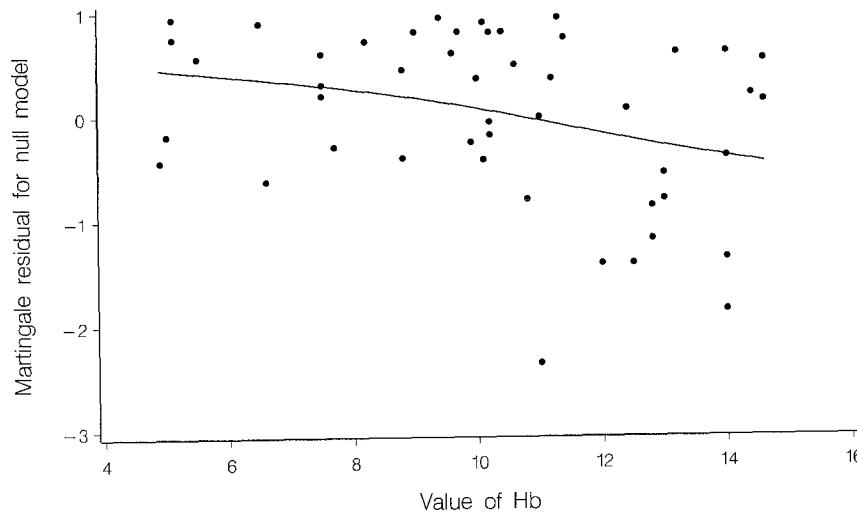


**Figure 4.5** Log-cumulative hazard plot of the Cox-Snell residuals.

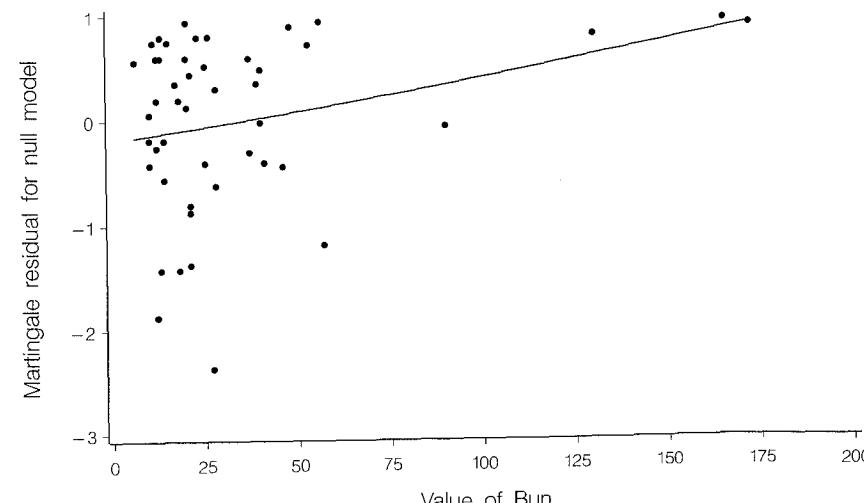


**Figure 4.6** Deviance residuals plotted against the risk score.

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.



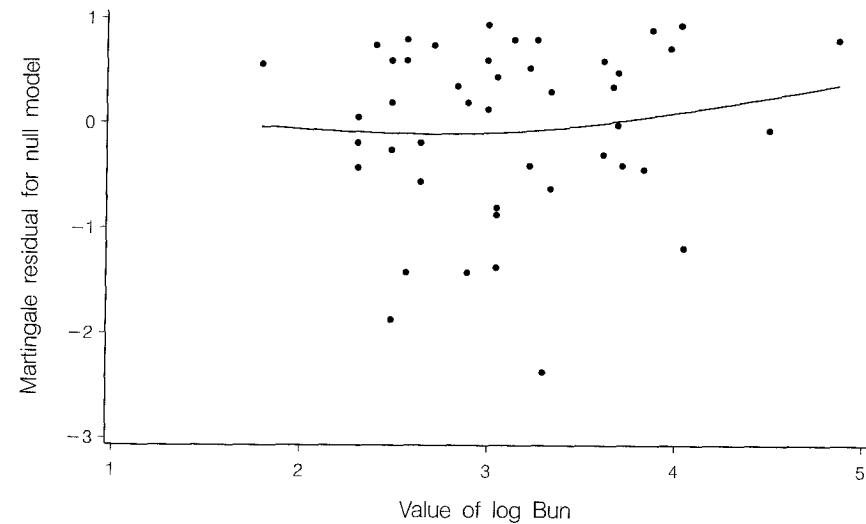
**Figure 4.7** Plot of the martingale residuals for the null model against the values of  $Hb$ , with a smoothed curve superimposed.



**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  $\beta$ -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  $\exp(\beta' \mathbf{x})$  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  $j$ th parameter estimate,  $j = 1, 2, \dots, p$ , 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  $i$ th observation is omitted, for  $i = 1, 2, \dots, n$ . Suppose that the value of the  $j$ th parameter estimate on omitting the  $i$ th 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  $\mathbf{r}'_{Si}$  denote the vector of values of the score residuals for the  $i$ th observation, so that  $\mathbf{r}'_{Si} = (r_{S1i}, r_{S2i}, \dots, r_{Sp_i})$ , where  $r_{Sji}$ ,  $j = 1, 2, \dots, p$ , is the  $i$ th 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  $i$ th observation, is then the  $j$ th component of the vector

$$\mathbf{r}'_{Si} \text{ var}(\hat{\beta}),$$

$\text{var}(\hat{\beta})$  being the variance-covariance matrix of the vector of parameter estimates in the fitted Cox regression model. The  $j$ th 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  $j$ th 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  $\hat{\beta}/\text{se}(\hat{\beta})$ , on omitting the  $i$ th 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  $i$ th 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  $i$ th 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,  $\Delta_i\hat{\beta}_1$  and  $\Delta_i\hat{\beta}_2$ , are obtained. These are given in Table 4.4.

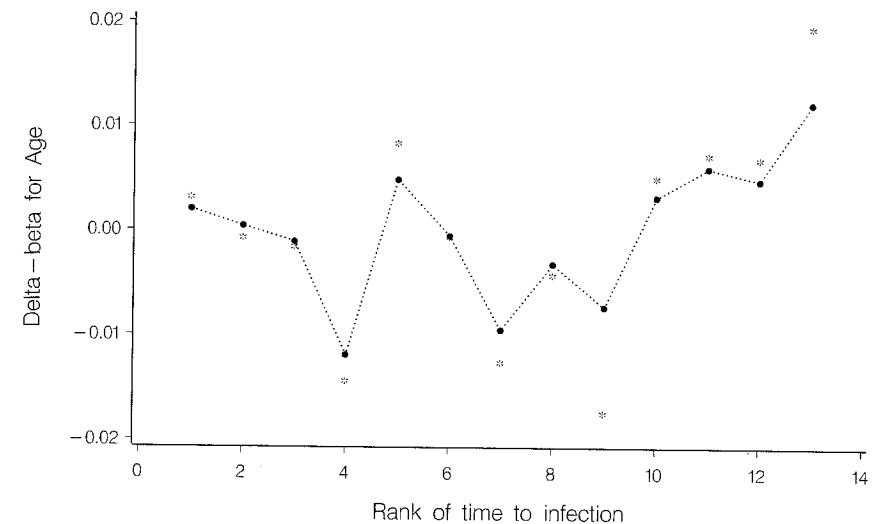
**Table 4.4** Approximate delta-betas for Age ( $\hat{\beta}_1$ ), and Sex ( $\hat{\beta}_2$ ).

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

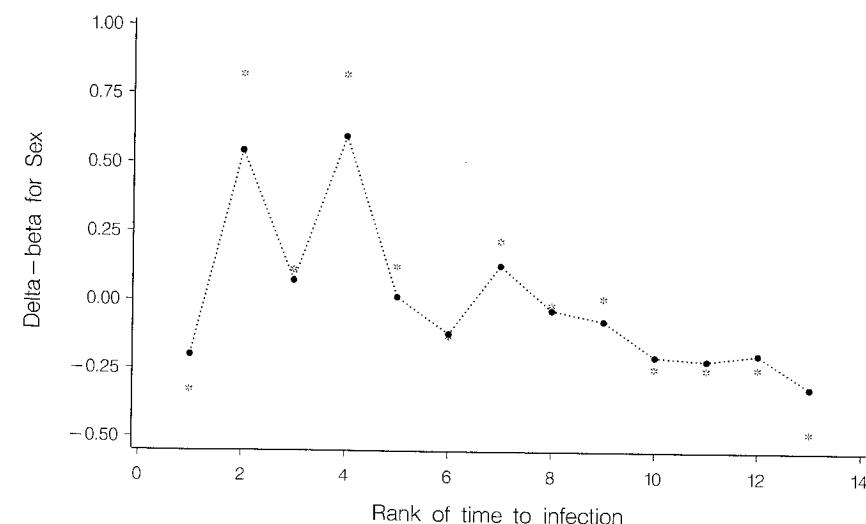
The largest delta-beta for *Age* occurs for patient number 13, but there are other delta-betas with similar values. The actual change in the parameter estimate on omitting the data for this patient is 0.0195, and so omission of this observation increases the hazard of infection relative to the baseline 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.

occurs for patient number 8. That for *Sex* is 0.276, which occurs for patient number 2. These differences are about a quarter of the value of the standard error of each parameter estimate.

#### 4.3.2 Influence of observations on the set of parameter estimates

It may happen that the structure of the fitted model is particularly sensitive to one or more observations in the data set. Such observations can be detected using diagnostics that are designed to highlight observations that influence the complete set of parameter estimates in the linear predictor. These diagnostics therefore reflect the influence that individual observations have on the risk score, and give information that is additional to that provided by the delta-betas. In particular, excluding a given observation from the data set may not have a great influence on any particular parameter estimate, and so will not be revealed from a study of the delta-beta statistics. However, the change in the set of parameter estimates might be such that the form of the estimated hazard function, or values of summary statistics based on the fitted model, change markedly when that observation is removed. Statistics for assessing the influence of observations on the set of parameter estimates also have the advantage that there is a single value of the diagnostic for each observation. This makes them easier to use than diagnostics such as the delta-betas.

A number of diagnostics for assessing the influence of each observation on the set of parameter estimates have been proposed. In this section, two will be described, but references to others will be given in the concluding section of this chapter.

One way of assessing the influence of each observation on the overall fit of the model is to examine the amount by which the value of minus twice the logarithm of the maximised partial likelihood,  $-2 \log \hat{L}$ , under a fitted model, changes when each observation in turn is left out. Write  $-2 \log L(\hat{\beta})$  for the value of the maximised log-likelihood when the model is fitted to all  $n$  observations, and  $-2 \log L(\hat{\beta}_{(i)})$  for the value of the maximised log-likelihood of the  $n$  observations when the parameter estimates are computed after omitting the  $i$ th observation from the fit. The diagnostic

$$2 \left\{ \log L(\hat{\beta}) - \log L(\hat{\beta}_{(i)}) \right\}$$

can then be useful in the study of influence.

Pettitt and Bin Daud (1989) show that an approximation to this *likelihood displacement* is

$$LD_i = \mathbf{r}'_{Si} \text{var}(\hat{\beta}) \mathbf{r}_{Si}, \quad (4.15)$$

where  $\mathbf{r}_{Si}$  is the  $p \times 1$  vector of score residuals, whose  $j$ th component is given in equation (4.13), and  $\text{var}(\hat{\beta})$  is the variance-covariance matrix of  $\hat{\beta}$ , the vector of parameter estimates. The values of this statistic may therefore be straightforwardly obtained from terms used in computing the delta-betas for each explanatory variable in the model. An index plot, or a plot of the likelihood displacements against the rank order of the survival times, provides

an informative visual summary of the values of the diagnostic. Observations that have relatively large values of the diagnostic are influential. Plots against explanatory variables are not recommended, since, as demonstrated by Pettitt and Bin Daud (1989), these plots can have a deterministic pattern, even when the fitted model is correct.

Another diagnostic that can be used to assess the impact of each observation on the set of parameter estimates is based on the  $n \times n$  symmetric matrix

$$\mathbf{B} = \Theta' \text{var}(\hat{\beta}) \Theta,$$

where  $\Theta'$  is the  $n \times p$  matrix formed from the vectors  $\mathbf{r}_{Si}$ , for  $i = 1, 2, \dots, p$ . An argument from linear algebra shows that the absolute values of the elements of the  $n \times 1$  eigenvector associated with the largest eigenvalue of the matrix  $\mathbf{B}$ , standardised to have unit length by dividing each component by the square root of the sum of squares of all the components of the eigenvector, is a measure of the sensitivity of the fit of the model to each of the  $n$  observations in the data set. Denoting this eigenvector by  $\mathbf{l}_{\max}$ , the  $i$ th element of  $\mathbf{l}_{\max}$  is a measure of the influence of the  $i$ th observation on the set of parameter estimates. The sign of this diagnostic is immaterial, and so plots based on the absolute values,  $|\mathbf{l}_{\max}|$  are commended for general use. Index plots of these values, plots against the rank order of the survival times, and against explanatory variables in the model, can all be useful in the assessment of influence.

The standardisation to unit length means that the squares of the elements of  $\mathbf{l}_{\max}$  must sum to 1.0. Observations for which the squares of the elements of the eigenvector account for a substantial proportion of the total sum of squares of unity will then be those that are most influential. Large elements of this eigenvector will therefore correspond to observations that have most effect on the value of the likelihood function. A final point to note is that unlike other diagnostics, a plot of the elements of  $\mathbf{l}_{\max}$  against explanatory variables will not have a deterministic pattern if the fitted model is correct. This means that plots of the absolute values of the elements of  $\mathbf{l}_{\max}$  against explanatory variables can be useful in assessing whether there are particular ranges of values of the variates over which the model does not fit well.

#### Example 4.7 Infection in patients on dialysis

The data first given in Example 4.1 will again be used to illustrate the use of diagnostics designed to reveal observations that influence the complete set of parameter estimates. In Table 4.5, the approximate likelihood displacements from equation (4.15), and the elements of the vector  $|\mathbf{l}_{\max}|$ , are given.

The observations that most affect the value of the maximised log-likelihood when they are omitted are those corresponding to patients 2 and 4. The value of the likelihood displacement diagnostic is also quite large for patient number 13. This means that the set of parameter estimates are most affected by the removal of either of these three patients from the data base.

The fourth element of  $\mathbf{l}_{\max}$  is the largest in absolute value, and indicates that omitting the data from patient number 4 has the greatest effect on the

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

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

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

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

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

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

$$\text{Omitting patient number 13: } 0.011 Age_i - 2.234 Sex_i.$$

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

$$0.030 Age_i - 2.711 Sex_i.$$

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

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

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

#### 4.3.3 Treatment of influential observations

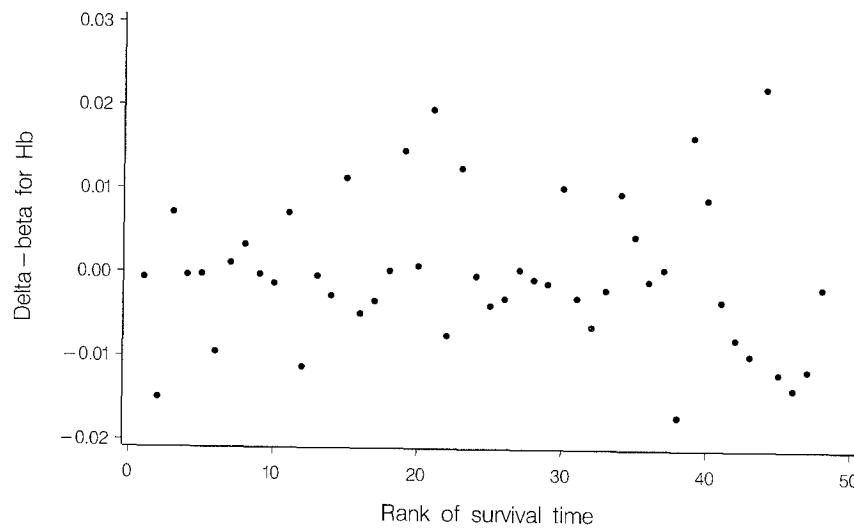
Once observations have been found to be unduly influential, it is difficult to offer any firm advice on what should be done about them. So much depends on the scientific background to the study.

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

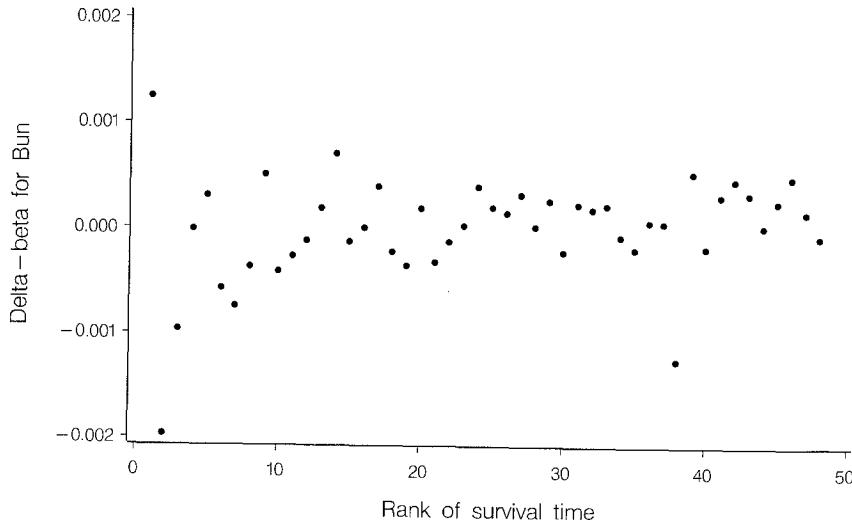
In many situations it will not be possible to confirm that the data corresponding to an influential observation are valid. Certainly, influential observations should not then be rejected outright. In these circumstances, the most appropriate course of action will be to establish the actual effect on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

#### Example 4.8 Survival of multiple myeloma patients

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



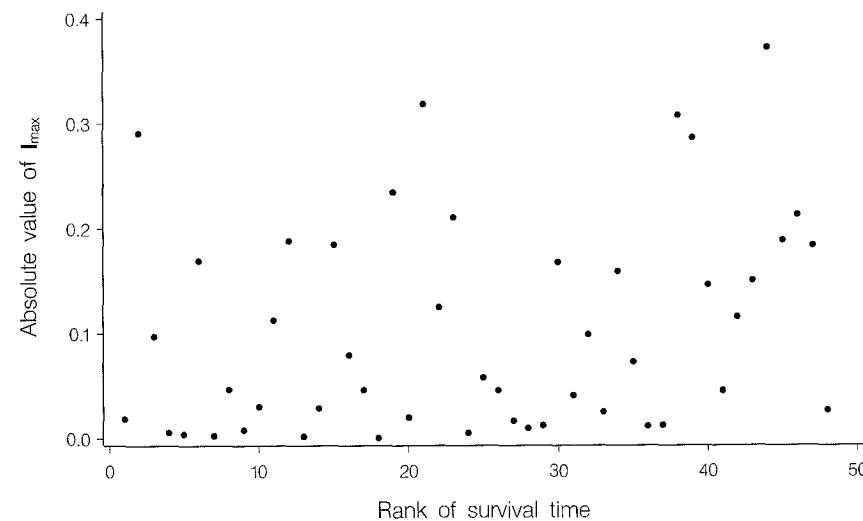
**Figure 4.12** Plot of the delta-betas for  $Hb$  against rank order of survival time.



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

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

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



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

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

#### 4.4 Testing the assumption of proportional hazards

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

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

#### 4.4.1 The log-cumulative hazard plot

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

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

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

$$\int_0^t h_i(u) du = \exp(\beta' \mathbf{x}_i) \int_0^t h_0(u) du,$$

and so, using equation (1.6),

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

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

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

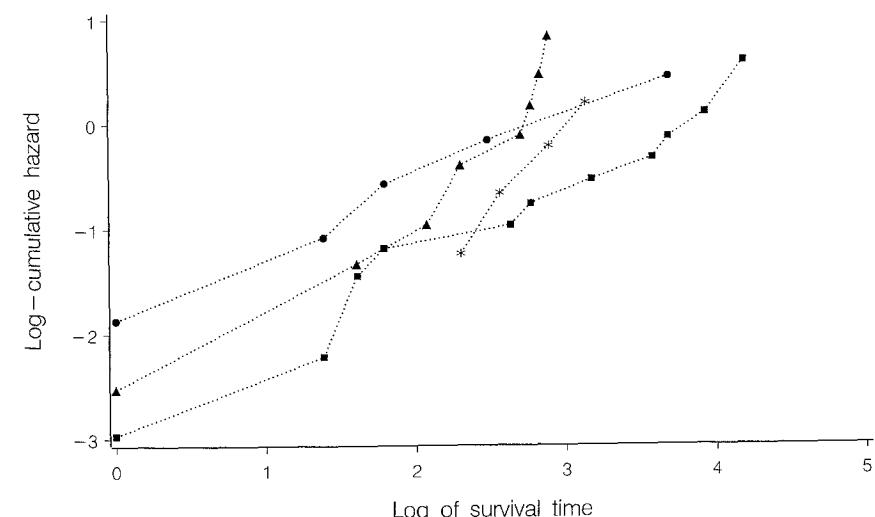
from which it follows that differences in the log-cumulative hazard functions do not depend on time. This means that if the log-cumulative hazard functions for individuals with different values of their explanatory variables are plotted against time, the curves so formed will be parallel if the proportional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

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

each level. On the other hand, the plot will be based on very few observations at the later survival times, and in more highly structured data sets, a different approach needs to be taken.

#### Example 4.9 Survival of multiple myeloma patients

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



**Figure 4.15** Log-cumulative hazard plot for multiple myeloma patients in four groups defined by  $Hb \leq 7$  ( $\bullet$ ),  $7 < Hb \leq 10$  ( $\blacksquare$ ),  $10 < Hb \leq 13$  ( $\blacktriangle$ ) and  $Hb > 13$  ( $*$ ).

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

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

#### 4.4.2\* Use of Schoenfeld residuals

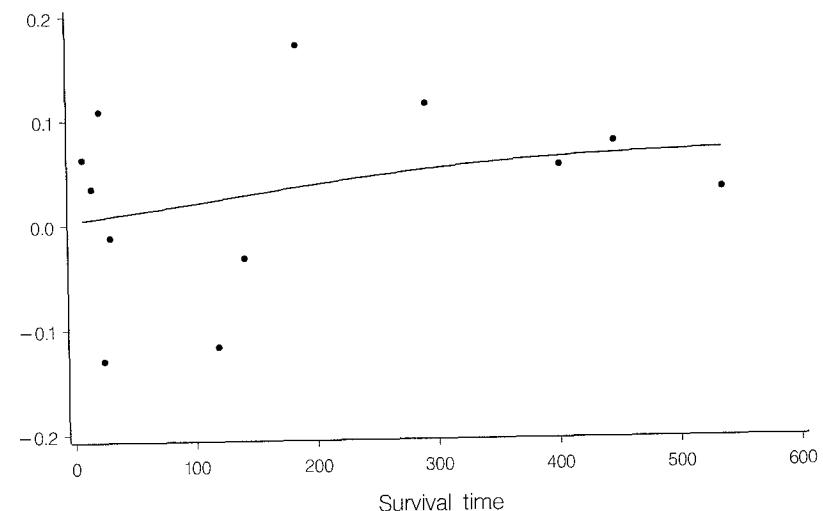
Hazards are said to be proportional if ratios of hazards are independent of time. If there are one or more explanatory variables in the model whose coefficients vary with time, or if there are explanatory variables that are time-dependent, the proportional hazards assumption will be violated. We therefore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

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

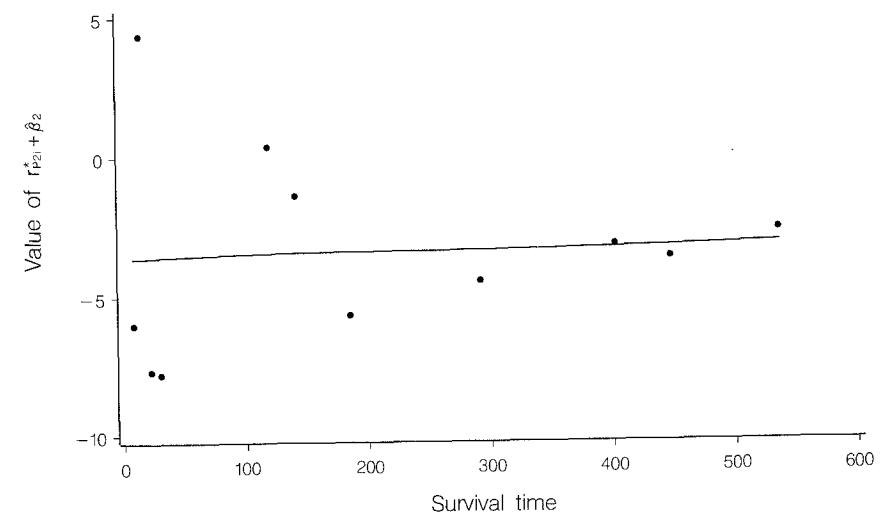
#### Example 4.10 Infection in patients on dialysis

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

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



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



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

#### 4.4.3\* Adding a time-dependent variable

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

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

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

where  $x_{1i}$  is the value of an indicator variable  $X_1$  that is zero for the standard treatment and unity for the new treatment. The relative hazard of death at any time for a patient on the new treatment, relative to one on the standard, is then  $e^{\beta_1}$ , which is independent of the survival time.

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

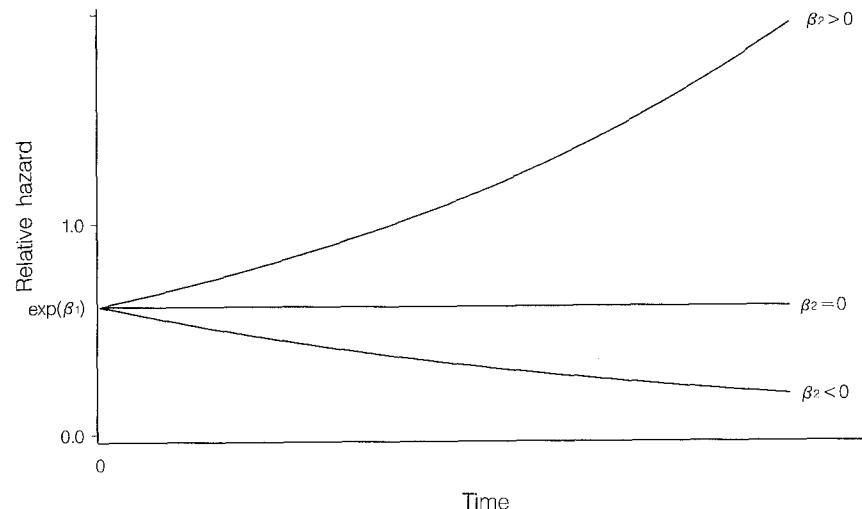
$$h_i(t) = \exp(\beta_1 x_{1i} + \beta_2 x_{2i}) h_0(t), \quad (4.18)$$

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

$$\exp(\beta_1 + \beta_2 t), \quad (4.19)$$

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

In order to aid in both the computation and interpretation of the parameters in the model of equation (4.18), the variable  $X_2$  can be defined in terms of the deviation from some time,  $t_0$ . The estimated values of  $\beta_1$  and  $\beta_2$  will then tend to be less highly correlated, and maximisation of the appropriate likelihood function will be less difficult. If  $X_2$  is taken to be such that  $X_2 = X_1(t - t_0)$ , the value of  $X_2$  is  $t - t_0$  for the new treatment and zero for the standard. The



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

relative hazard now becomes

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

In the model of equation (4.18), the quantity  $e^{\beta_1}$  is the hazard of death at time  $t_0$  for an individual on the new treatment relative to one on the standard. In practical applications,  $t_0$  will generally be chosen to provide a convenient interpretation for the time at which this relative hazard is applicable. For example, taking  $t_0$  to be the mean or median survival time means that  $\exp(\hat{\beta}_1)$  is the estimated relative hazard of death at this time.

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

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

where  $x_{2i}$  is the value of  $X$  for that individual. The hazard of death at time  $t$  for an individual for whom  $X = x_i + 1$ , relative to an individual for whom  $X = x_i$ , is then  $\exp(\beta_1 + \beta_2 t)$ , as in equation (4.19).

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

hypothesis is that the hazard ratio is dependent on the logarithm of time. Using  $\log t$  in the definition of a time-dependent variable is also helpful when the numerical values of the survival times are large, such as when survival in a long-term study is measured in days. There may then be computational problems associated with calculating the value of  $\exp(\beta_2 x_2)$  in equation (4.18), which are resolved by using  $\log t$  in place of  $t$  in the definition of  $X_2$ .

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

#### *Example 4.11 Infection in patients on dialysis*

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

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

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

#### 4.5 Recommendations

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

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

residuals against the survival times, the rank order of the survival times, or explanatory variables may also be useful.

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

#### 4.6 Further reading

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

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

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

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

theory of counting processes, Fleming and Harrington (1991) and Therneau and Grambsch (2000) show how different types of residual can be used, and give detailed practical examples. Two other types of residual, introduced by Nardi and Schemper (1999), are particularly suitable for the detection of outlying survival times.

Influence diagnostics for the Cox regression model have been considered by many authors, but the major papers are those of Cain and Lange (1984), Reid and Crépeau (1985), Storer and Crowley (1985), Pettitt and Bin Daud (1989) and Weissfeld (1990). Pettitt and Bin Daud (1990) show how time-dependence in the Cox proportional hazards model can be detected by smoothing the Schoenfeld residuals. The LOWESS smoother was introduced by Cleveland (1979), and the algorithm is also presented in Collett (2003).

Some other graphical methods for evaluating survival models, not mentioned in this chapter, have been proposed by Cox (1979) and Arjas (1988). Gray (1990) describes the use of smoothed estimates of cumulative hazard functions in evaluating the fit of a Cox model.

Most of the diagnostic procedures presented in this chapter rely on an informal evaluation of tabular or graphical presentations of particular statistics. In addition to these procedures, a variety of significance tests have been proposed that can be used to assess the goodness of fit of the model. Examples include the methods of Schoenfeld (1980), Andersen (1982), Nagelkerke *et al.* (1984), Ciampi and Etezadi-Amoli (1985), Moreau *et al.* (1985), Gill and Schumacher (1987), O'Quigley and Pessione (1989), Quantin *et al.* (1996), Grønnesby and Borgan (1996), and Verweij *et al.* (1998). Reviews of some of these goodness of fit tests for the Cox regression model are included in Lin and Wei (1991) and Quantin *et al.* (1996). Many of these test involve statistics that are quite complicated, and the procedures are not widely in computer software for survival analysis. A more simple procedure for evaluating the overall fit of a model has been proposed by May and Hosmer (1998).

## CHAPTER 5

# Parametric proportional hazards models

When the Cox regression model is used in the analysis of survival data, there is no need to assume a particular form of probability distribution for the survival times. As a result, the hazard function is not restricted to a specific functional form, and the model has flexibility and widespread applicability. On the other hand, if the assumption of a particular probability distribution for the data is valid, inferences based on such an assumption will be more precise. In particular, estimates of quantities such as relative hazards and median survival times will tend to have smaller standard errors than they would in the absence of a distributional assumption. Models in which a specific probability distribution is assumed for the survival times are known as *parametric models*, and parametric versions of the proportional hazards model, described in Chapter 3, are the subject of this chapter.

A probability distribution that plays a central role in the analysis of survival data is the Weibull distribution, introduced by W. Weibull in 1951 in the context of industrial reliability testing. Indeed, this distribution is as central to the parametric analysis of survival data as the normal distribution is in linear modelling. Proportional hazards models based on the Weibull distribution are therefore considered in some detail.

### 5.1 Models for the hazard function

Once a distributional model for survival times has been specified in terms of a probability density function, the corresponding survivor and hazard functions can be obtained from the relations

$$S(t) = 1 - \int_0^t f(u) du, \quad (5.1)$$

and

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \{\log S(t)\}, \quad (5.2)$$

where  $f(t)$  is the probability density function of the survival times. These relationships were derived in Section 1.3. An alternative approach is to specify a functional form for the hazard function, from which the survivor function and probability density functions can be determined from the equations

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

and

$$f(t) = h(t)S(t) = -\frac{dS(t)}{dt}, \quad (5.4)$$

where

$$H(t) = \int_0^t h(u) du$$

is the integrated hazard function.

### 5.1.1 The exponential distribution

The simplest model for the hazard function is to assume that it is constant over time. The hazard of death at any time after the time origin of the study is then the same, irrespective of the time elapsed. Under this model, the hazard function may be written as

$$h(t) = \lambda,$$

for  $0 \leq t < \infty$ . The parameter  $\lambda$  is a positive constant that would be estimated by fitting the model to the observed data. From equation (5.3), the corresponding survivor function is

$$\begin{aligned} S(t) &= \exp \left\{ - \int_0^t \lambda du \right\}, \\ &= e^{-\lambda t}, \end{aligned} \quad (5.5)$$

and so the implied probability density function of the survival times is

$$f(t) = \lambda e^{-\lambda t}, \quad (5.6)$$

for  $0 \leq t < \infty$ . This is the probability density function of a random variable  $T$  that has an *exponential distribution* with a mean of  $\lambda^{-1}$ . It is sometimes convenient to write  $\mu = \lambda^{-1}$ , so that the hazard function is  $\mu^{-1}$ , and the survival time distribution has a mean of  $\mu$ . However, the former specification of the hazard function will generally be used in this book.

The median of the exponential distribution,  $t(50)$ , is such that  $S\{t(50)\} = 0.5$ , that is,

$$\exp\{-\lambda t(50)\} = 0.5,$$

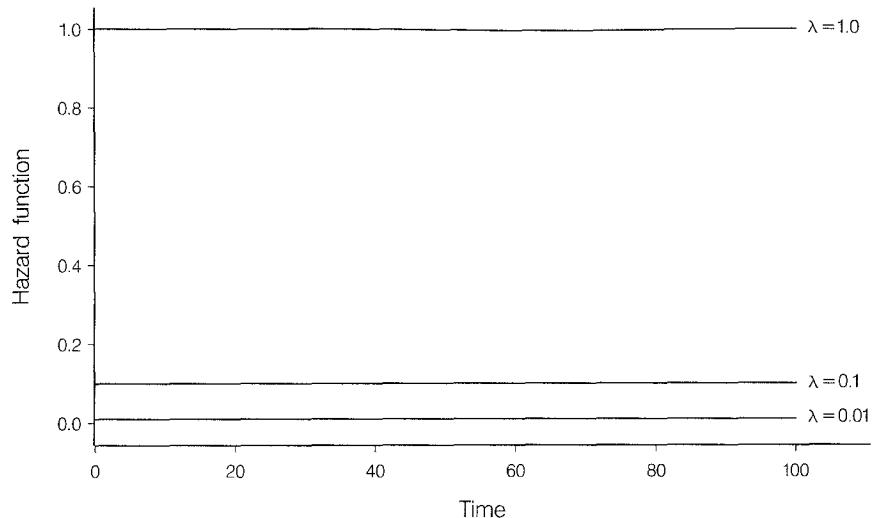
so that

$$t(50) = \frac{1}{\lambda} \log 2.$$

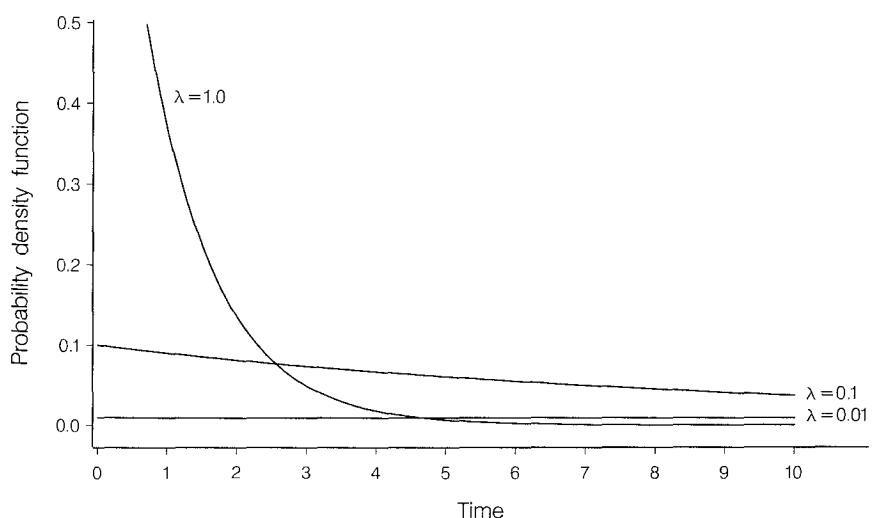
More generally, the  $p$ th percentile of the survival time distribution is the value  $t(p)$  such that  $S\{t(p)\} = 1 - (p/100)$ , and using equation (5.5), this is

$$t(p) = \frac{1}{\lambda} \log \left( \frac{100}{100-p} \right).$$

A plot of the hazard function for three values of  $\lambda$ , namely 1.0, 0.1 and 0.01, is given in Figure 5.1, and the corresponding probability density functions are shown in Figure 5.2. For these values of  $\lambda$ , the means of the corresponding exponential distributions are 1, 10 and 100, and the median survival times are 0.69, 6.93 and 69.31, respectively.



**Figure 5.1** Hazard functions for exponential distributions with  $\lambda = 1.0$ , 0.1 and 0.01.



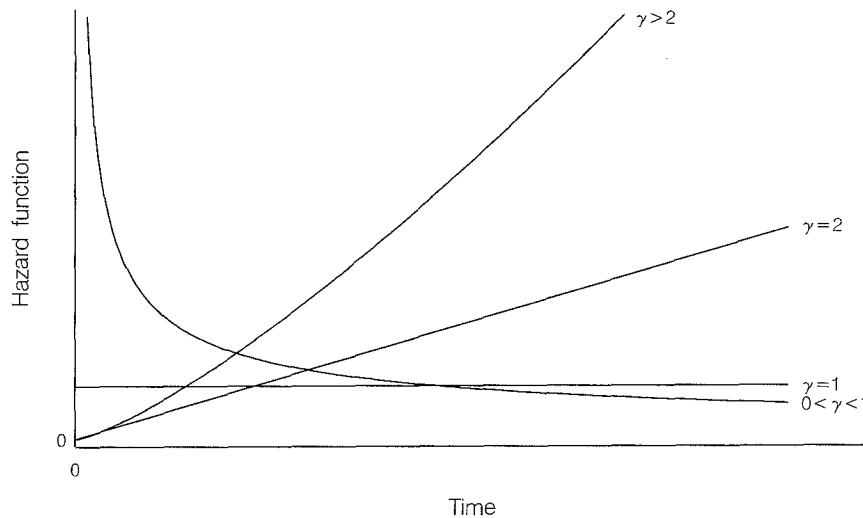
**Figure 5.2** Probability density functions for exponential distributions with  $\lambda = 1.0$ , 0.1 and 0.01.

### 5.1.2 The Weibull distribution

In practice, the assumption of a constant hazard function, or equivalently of exponentially distributed survival times, is rarely tenable. A more general form of hazard function is such that

$$h(t) = \lambda\gamma t^{\gamma-1}, \quad (5.7)$$

for  $0 \leq t < \infty$ , a function that depends on two parameters  $\lambda$  and  $\gamma$ , which are both greater than zero. In the particular case where  $\gamma = 1$ , the hazard function takes a constant value  $\lambda$ , and the survival times have an exponential distribution. For other values of  $\gamma$ , the hazard function increases or decreases monotonically, that is, it does not change direction. The shape of the hazard function depends critically on the value of  $\gamma$ , and so  $\gamma$  is known as the *shape parameter*, while the parameter  $\lambda$  is a *scale parameter*. The general form of this hazard function for different values of  $\gamma$  is shown in Figure 5.3.



**Figure 5.3** The form of the Weibull hazard function,  $h(t) = \lambda\gamma t^{\gamma-1}$ , for different values of  $\gamma$ .

For this particular choice of hazard function, the survivor function is given by

$$S(t) = \exp \left\{ - \int_0^t \lambda\gamma u^{\gamma-1} du \right\} = \exp(-\lambda t^\gamma). \quad (5.8)$$

The corresponding probability density function is then

$$f(t) = \lambda\gamma t^{\gamma-1} \exp(-\lambda t^\gamma),$$

for  $0 \leq t < \infty$ , which is the density of a random variable that has a *Weibull distribution* with scale parameter  $\lambda$  and shape parameter  $\gamma$ . This distribution will be denoted  $W(\lambda, \gamma)$ . The right-hand tail of this distribution is longer than the left-hand one, and so the distribution is positively skewed.

### ASSESSING THE SUITABILITY OF A PARAMETRIC MODEL

The mean, or expected value, of a random variable  $T$  that has a  $W(\lambda, \gamma)$  distribution can be shown to be given by

$$E(T) = \lambda^{-1/\gamma} \Gamma(\gamma^{-1} + 1),$$

where  $\Gamma(x)$  is the gamma function defined by the integral

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du.$$

The value of this integral is  $(x-1)!$ , and so for integer values of  $x$  it can easily be calculated. For non-integer values of  $x$ , tables of the gamma function, such as those in Abramowitz and Stegun (1972), or suitable computer software, will be needed to compute the mean. However, since the Weibull distribution is skewed, a more appropriate, and more tractable, summary of the location of the distribution is the median survival time. This is the value  $t(50)$  such that  $S\{t(50)\} = 0.5$ , so that

$$\exp\{-\lambda[t(50)]^\gamma\} = 0.5,$$

and

$$t(50) = \left\{ \frac{1}{\lambda} \log 2 \right\}^{1/\gamma}.$$

More generally, the  $p$ th percentile of the Weibull distribution,  $t(p)$ , is such that

$$t(p) = \left\{ \frac{1}{\lambda} \log \left( \frac{100}{100-p} \right) \right\}^{1/\gamma}. \quad (5.9)$$

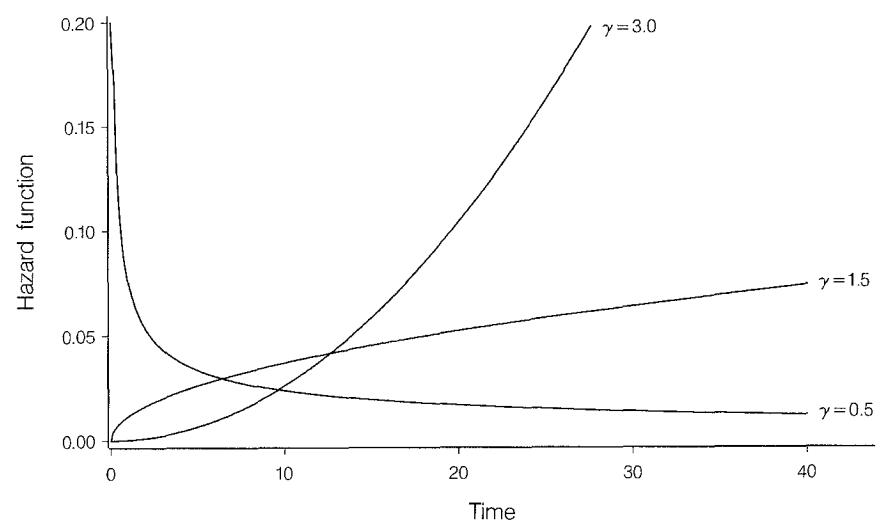
The median and other percentiles of the Weibull distribution are therefore much simpler to compute than the mean of the distribution.

The hazard function and corresponding probability density function for Weibull distributions with a median of 20, and shape parameters  $\gamma = 0.5$ , 1.5 and 3.0, are shown in Figures 5.4 and 5.5, respectively. The corresponding value of the scale parameter,  $\lambda$ , for these three Weibull distributions is 0.15, 0.0078 and 0.000087, respectively.

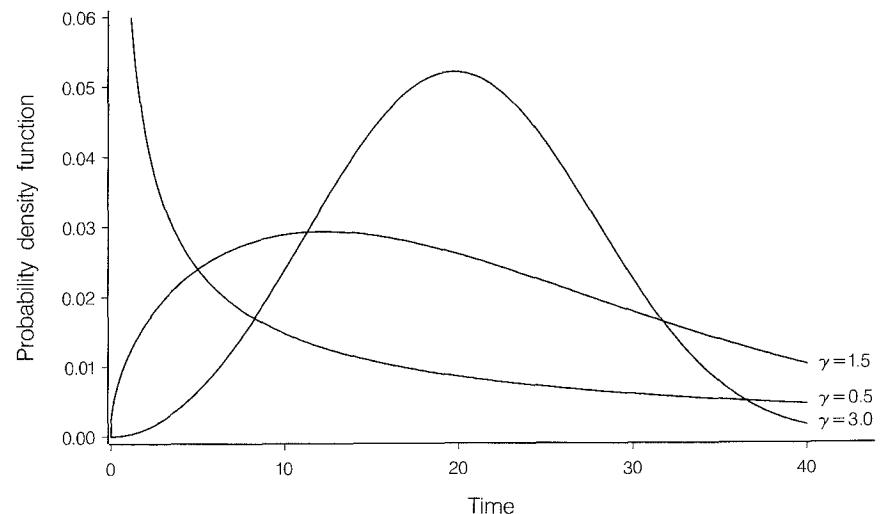
Since the Weibull hazard function can take a variety of forms, depending on the value of the shape parameter,  $\gamma$ , and appropriate summary statistics can be easily obtained, this distribution is widely used in the parametric analysis of survival data.

### 5.2 Assessing the suitability of a parametric model

Prior to fitting a model based on an assumed parametric form for the hazard function, a preliminary study of the validity of this assumption should be carried out. One approach would be to estimate the hazard function using the methods outlined in Section 2.3. If the hazard function were reasonably constant over time, this would indicate that the exponential distribution might be a suitable model for the data. On the other hand, if the hazard function increased or decreased monotonically with increasing survival time, a model based on the Weibull distribution would be indicated.



**Figure 5.4** Hazard functions for a Weibull distribution with a median of 20 and  $\gamma = 0.5, 1.5$  and  $3.0$ .



**Figure 5.5** Probability density functions for a Weibull distribution with a median of 20 and  $\gamma = 0.5, 1.5$  and  $3.0$ .

A more informative way of assessing whether a particular distribution for the survival times is plausible is to compare the survivor function for the data with that of a chosen model. This is greatly helped by transforming the survivor function to produce a plot that should give a straight line if the assumed model is appropriate.

Suppose that a single sample of survival data is available, and that a Weibull distribution for the survival times is contemplated. Since the survivor function for a Weibull distribution, with scale parameter  $\lambda$  and shape parameter  $\gamma$ , is given by

$$S(t) = \exp\{-\lambda t^\gamma\},$$

taking the logarithm of  $S(t)$ , multiplying by  $-1$ , and taking logarithms a second time, gives

$$\log\{-\log S(t)\} = \log \lambda + \gamma \log t. \quad (5.10)$$

We now substitute the Kaplan-Meier estimate of the survivor function,  $\hat{S}(t)$ , for  $S(t)$  in equation (5.10). If the Weibull assumption is tenable,  $\hat{S}(t)$  will be “close” to  $S(t)$ , and a plot of  $\log\{-\log \hat{S}(t)\}$  against  $\log t$  would then give an approximately straight line. From equation (1.7), the cumulative hazard function,  $H(t)$ , is  $-\log S(t)$  and so  $\log\{-\log S(t)\}$  is the log-cumulative hazard. A plot of the values of  $\log\{-\log \hat{S}(t)\}$  against  $\log t$  is a log-cumulative hazard plot, introduced in Section 4.4.1 of Chapter 4.

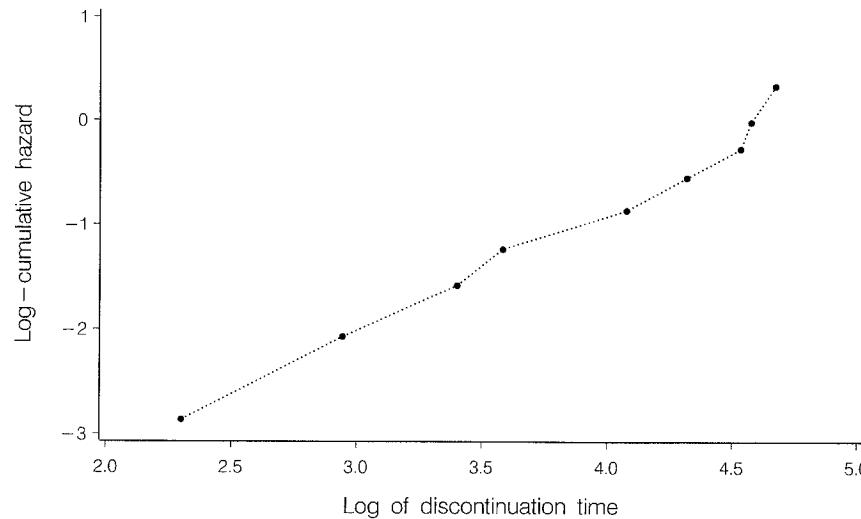
If the log-cumulative hazard plot gives a straight line, the plot can be used to provide a rough estimate of the two parameters of the Weibull distribution. Specifically, from equation (5.10), the intercept and slope of the straight line will be  $\log \lambda$  and  $\gamma$ , respectively. Thus, the slope of the line in a log-cumulative hazard plot gives an estimate of the shape parameter, and the exponent of the intercept provides an estimate of the scale parameter. Note that if the slope of the log-cumulative hazard plot is close to unity, the survival times could have an exponential distribution.

#### *Example 5.1 Time to discontinuation of the use of an IUD*

In Example 2.3, the Kaplan-Meier estimate of the survivor function,  $\hat{S}(t)$ , for the data on the time to discontinuation of an IUD, was obtained. A log-cumulative hazard plot for these data, that is, a plot of  $\log\{-\log \hat{S}(t)\}$  against  $\log t$ , is shown in Figure 5.6.

The plot indicates that there is a straight line relationship between the log-cumulative hazard and  $\log t$ , confirming that the Weibull distribution is an appropriate model for the discontinuation times. From the graph, the intercept of the line is approximately  $-6.0$  and the slope is approximately  $1.25$ . Approximate estimates of the parameters of the Weibull distribution are therefore  $\lambda^* = \exp(-6.0) = 0.002$  and  $\gamma^* = 1.25$ . The estimated value of  $\gamma$ , the shape parameter of the Weibull distribution, is quite close to unity, suggesting that the discontinuation times might be adequately modelled by an exponential distribution.

These informal estimates of  $\lambda$  and  $\gamma$  can be used to estimate the parameters



**Figure 5.6** Log-cumulative hazard plot for the data from Example 1.1.

of the distribution, and hence functions of these estimates, such as the median of the survival time distribution. However, this graphical approach does not lead to a measure of the precision with which the quantities have been estimated. In view of this limitation, a more formal way of fitting parametric models to survival data is developed in the next section.

### 5.3 Fitting a parametric model to a single sample

Parametric models can be fitted to an observed set of survival data using the method of maximum likelihood, outlined in Section 3.3. Consider first the situation where actual survival times have been observed for  $n$  individuals, so that there are no censored observations. If the probability density function of the random variable associated with survival time is  $f(t)$ , the likelihood of the  $n$  observations  $t_1, t_2, \dots, t_n$  is simply the product

$$\prod_{i=1}^n f(t_i).$$

This likelihood will be a function of the unknown parameters in the probability density function, and the maximum likelihood estimates of these parameters are those values for which the likelihood function is a maximum. In practice, it is generally more convenient to work with the logarithm of the likelihood function. Those values of the unknown parameters in the density function that maximise the log-likelihood are of course the same values that maximise the likelihood function itself.

We now consider the more usual situation where the survival data include one or more censored survival times. Specifically, suppose that  $r$  of the  $n$

individuals die at times  $t_1, t_2, \dots, t_r$ , and that the survival times of the remaining  $n - r$  individuals,  $t_1^*, t_2^*, \dots, t_{n-r}^*$ , are right-censored. The  $r$  death times contribute a term of the form

$$\prod_{j=1}^r f(t_j)$$

to the overall likelihood function. Naturally, we cannot ignore information about the survival experience of the  $n - r$  individuals for whom a censored survival time has been recorded. If a survival time is censored at time  $t^*$ , say, we know that the lifetime of the individual is at least  $t^*$ , and the probability of this event is  $P(T \geq t^*)$ , which is  $S(t^*)$ . Thus each censored observation contributes a term of this form to the likelihood of the  $n$  observations. The total likelihood function is therefore

$$\prod_{j=1}^r f(t_j) \prod_{l=1}^{n-r} S(t_l^*), \quad (5.11)$$

in which the first product is taken over the  $r$  death times and the second over the  $n - r$  censored survival times.

More compactly, suppose that the data are regarded as  $n$  pairs of observations, where the pair for the  $i$ th individual is  $(t_i, \delta_i)$ ,  $i = 1, 2, \dots, n$ . In this notation,  $\delta_i$  is an indicator variable that takes the value zero when the survival time  $t_i$  is censored and unity when  $t_i$  is an uncensored survival time. The likelihood function can then be written as

$$\prod_{i=1}^n \{f(t_i)\}^{\delta_i} \{S(t_i)\}^{1-\delta_i}. \quad (5.12)$$

This function, which is equivalent to that in expression (5.11), can then be maximised with respect to the unknown parameters in the density and survivor functions. A more careful derivation of this likelihood function is given in Appendix B, which shows the relevance of the assumption of non-informative censoring, mentioned in Section 1.1 of Chapter 1.

An alternative expression for the likelihood function can be obtained by writing expression (5.12) in the form

$$\prod_{i=1}^n \left\{ \frac{f(t_i)}{S(t_i)} \right\}^{\delta_i} S(t_i),$$

so that, from equation (1.3) of Chapter 1, this becomes

$$\prod_{i=1}^n \{h(t_i)\}^{\delta_i} S(t_i). \quad (5.13)$$

This version of the likelihood function is particularly useful when the probability density function has a complicated form, as it often does. Estimates of the unknown parameters in this likelihood function are then found by maximising the logarithm of the likelihood function.

We now consider fitting exponential and Weibull distributions to a single sample of survival data.

### 5.3.1\* Fitting the exponential distribution

Suppose that the survival times of  $n$  individuals,  $t_1, t_2, \dots, t_n$ , are assumed to have an exponential distribution with mean  $\lambda^{-1}$ . Further suppose that the data give the actual death times of  $r$  individuals, and that the remaining  $n - r$  survival times are right-censored.

For the exponential distribution,

$$f(t) = \lambda e^{-\lambda t}, \quad S(t) = e^{-\lambda t},$$

and on substituting into expression (5.12), the likelihood function for the  $n$  observations is given by

$$L(\lambda) = \prod_{i=1}^n (\lambda e^{-\lambda t_i})^{\delta_i} (e^{-\lambda t_i})^{1-\delta_i},$$

where  $\delta_i$  is zero if the survival time of the  $i$ th individual is censored and unity otherwise. After some simplification,

$$L(\lambda) = \prod_{i=1}^n \lambda^{\delta_i} e^{-\lambda t_i},$$

and the corresponding log-likelihood function is

$$\log L(\lambda) = \sum_{i=1}^n \delta_i \log \lambda - \lambda \sum_{i=1}^n t_i.$$

Since the data contain  $r$  deaths,  $\sum_{i=1}^n \delta_i = r$  and the log-likelihood function becomes

$$\log L(\lambda) = r \log \lambda - \lambda \sum_{i=1}^n t_i.$$

We now need to identify the value  $\hat{\lambda}$ , for which the log-likelihood function is a maximum. Differentiation with respect to  $\lambda$  gives

$$\frac{d \log L(\lambda)}{d \lambda} = \frac{r}{\lambda} - \sum_{i=1}^n t_i,$$

and equating the derivative to zero and evaluating it at  $\hat{\lambda}$  gives

$$\hat{\lambda} = r / \sum_{i=1}^n t_i \quad (5.14)$$

for the maximum likelihood estimator of  $\lambda$ .

The mean of an exponential distribution is  $\mu = \lambda^{-1}$ , and so the maximum

likelihood estimator of  $\mu$  is

$$\hat{\mu} = \hat{\lambda}^{-1} = \frac{1}{r} \sum_{i=1}^n t_i.$$

This estimator of  $\mu$  is the total time survived by the  $n$  individuals in the data set divided by the number of deaths observed. The estimator therefore has intuitive appeal as an estimate of the mean lifetime from censored survival data.

The standard error of either  $\hat{\lambda}$  or  $\hat{\mu}$  can be obtained from the second derivative of the log-likelihood function, using a result from the theory of maximum likelihood estimation given in Appendix A. Differentiating  $\log L(\lambda)$  a second time gives

$$\frac{d^2 \log L(\lambda)}{d \lambda^2} = -\frac{r}{\lambda^2},$$

and so the asymptotic variance of  $\hat{\lambda}$  is

$$\text{var}(\hat{\lambda}) = \left\{ -E \left( \frac{d^2 \log L(\lambda)}{d \lambda^2} \right) \right\}^{-1} = \frac{\lambda^2}{r}.$$

Consequently, the standard error of  $\hat{\lambda}$  is given by

$$\text{se}(\hat{\lambda}) = \hat{\lambda} / \sqrt{r}. \quad (5.15)$$

This result could be used to obtain a confidence interval for the mean survival time. In particular, the limits of a  $100(1 - \alpha)\%$  confidence interval for  $\lambda$  are  $\hat{\lambda} \pm z_{\alpha/2} \text{se}(\hat{\lambda})$ , where  $z_{\alpha/2}$  is the upper  $\alpha/2$ -point of the standard normal distribution.

In presenting the results of a survival analysis, the estimated survivor and hazard functions, and the median and other percentiles of the distribution of survival times, are useful. Once an estimate of  $\lambda$  has been found, all these functions can be estimated using the results given in Section 5.1.1. In particular, under the assumed exponential distribution, the estimated hazard function is  $\hat{h}(t) = \hat{\lambda}$  and the estimated survivor function is  $\hat{S}(t) = \exp(-\hat{\lambda}t)$ . In addition, the estimated  $p$ th percentile is given by

$$\hat{t}(p) = \frac{1}{\hat{\lambda}} \log \left( \frac{100}{100-p} \right), \quad (5.16)$$

and the estimated median survival time is

$$\hat{t}(50) = \hat{\lambda}^{-1} \log 2. \quad (5.17)$$

The standard error of an estimate of the  $p$ th percentile of the distribution of survival times can be found using the result for the approximate variance of a function of a random variable given in equation (2.9) of Chapter 2. According to this result, an approximation to the variance of a function  $g(\hat{\lambda})$  of  $\hat{\lambda}$  is such that

$$\text{var}\{g(\hat{\lambda})\} \approx \left\{ \frac{dg(\hat{\lambda})}{d\hat{\lambda}} \right\}^2 \text{var}(\hat{\lambda}). \quad (5.18)$$

Using this result, the approximate variance of the estimated  $p$ th percentile is given by

$$\text{var}\{\hat{t}(p)\} \approx \left\{ -\frac{1}{\hat{\lambda}^2} \log\left(\frac{100}{100-p}\right) \right\}^2 \text{var}(\hat{\lambda}).$$

On simplifying this and taking the square root, we get

$$\text{se}\{\hat{t}(p)\} = \frac{1}{\hat{\lambda}^2} \log\left(\frac{100}{100-p}\right) \text{se}(\hat{\lambda}),$$

and on further substituting for  $\text{se}(\hat{\lambda})$  from equation (5.15) and  $\hat{t}(p)$  from equation (5.16), we find

$$\text{se}\{\hat{t}(p)\} = \hat{t}(p)/\sqrt{r}. \quad (5.19)$$

In particular, the standard error of the estimated median survival time is

$$\text{se}\{\hat{t}(50)\} = \hat{t}(50)/\sqrt{r}. \quad (5.20)$$

Confidence intervals for a true percentile are best obtained from exponentiating the confidence limits for the logarithm of the percentile. This procedure ensures that confidence limits for the percentile will be non-negative. Again making use of the result in equation (5.18), the standard error of  $\log \hat{t}(p)$  is given by

$$\text{se}\{\log \hat{t}(p)\} = \hat{t}(p)^{-1} \text{se}\{\hat{t}(p)\},$$

and after substituting for  $\text{se}\{\hat{t}(p)\}$  from equation (5.19), this standard error becomes

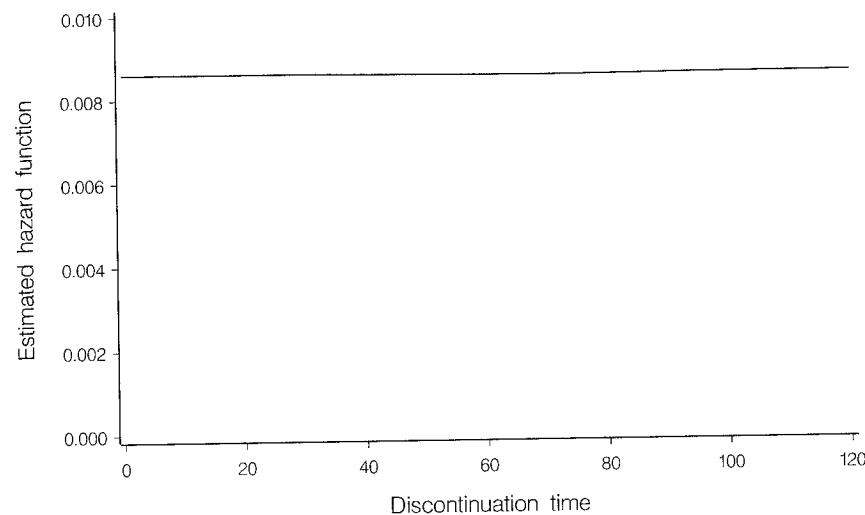
$$\text{se}\{\log \hat{t}(p)\} = 1/\sqrt{r}.$$

Using this result,  $100(1 - \alpha)\%$  confidence limits for the  $100p$ th percentile are  $\exp\{\log \hat{t}(p) \pm z_{\alpha/2}/\sqrt{r}\}$ , that is,  $\hat{t}(p) \exp\{\pm z_{\alpha/2}/\sqrt{r}\}$ , where  $z_{\alpha/2}$  is the upper  $\alpha/2$ -point of the standard normal distribution.

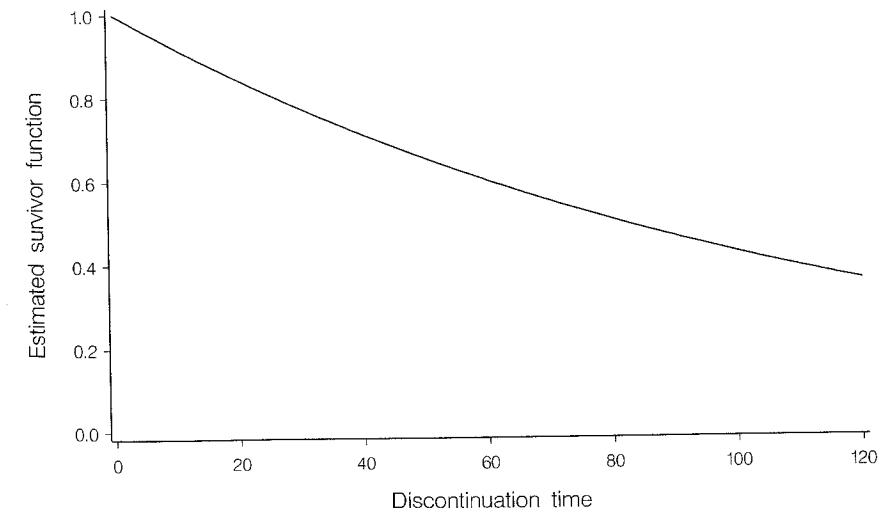
### Example 5.2 Time to discontinuation of the use of an IUD

In this example, the data of Example 1.1 on the times to discontinuation of an IUD for 18 women are analysed under the assumption of a constant hazard of discontinuation. An exponential distribution is therefore fitted to the discontinuation times. For these data, the total of the observed and right-censored discontinuation times is 1046 days, and the number of uncensored times is 9. Therefore, using equation (5.14),  $\hat{\lambda} = 9/1046 = 0.0086$ , and the standard error of  $\hat{\lambda}$  from equation (5.15) is  $\text{se}(\hat{\lambda}) = 0.0086/\sqrt{9} = 0.0029$ . The estimated hazard function is therefore  $\hat{h}(t) = 0.0086$ ,  $t > 0$ , and the estimated survivor function is  $\hat{S}(t) = \exp(-0.0086 t)$ . The estimated hazard and survivor functions are shown in Figures 5.7 and 5.8, respectively.

Estimates of the median and other percentiles of the distribution of discontinuation times can be found from Figure 5.8, but more accurate estimates are obtained from equation (5.16). In particular, using equation (5.17), the median discontinuation time is 81 days, and an estimate of the 90th percentile of the distribution of discontinuation times is, from equation (5.16),  $\hat{t}(90) = \log 10/0.0086 = 267.61$ . This means that on the assumption that the



**Figure 5.7** Estimated hazard function on fitting the exponential distribution.



**Figure 5.8** Estimated survivor function on fitting the exponential distribution.

risk of discontinuing the use of an IUD is independent of time, 90% of women will have a discontinuation time of less than 268 days.

From equation (5.20), the standard error of the estimated median time to discontinuation is  $80.56/\sqrt{9}$ , that is, 26.85 days. The limits of a 95% confidence interval for the true median discontinuation time are

$$80.56 \exp\{\pm 1.96/\sqrt{9}\},$$

and so the interval is from 42 days to 155 days. Confidence intervals for other percentiles can be calculated in a similar manner.

### 5.3.2\* Fitting the Weibull distribution

The survival times of  $n$  individuals are now taken to be a censored sample from a Weibull distribution with scale parameter  $\lambda$  and shape parameter  $\gamma$ . Suppose that there are  $r$  deaths among the  $n$  individuals and  $n - r$  right-censored survival times. We can again use expression (5.12) to obtain the likelihood of the sample data. The probability density, survivor and hazard function of a  $W(\lambda, \gamma)$  distribution are given by

$$f(t) = \lambda \gamma t^{\gamma-1} \exp(-\lambda t^\gamma), \quad S(t) = \exp(-\lambda t^\gamma), \quad h(t) = \lambda \gamma t^{\gamma-1},$$

and so, from expression (5.12), the likelihood of the  $n$  survival times is

$$\prod_{i=1}^n \left\{ \lambda \gamma t_i^{\gamma-1} \exp(-\lambda t_i^\gamma) \right\}^{\delta_i} \{ \exp(-\lambda t_i^\gamma) \}^{1-\delta_i},$$

where  $\delta_i$  is zero if the  $i$ th survival time is censored and unity otherwise. Equivalently, from expression (5.13), the likelihood function is

$$\prod_{i=1}^n \left\{ \lambda \gamma t_i^{\gamma-1} \right\}^{\delta_i} \exp(-\lambda t_i^\gamma).$$

This is regarded as a function of  $\lambda$  and  $\gamma$ , the unknown parameters in the Weibull distribution, and so can be written  $L(\lambda, \gamma)$ . The corresponding log-likelihood function is given by

$$\log L(\lambda, \gamma) = \sum_{i=1}^n \delta_i \log(\lambda \gamma) + (\gamma - 1) \sum_{i=1}^n \delta_i \log t_i - \lambda \sum_{i=1}^n t_i^\gamma,$$

and noting that  $\sum_{i=1}^n \delta_i = r$ , the log-likelihood becomes

$$\log L(\lambda, \gamma) = r \log(\lambda \gamma) + (\gamma - 1) \sum_{i=1}^n \delta_i \log t_i - \lambda \sum_{i=1}^n t_i^\gamma.$$

The maximum likelihood estimates of  $\lambda$  and  $\gamma$  are found by differentiating this function with respect to  $\lambda$  and  $\gamma$ , equating the derivatives to zero, and evaluating them at  $\hat{\lambda}$  and  $\hat{\gamma}$ . The resulting equations are

$$\frac{r}{\hat{\lambda}} - \sum_{i=1}^n t_i^{\hat{\gamma}} = 0, \quad (5.21)$$

and

$$\frac{r}{\hat{\gamma}} + \sum_{i=1}^n \delta_i \log t_i - \hat{\lambda} \sum_{i=1}^n t_i^{\hat{\gamma}} \log t_i = 0. \quad (5.22)$$

From equation (5.21),

$$\hat{\lambda} = r / \sum_{i=1}^n t_i^{\hat{\gamma}}, \quad (5.23)$$

and on substituting for  $\hat{\lambda}$  in equation (5.22), we get the equation

$$\frac{r}{\hat{\gamma}} + \sum_{i=1}^n \delta_i \log t_i - \frac{r}{\sum_i t_i^{\hat{\gamma}}} \sum_{i=1}^n t_i^{\hat{\gamma}} \log t_i = 0. \quad (5.24)$$

This is a non-linear equation in  $\hat{\gamma}$ , which can only be solved using an iterative numerical procedure. Once the estimate,  $\hat{\gamma}$ , which satisfies equation (5.24), has been found, equation (5.23) can be used to obtain  $\hat{\lambda}$ .

In practice, a numerical procedure, such as the Newton-Raphson algorithm, is used to find the values  $\hat{\lambda}$  and  $\hat{\gamma}$  which maximise the likelihood function simultaneously. This procedure was described in Section 3.3.3 of Chapter 3, in connection with fitting the Cox regression model. In that section it was noted that an important by-product of the Newton-Raphson procedure is an approximation to the variance-covariance matrix of the parameter estimates, from which their standard errors can be obtained.

Once estimates of the parameters  $\lambda$  and  $\gamma$  have been found from fitting the Weibull distribution to the observed data, percentiles of the survival time distribution can be estimated using equation (5.9). The estimated  $p$ th percentile of the distribution is

$$\hat{t}(p) = \left\{ \frac{1}{\hat{\lambda}} \log \left( \frac{100}{100-p} \right) \right\}^{1/\hat{\gamma}}, \quad (5.25)$$

and so the estimated median survival time is given by

$$\hat{t}(50) = \left\{ \frac{1}{\hat{\lambda}} \log 2 \right\}^{1/\hat{\gamma}}. \quad (5.26)$$

The standard error of the estimated  $p$ th percentile can be obtained using a generalisation of the result in equation (5.18) to the case where the approximate variance of a function of two estimates is required. Details of the derivation are given in Appendix C, where it is shown that

$$\begin{aligned} \text{se } \{\hat{t}(p)\} &= \frac{\hat{t}(p)}{\hat{\lambda} \hat{\gamma}^2} \left\{ \hat{\gamma}^2 \text{var}(\hat{\lambda}) + \hat{\lambda}^2 (c_p - \log \hat{\lambda})^2 \text{var}(\hat{\gamma}) \right. \\ &\quad \left. + 2\hat{\lambda}\hat{\gamma} (c_p - \log \hat{\lambda}) \text{cov}(\hat{\lambda}, \hat{\gamma}) \right\}^{\frac{1}{2}}, \end{aligned} \quad (5.27)$$

where

$$c_p = \log \log \left( \frac{100}{100-p} \right).$$

The variances of  $\hat{\lambda}$  and  $\hat{\gamma}$ , and their covariance, are found from the variance-covariance matrix of the estimates.

As before, a confidence interval for the true value of the  $p$ th percentile,  $t(p)$ , is best obtained from the corresponding interval for  $\log t(p)$ . The standard error of  $\log \hat{t}(p)$  is

$$\text{se}\{\log \hat{t}(p)\} = \frac{1}{\hat{t}(p)} \text{se}\{\hat{t}(p)\}, \quad (5.28)$$

and  $100(1 - \alpha)\%$  confidence limits for  $\log t(p)$  are

$$\log \hat{t}(p) \pm z_{\alpha/2} \text{se}\{\log \hat{t}(p)\}.$$

Corresponding interval estimates for  $t(p)$  are found by exponentiating these limits. For example, the limits of a  $100(1 - \alpha)\%$  confidence interval for the median survival time,  $t(50)$ , are  $\hat{t}(50) \exp[\pm z_{\alpha/2} \text{se}\{\log \hat{t}(50)\}]$ .

There is a substantial amount of arithmetic involved in these calculations, and care must be taken to ensure that significant figures are not lost during the course of the calculation. For this reason, it is better to perform the calculations using a suitable computer program.

#### *Example 5.3 Time to discontinuation of the use of an IUD*

In Example 5.1, it was found that an exponential distribution provides a satisfactory model for the data on the discontinuation times of 18 IUD users. For comparison, a Weibull distribution will be fitted to the same data set. The distribution can be fitted using computer software, and from the resulting output, the estimated scale parameter of the distribution is found to be  $\hat{\lambda} = 0.000454$ , while the estimated shape parameter is  $\hat{\gamma} = 1.676$ . The standard errors of these estimates are given by  $\text{se}(\hat{\lambda}) = 0.000965$  and  $\text{se}(\hat{\gamma}) = 0.460$ , respectively. Note that approximate confidence limits for the shape parameter,  $\gamma$ , found using  $\hat{\gamma} \pm 1.96 \text{se}(\hat{\gamma})$ , include unity, suggesting that the exponential distribution would provide a satisfactory model for the discontinuation times.

The estimated hazard and survivor functions are obtained by substituting these estimates into equations (5.7) and (5.8), whence

$$\hat{h}(t) = \hat{\lambda} \hat{\gamma} t^{\hat{\gamma}-1},$$

and

$$\hat{S}(t) = \exp(-\hat{\lambda} t^{\hat{\gamma}}).$$

These two functions are shown in Figures 5.9 and 5.10.

Although percentiles of the discontinuation time can be read from the estimated survivor function in Figure 5.10, they are better estimated using equation (5.25). Hence, under the Weibull distribution, the median discontinuation time can be estimated using equation (5.26), and is given by

$$\hat{t}(50) = \left\{ \frac{1}{0.000454} \log 2 \right\}^{1/1.676} = 79.27.$$

As a check, notice that this is perfectly consistent with the value of the discontinuation time corresponding to  $\hat{S}(t) = 0.5$  in Figure 5.10. The standard

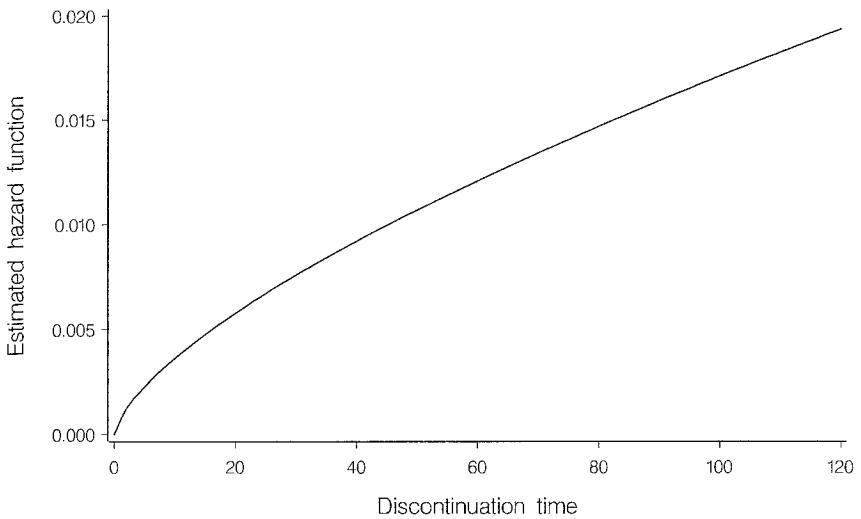


Figure 5.9 Estimated hazard function on fitting the Weibull distribution.

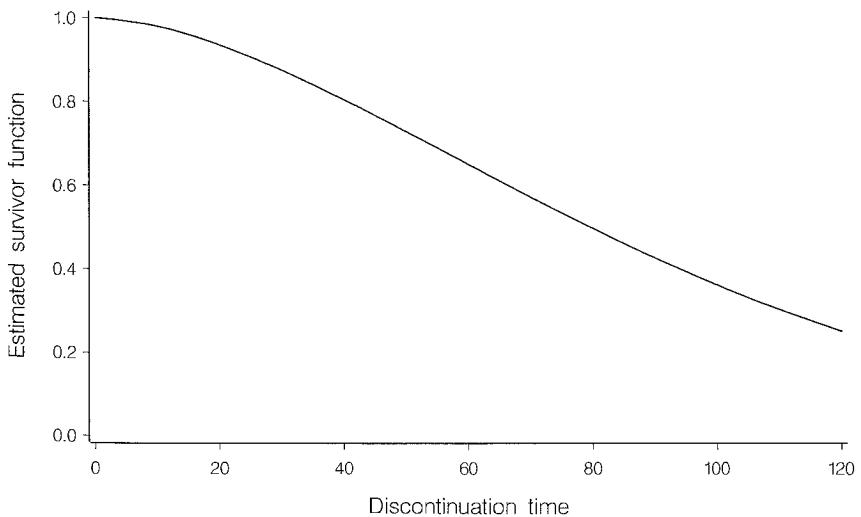


Figure 5.10 Estimated survivor function on fitting the Weibull distribution.

error of this estimate, from equation (5.27) is, after much arithmetic, found to be

$$\text{se}\{\hat{t}(50)\} = 15.795.$$

In order to obtain a 95% confidence interval for the median discontinuation time, the standard error of  $\log \hat{t}(50)$  is required. From equation (5.28),

$$\text{se}\{\log \hat{t}(50)\} = \frac{15.795}{79.272} = 0.199,$$

and so the required confidence limits for the log median discontinuation time are  $\log 79.272 \pm 1.96 \times 0.199$ , that is, (3.982, 4.763). The corresponding interval estimate for the true median discontinuation time is (53.64, 117.15), so that there is a 95% chance that the interval from 54 days to 117 days includes the true value of the median discontinuation time. This interval is rather wide because of the small number of actual discontinuation times in the data set.

It is interesting to compare these results with those found in Example 5.2, where the discontinuation times were modelled using an exponential distribution. The estimated median survival times are very similar, at 80.6 days for the exponential and 79.3 days for the Weibull model. However, the standard error of the estimated median survival time is 26.8 days when the times are assumed to have an exponential distribution, and only 15.8 days under the Weibull model. The median is therefore estimated more precisely when the discontinuation times are assumed to have a Weibull distribution.

Other percentiles of the discontinuation time distribution, and accompanying standard errors and confidence intervals, can be found in a similar fashion. For example, the 90th percentile, that is, the time beyond which 10% of those in the study continue with the use of the IUD, is 162.23 days, and 95% confidence limits for the true percentile are from 95.41 to 275.84 days. Notice that the width of this confidence interval is larger than that for the median discontinuation time, reflecting the fact that the median is more precisely estimated than other percentiles.

#### 5.4 A model for the comparison of two groups

We saw in Section 3.1 that a convenient general model for comparing two groups of survival times is the proportional hazards model. Here, the two groups will be labelled Group I and Group II, and  $X$  will be an indicator variable that takes the value zero if an individual is in Group I and unity if an individual is in Group II. Under the proportional hazards model, the hazard of death at time  $t$  for the  $i$ th individual is given by

$$h_i(t) = e^{\beta x_i} h_0(t), \quad (5.29)$$

where  $x_i$  is the value of  $X$  for the  $i$ th individual. Consequently, the hazard at time  $t$  for an individual in Group I is  $h_0(t)$ , and that for an individual in Group II is  $\psi h_0(t)$ , where  $\psi = \exp(\beta)$ . The quantity  $\beta$  is then the logarithm of the ratio of the hazard for an individual in Group II, to that of an individual in Group I.

We will now make the additional assumption that the survival times for the individuals in Group I have a Weibull distribution with scale parameter  $\lambda$  and shape parameter  $\gamma$ . Using equation (5.29), the hazard function for the individuals in this group is  $h_0(t)$ , where

$$h_0(t) = \lambda \gamma t^{\gamma-1}.$$

Now, also from equation (5.29), the hazard function for those in Group II is  $\psi h_0(t)$ , that is,

$$\psi \lambda \gamma t^{\gamma-1}.$$

This is the hazard function for a Weibull distribution with scale parameter  $\psi \lambda$  and shape parameter  $\gamma$ . We therefore have the result that if the survival times of individuals in one group have a Weibull distribution with shape parameter  $\gamma$ , and the hazard of death at time  $t$  for an individual in the second group is proportional to that of an individual in the first, the survival times of those in the second group will also have a Weibull distribution with shape parameter  $\gamma$ . The Weibull distribution is then said to have the *proportional hazards property*. This property is another reason for the importance of the Weibull distribution in the analysis of survival data.

##### 5.4.1 The log-cumulative hazard plot

When a single sample of survival times has a Weibull distribution  $W(\lambda, \gamma)$ , the log-cumulative hazard plot described in Section 5.2 will give a straight line with intercept  $\log \lambda$  and slope  $\gamma$ . It then follows that if the survival times in a second group have a  $W(\psi \lambda, \gamma)$  distribution, as they would under the proportional hazards model in equation (5.29), the log-cumulative hazard plot will give a straight line, also of slope  $\gamma$ , but with intercept  $\log \psi + \log \lambda$ . If the estimated log-cumulative hazard function is plotted against the logarithm of the survival time for individuals in two groups, parallel straight lines would mean that the assumptions of a proportional hazards model and Weibull survival times were tenable. The vertical separation of the two lines provides an estimate of  $\beta = \log \psi$ , the logarithm of the relative hazard.

If the two lines in a log-cumulative hazard plot are essentially straight, but not parallel, this means that the shape parameter,  $\gamma$ , is different in the two groups, and the hazards are no longer proportional. If the lines are not particularly straight, the Weibull model may not be appropriate. However, if the curves can be taken to be parallel, this would mean that the proportional hazards model is valid, and the Cox regression model discussed in Chapter 3 might be more satisfactory.

##### Example 5.4 Prognosis for women with breast cancer

In this example, we investigate whether the Weibull proportional hazards model is likely to be appropriate for the data of Example 1.2 on the survival times of breast cancer patients. These data relate to women classified according to whether their tumours were positively or negatively stained. The Kaplan-Meier estimate of the survivor functions for the women in each group

were shown in Figure 2.9. From these estimates, the log-cumulative hazards can be estimated and plotted against  $\log t$ . The resulting log-cumulative hazard plot is shown in Figure 5.11.

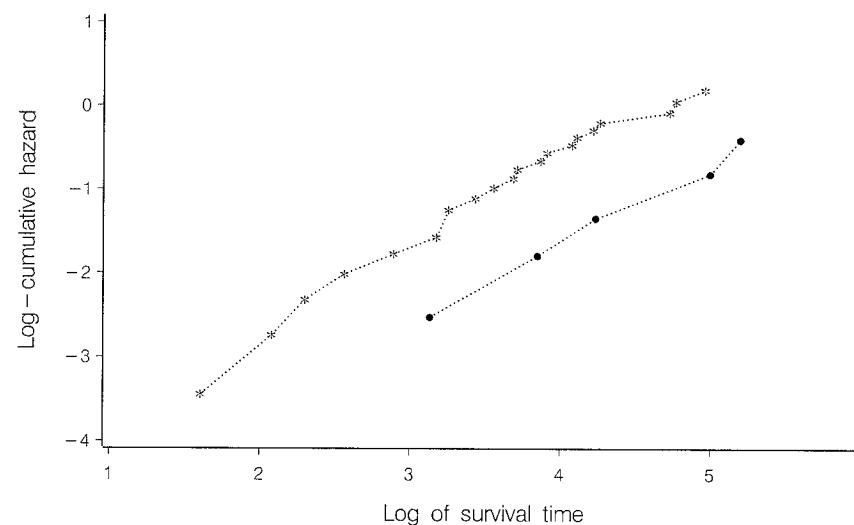


Figure 5.11 Log-cumulative hazard plot for women with tumours that were positively stained (\*) and negatively stained (●).

In this figure, the lines corresponding to the two staining groups are reasonably straight. This means that the assumption of Weibull distributions for the survival times of the women in each group is quite plausible. Moreover, the gradients of the two lines are very similar, which means that the proportional hazards model is valid. The vertical separation of the two lines provides an estimate of the log relative hazard. From Figure 5.11, the vertical distance between the two straight lines is approximately 1.0, and so a rough estimate of the hazard ratio is  $e^{1.0} = 2.72$ . Women in the positively stained group would appear to have nearly three times the risk of death at any time compared to those in the negatively stained group. More accurate estimates of the relative hazard will be obtained from fitting exponential and Weibull models to the data of this example, in Examples 5.5 and 5.6.

#### 5.4.2\* Fitting the model

The proportional hazards model in equation (5.29) can be fitted using the method of maximum likelihood. To illustrate the process, we consider the situation where the survival times in each group have an exponential distribution.

Suppose that the observations from  $n_1$  individuals in Group I can be expressed as  $(t_{i1}, \delta_{i1})$ ,  $i = 1, 2, \dots, n_1$ , where  $\delta_{i1}$  takes the value zero if the survival time of the  $i$ th individual in that group is censored, and unity if that

time is a death time. Similarly, let  $(t_{i'2}, \delta_{i'2})$ ,  $i' = 1, 2, \dots, n_2$ , be the observations from the  $n_2$  individuals in Group II. For individuals in Group I, the hazard function will be taken to be  $\lambda$ , and the probability density function and survivor function are given by

$$f(t_{i1}) = \lambda e^{-\lambda t_{i1}}, \quad S(t_{i1}) = e^{-\lambda t_{i1}}.$$

For those in Group II, the hazard function is  $\psi\lambda$ , and the probability density function and survivor function are

$$f(t_{i'2}) = \psi\lambda e^{-\psi\lambda t_{i'2}}, \quad S(t_{i'2}) = e^{-\psi\lambda t_{i'2}}.$$

Using equation (5.12), the likelihood of the  $n_1 + n_2$  observations,  $L(\psi, \lambda)$ , is

$$\prod_{i=1}^{n_1} \{\lambda e^{-\lambda t_{i1}}\}^{\delta_{i1}} \{e^{-\lambda t_{i1}}\}^{1-\delta_{i1}} \prod_{i'=1}^{n_2} \{\psi\lambda e^{-\psi\lambda t_{i'2}}\}^{\delta_{i'2}} \{e^{-\psi\lambda t_{i'2}}\}^{1-\delta_{i'2}},$$

which simplifies to

$$\prod_{i=1}^{n_1} \lambda^{\delta_{i1}} e^{-\lambda t_{i1}} \prod_{i'=1}^{n_2} (\psi\lambda)^{\delta_{i'2}} e^{-\psi\lambda t_{i'2}}.$$

If the numbers of actual death times in the two groups are  $r_1$  and  $r_2$ , respectively, then  $r_1 = \sum_i \delta_{i1}$  and  $r_2 = \sum_{i'} \delta_{i'2}$ , and the log-likelihood function is given by

$$\log L(\psi, \lambda) = r_1 \log \lambda - \lambda \sum_{i=1}^{n_1} t_{i1} + r_2 \log(\psi\lambda) - \psi\lambda \sum_{i'=1}^{n_2} t_{i'2}.$$

Now write  $T_1$  and  $T_2$  for the total known time survived by the individuals in Groups I and II, respectively. Then,  $T_1$  and  $T_2$  are the totals of uncensored and censored survival times in each group, so that the log-likelihood function becomes

$$\log L(\psi, \lambda) = (r_1 + r_2) \log \lambda + r_2 \log \psi - \lambda(T_1 + \psi T_2).$$

In order to obtain the values  $\hat{\psi}, \hat{\lambda}$  for which this function is a maximum, we differentiate with respect to  $\psi$  and  $\lambda$ , and set the derivatives equal to zero. The resulting equations that are satisfied by  $\hat{\psi}, \hat{\lambda}$  are

$$\frac{r_2}{\hat{\psi}} - \hat{\lambda}T_2 = 0, \tag{5.30}$$

$$\frac{r_1 + r_2}{\hat{\lambda}} - (T_1 + \hat{\psi}T_2) = 0. \tag{5.31}$$

From equation (5.30),

$$\hat{\lambda} = \frac{r_2}{\hat{\psi}T_2},$$

and on substituting for  $\hat{\lambda}$  in equation (5.31) we get

$$\hat{\psi} = \frac{r_2 T_1}{r_1 T_2}. \tag{5.32}$$

Then, from equation (5.30),

$$\hat{\lambda} = r_1/T_1.$$

Both of these estimates have an intuitive justification. The estimated value of  $\lambda$  is the reciprocal of the average time survived by individuals in Group I, while the estimated relative hazard,  $\hat{\psi}$ , is the ratio of the average times survived by the individuals in the two groups.

The asymptotic variance-covariance matrix of the parameter estimates is the inverse of the information matrix, whose elements are found from the second derivatives of the log-likelihood function; see Appendix A. We have that

$$\frac{d^2 \log L(\psi, \lambda)}{d\psi^2} = -\frac{r_2}{\psi^2}, \quad \frac{d^2 \log L(\psi, \lambda)}{d\lambda^2} = -\frac{r_1 + r_2}{\lambda^2}, \quad \frac{d^2 \log L(\psi, \lambda)}{d\lambda d\psi} = -T_2,$$

and the information matrix is the matrix of negative expected values of these partial derivatives. The only second derivative for which expectations need to be obtained is the derivative with respect to  $\lambda$  and  $\psi$ , for which  $E(T_2)$  is required. This is straightforward when the survival times have an exponential distribution, but as shown in Section 5.1.2, the expected value of a survival time that has a Weibull distribution is much more difficult to calculate. For this reason, the information matrix is usually approximated by using the observed values of the negative second partial derivatives. The observed information matrix is thus

$$\mathbf{I}(\psi, \lambda) = \begin{pmatrix} r_2/\psi^2 & T_2 \\ T_2 & (r_1 + r_2)/\lambda^2 \end{pmatrix},$$

and the inverse of this matrix is

$$\frac{1}{(r_1 + r_2)r_2 - T_2^2\psi^2\lambda^2} \begin{pmatrix} (r_1 + r_2)\psi^2 & -T_2\psi^2\lambda^2 \\ -T_2\psi^2\lambda^2 & r_2\lambda^2 \end{pmatrix}.$$

The standard errors of  $\hat{\psi}$  and  $\hat{\lambda}$  are found by substituting  $\hat{\psi}$  and  $\hat{\lambda}$  for  $\psi$  and  $\lambda$  in this matrix, and taking square roots. Thus, the standard error of  $\hat{\psi}$  is given by

$$se(\hat{\psi}) = \sqrt{\frac{(r_1 + r_2)\hat{\psi}^2}{(r_1 + r_2)r_2 - T_2^2\hat{\psi}^2\hat{\lambda}^2}}.$$

On substituting for  $\hat{\psi}$  and  $\hat{\lambda}$  in the denominator of this expression, this standard error simplifies to

$$\hat{\psi} \sqrt{\frac{r_1 + r_2}{r_1 r_2}}. \quad (5.33)$$

Similarly, the standard error of  $\hat{\lambda}$  turns out to be given by

$$se(\hat{\lambda}) = \hat{\lambda}/\sqrt{r_1}.$$

The standard error of these estimates cannot be used directly in the construction of confidence intervals for  $\psi$  and  $\lambda$ . The reason for this is that the values of both parameters must be positive and their estimated values will tend to have skewed distributions. This means that the assumption of normality, used

in constructing a confidence interval, would not be justified. The distribution of the logarithm of an estimate of either  $\psi$  or  $\lambda$  is much more likely to be symmetric, and so confidence limits for the logarithm of the parameter are found using the standard error of the logarithm of the parameter estimate. The resulting confidence limits are then exponentiated to give an interval estimate for the parameter itself.

The standard error of the logarithm of a parameter estimate can be found using the general result given in equation (5.18). Thus the approximate variance of  $\log \hat{\psi}$  is

$$\text{var}(\log \hat{\psi}) \approx \hat{\psi}^{-2} \text{var}(\hat{\psi}),$$

and so the standard error of  $\log \hat{\psi}$  is given by

$$se(\log \hat{\psi}) \approx \hat{\psi}^{-1} se(\hat{\psi}) = \sqrt{\frac{r_1 + r_2}{r_1 r_2}}. \quad (5.34)$$

A  $100(1 - \alpha)\%$  confidence interval for the logarithm of the relative hazard has limits  $\log \hat{\psi} \pm z_{\alpha/2} se(\log \hat{\psi})$ , and confidence limits for the hazard ratio  $\psi$  are found by exponentiating these limits for  $\log \psi$ . If required, a confidence interval for  $\lambda$  can be found in a similar manner.

#### *Example 5.5 Prognosis for women with breast cancer*

The theoretical results developed in this section will now be illustrated using the data on the survival times of breast cancer patients. The survival times for the women in each group are assumed to have exponential distributions, so that the hazard of death, at any time, for a woman in the negatively stained group is a constant value,  $\lambda$ , while that for a woman in the positively stained group is  $\psi\lambda$ , where  $\psi$  is the hazard ratio.

From the data given in Table 1.2 of Chapter 1, the numbers of death times in the negatively and positively stained groups are, respectively,  $r_1 = 5$  and  $r_2 = 21$ . Also, the total time survived in each group is  $T_1 = 1652$  and  $T_2 = 2679$  months. Using equation (5.32), the estimated hazard of death for a woman in the positively stained group, relative to one in the negatively stained group, is

$$\hat{\psi} = \frac{21 \times 1652}{5 \times 2679} = 2.59,$$

so that a woman in the positively stained group has about two and a half times the risk of death at any given time, compared to a woman whose tumour was negatively stained. This is consistent with the estimated value of  $\psi$  of 2.72 from the graphical procedure used in Example 5.4.

Next, using equation (5.33), the standard error of the estimated hazard ratio is given by

$$se(\hat{\psi}) = 2.59 \sqrt{\frac{5 + 21}{5 \times 21}} = 1.289.$$

In order to obtain a 95% confidence interval for the true relative hazard, the standard error of  $\log \hat{\psi}$  is required. Using equation (5.34), this is found to be given by  $se(\log \hat{\psi}) = 0.498$ , and so 95% confidence limits for  $\log \psi$  are

$\log(2.59) \pm 1.96 \text{ se}(\log \hat{\psi})$ , that is,  $0.952 \pm (1.96 \times 0.498)$ . The confidence interval for the log relative hazard is  $(-0.024, 1.927)$ , and the corresponding interval estimate for the relative hazard itself is  $(e^{-0.024}, e^{1.927})$ , that is,  $(0.98, 6.87)$ . This interval only just includes unity, and suggests that women with positively stained tumours have a poorer prognosis than those whose tumours were negatively stained. This result is consistent with the result of the log-rank test in Example 2.12, where a  $P$ -value of 0.061 was obtained on testing the hypothesis of no group difference.

Computer software is generally required to fit a Weibull model to two groups of survival data, assuming proportional hazards. When the model in equation (5.29) is fitted using a computer package, estimates of  $\beta$ ,  $\lambda$  and  $\gamma$ , and their standard errors, can be obtained from the resulting output. Further calculation may then be needed to obtain an estimate of the relative hazard, and the standard error of this estimate. In particular, the estimated hazard ratio would be obtained as  $\hat{\psi} = \exp(\hat{\beta})$  and  $\text{se}(\hat{\psi})$  found from the equation

$$\text{se}(\hat{\psi}) = \exp(\hat{\beta}) \text{se}(\hat{\beta}),$$

a result that follows from equation (5.18).

The median and other percentiles of the survival time distributions in the two groups can be estimated from the values of  $\hat{\lambda}$  and  $\hat{\psi}$ . For example, from equation (5.25), the estimated  $p$ th percentile for those in Group I is found from

$$\hat{t}(p) = \left\{ \frac{1}{\hat{\lambda}} \log \left( \frac{100}{100-p} \right) \right\}^{1/\hat{\gamma}},$$

and that for individuals in Group II is

$$\hat{t}(p) = \left\{ \frac{1}{\hat{\psi}\hat{\lambda}} \log \left( \frac{100}{100-p} \right) \right\}^{1/\hat{\gamma}}.$$

An expression similar to that in (5.27) can be used to obtain the standard error of an estimated percentile for individuals in each group, once the variances and covariances of the parameter estimates in the model have been found. Specific results for the standard error of percentiles of the survival time distributions in each of the two groups will not be given. Instead, the general expression for the standard error of the  $p$ th percentile after fitting a Weibull model, given in equation (5.27), may be used.

#### Example 5.6 Prognosis for women with breast cancer

In Example 5.4, a Weibull proportional hazards model was found to be appropriate for the data on the survival times of two groups of breast cancer patients. Under this model, the hazard of death at time  $t$  is  $\lambda\gamma t^{\gamma-1}$  for a negatively stained patient and  $\psi\lambda\gamma t^{\gamma-1}$  for a patient who is positively stained.

The estimated value of the shape parameter of the fitted Weibull distribution is  $\hat{\gamma} = 0.937$ . The estimated scale parameter for women in Group I is  $\hat{\lambda} = 0.00414$  and that for women in Group II is  $\hat{\lambda}\hat{\psi} = 0.0105$ . The estimated hazard ratio under this Weibull model is  $\hat{\psi} = 2.55$ , which is not very different

from the value obtained in Example 5.5, on the assumption of exponentially distributed survival times.

Putting  $\hat{\gamma} = 0.937$  and  $\hat{\lambda} = 0.00414$  in equation (5.26) gives 235.89 for the median survival time of those in Group I. The estimated median survival time for women in Group II is found by putting  $\hat{\gamma} = 0.937$  and  $\hat{\lambda} = 0.0105$  in that equation, and gives 87.07 for the estimated median survival time of those women. The median survival time of women whose tumour was positively stained is about one third that of those whose tumour was negatively stained.

Using the general result for the standard error of the median survival time, given in equation (5.27), the standard error of the two medians is found by taking  $p = 50$ ,  $\hat{\gamma} = 0.937$  and  $\hat{\lambda} = 0.00414$  and 0.0105 in turn. They turn out to be 114.126 and 20.550, respectively.

As in Section 5.3.2, 95% confidence limits for the true median survival times for each group of women are best obtained by working with the logarithm of the median. The standard error of  $\log \hat{t}(50)$  is found using equation (5.28), from which

$$\text{se}\{\hat{t}(50)\} = \frac{1}{\hat{t}(50)} \text{se}\{\hat{t}(50)\}.$$

Confidence limits for  $\log t(50)$  are then exponentiated to give the corresponding confidence limits for  $t(50)$  itself.

In this example, 95% confidence intervals for the true median survival times of the two groups of women are (91.4, 608.9) and (54.8, 138.3), respectively. Notice that the confidence interval for the median survival time of patients with positive staining is much narrower than that for women with negative staining. This is due to there being a relatively small number of uncensored survival times in the women whose tumours were negatively stained.

## 5.5 The Weibull proportional hazards model

The model in equation (5.29) for the comparison of two groups of survival data can easily be generalised to give a model that is similar in form to the Cox regression model described in Section 3.1.2. Suppose that the values  $x_1, x_2, \dots, x_p$  of  $p$  explanatory variables,  $X_1, X_2, \dots, X_p$ , are recorded for each of  $n$  individuals. Under the proportional hazards model, the hazard of death at time  $t$  for the  $i$ th individual is

$$h_i(t) = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}) h_0(t), \quad (5.35)$$

for  $i = 1, 2, \dots, n$ . Although this model has a similar appearance to that given in equation (3.3), there is one fundamental difference, which concerns the specification of the baseline hazard function  $h_0(t)$ . In the Cox regression model, the form of  $h_0(t)$  is unspecified, and the shape of the function is essentially determined by the actual data. In the model being considered in this section, the survival times are assumed to have a Weibull distribution, and this imposes a particular parametric form on  $h_0(t)$ .

Consider an individual for whom the values of the  $p$  explanatory variables in the model of equation (5.35) are all equal to zero. The hazard function for

such an individual is  $h_0(t)$ . If the survival time of this individual has a Weibull distribution with scale parameter  $\lambda$  and shape parameter  $\gamma$ , then their hazard function is such that

$$h_0(t) = \lambda \gamma t^{\gamma-1}.$$

Using equation (5.35), the hazard function for the  $i$ th individual in the study is then given by

$$h_i(t) = \exp(\beta' \mathbf{x}_i) \lambda \gamma t^{\gamma-1}, \quad (5.36)$$

where  $\beta' \mathbf{x}_i$  stands for  $\beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}$ . From the form of this hazard function, we can see that the survival time of the  $i$ th individual in the study has a Weibull distribution with scale parameter  $\lambda \exp(\beta' \mathbf{x}_i)$  and shape parameter  $\gamma$ . This again is a manifestation of the proportional hazards property of the Weibull distribution. This result shows that the effect of the explanatory variates in the model is to alter the scale parameter of the distribution, while the shape parameter remains constant.

The survivor function corresponding to the hazard function given in equation (5.36) is found using equation (1.5), and turns out to be

$$S_i(t) = \exp\{-\exp(\beta' \mathbf{x}_i) \lambda t^\gamma\}. \quad (5.37)$$

### 5.5.1\* Fitting the model

The Weibull proportional hazards model is fitted by constructing the likelihood function of the  $n$  observations, and maximising this function with respect to the unknown parameters,  $\beta_1, \beta_2, \dots, \beta_p, \lambda$  and  $\gamma$ . Since the hazard function and survivor function differ for each individual, the likelihood function in expression (5.13) is now written as

$$\prod_{i=1}^n \{h_i(t_i)\}^{\delta_i} S_i(t_i). \quad (5.38)$$

The logarithm of the likelihood function, rather than the likelihood itself, is maximised with respect to the unknown parameters, and from expression (5.38), this is

$$\sum_{i=1}^n \{\delta_i \log h_i(t_i) + \log S_i(t_i)\}.$$

On substituting for  $h_i(t_i)$  and  $S_i(t_i)$  from equations (5.36) and (5.37), the log-likelihood becomes

$$\sum_{i=1}^n [\delta_i \{\beta' \mathbf{x}_i + \log(\lambda \gamma) + (\gamma - 1) \log t_i\} - \lambda \exp(\beta' \mathbf{x}_i) t^\gamma],$$

which can be written as

$$\sum_{i=1}^n [\delta_i \{\beta' \mathbf{x}_i + \log(\lambda \gamma) + \gamma \log t_i\} - \lambda \exp(\beta' \mathbf{x}_i) t^\gamma] - \sum_{i=1}^n \delta_i \log t_i. \quad (5.39)$$

The final term in this expression,  $-\sum_{i=1}^n \delta_i \log t_i$ , does not involve any of the unknown parameters, and can be omitted from the likelihood. The resulting

log-likelihood function is then

$$\sum_{i=1}^n [\delta_i \{\beta' \mathbf{x}_i + \log(\lambda \gamma) + \gamma \log t_i\} - \lambda \exp(\beta' \mathbf{x}_i) t^\gamma], \quad (5.40)$$

which differs from that obtained from the full log-likelihood, given in expression (5.39), by the value of  $\sum_{i=1}^n \delta_i \log t_i$ . When computer software is used to fit the Weibull proportional hazards model, the log-likelihood is generally computed from expression (5.40). This expression will also be used in the examples given in this book.

Computer software for fitting parametric proportional hazards models generally includes the standard errors of the parameter estimates, from which confidence intervals for relative hazards and the median and other percentiles of the survival time distribution can be found. Specifically, suppose that the estimates of the parameters in the model of equation (5.36) are  $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p, \hat{\lambda}$  and  $\hat{\gamma}$ . The estimated survivor function for the  $i$ th individual in the study, for whom the values of the explanatory variables in the model are  $x_{1i}, x_{2i}, \dots, x_{pi}$ , is then

$$\hat{S}_i(t) = \exp\{-\exp(\hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_p x_{pi}) \hat{\lambda} t^{\hat{\gamma}}\}, \quad (5.41)$$

and the corresponding estimated hazard function is

$$\hat{h}_i(t) = \exp(\hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_p x_{pi}) \hat{\lambda} \hat{\gamma} t^{\hat{\gamma}-1}.$$

Both of these functions can be estimated and plotted against  $t$ , for individuals with particular values of the explanatory variables in the model.

Generalising the result in equation (5.25) to the situation where the Weibull scale parameter is  $\lambda \exp(\beta' \mathbf{x}_i)$ , the estimated  $p$ th percentile of the survival time distribution for an individual, whose vector of explanatory variables is  $\mathbf{x}_i$ , is

$$\hat{t}(p) = \left\{ \frac{1}{\hat{\lambda} \exp(\hat{\beta}' \mathbf{x}_i)} \log \left( \frac{100}{100-p} \right) \right\}^{1/\hat{\gamma}}.$$

The estimated median survival time for such an individual is therefore

$$\hat{t}(50) = \left\{ \frac{\log 2}{\hat{\lambda} \exp(\hat{\beta}' \mathbf{x}_i)} \right\}^{1/\hat{\gamma}}. \quad (5.42)$$

The standard error of  $\hat{t}(p)$ , and confidence intervals for  $t(p)$ , can be found after first obtaining the standard error of  $\log \hat{t}(p)$ . The standard error of  $\log \hat{t}(p)$  is shown in equation (C.5) of Appendix C to be given by

$$\text{se}\{\log \hat{t}(p)\} = \hat{\gamma}^{-1} \sqrt{(d'_0 \mathbf{V} d_0)},$$

where  $\mathbf{V}$  is the variance-covariance matrix of the estimated values of the parameters  $\gamma, \lambda, \beta_1, \beta_2, \dots, \beta_p$ , and  $d_0$  is a vector whose  $p+2$  components are  $\hat{\lambda}^{-1}, \hat{\gamma}^{-1} \{c_p - \log \hat{\lambda} - \hat{\beta}' \mathbf{x}\}, x_1, x_2, \dots, x_p$ , with

$$c_p = \log \log \left( \frac{100}{100-p} \right).$$

If required, the standard error of  $\hat{t}(p)$  is found using

$$\text{se}\{\hat{t}(p)\} = \hat{t}(p) \text{se}\{\log \hat{t}(p)\},$$

and as usual, confidence intervals for  $t(p)$  can be found by exponentiating confidence limits for  $\log t(p)$ .

### 5.5.2\* Log-linear form of the model

Most computer software for fitting the Weibull proportional hazards model uses a different form of the model from that adopted in this chapter. The reason for this will be given in the next chapter, but for the moment we note that the model can be formulated as a log-linear model for  $T_i$ , the random variable associated with the survival time of the  $i$ th individual. In this version of the model, the survivor function of  $T_i$ , given in equation (5.37), is expressed as

$$S_i(t) = \exp\left\{-\exp\left(\frac{\log t - \mu - \boldsymbol{\alpha}' \mathbf{x}_i}{\sigma}\right)\right\}.$$

The correspondence between these two representations of the model is such that

$$\lambda = \exp(-\mu/\sigma), \quad \gamma = \sigma^{-1}, \quad \beta_j = -\alpha_j/\sigma,$$

for  $j = 1, 2, \dots, p$ , and  $\mu, \sigma$  are often termed the *intercept* and *scale parameter*, respectively; see Chapter 6 for fuller details.

Generally speaking, it will be more straightforward to use the log-linear form of the model to estimate the hazard and survivor functions, percentiles of the survival time distribution, and their standard errors. The relevant expressions are given as equations (6.17), (6.18) and (6.20) in Section 6.5 of Chapter 6.

However, the log-linear representation of the model makes it difficult to obtain confidence intervals for a log-hazard ratio,  $\beta$ , in a proportional hazards model, since only the standard error of the estimate of  $\alpha$  is given in the output. In particular, on fitting the Weibull proportional hazards model, the output provides the estimated value of  $\alpha = -\sigma\beta$ ,  $\hat{\alpha}$ , and the standard error of  $\hat{\alpha}$ . The corresponding estimate of  $\beta$  is easily found from  $\hat{\beta} = -\hat{\alpha}/\hat{\sigma}$ , but the standard error of  $\hat{\beta}$  is more complicated to calculate.

To obtain the standard error of  $\hat{\beta}$ , we can use the result that the approximate variance of a function of two parameter estimates,  $\hat{\theta}_1, \hat{\theta}_2$ , say, is

$$\left(\frac{\partial g}{\partial \hat{\theta}_1}\right)^2 \text{var}(\hat{\theta}_1) + \left(\frac{\partial g}{\partial \hat{\theta}_2}\right)^2 \text{var}(\hat{\theta}_2) + 2 \left(\frac{\partial g}{\partial \hat{\theta}_1} \frac{\partial g}{\partial \hat{\theta}_2}\right) \text{cov}(\hat{\theta}_1, \hat{\theta}_2). \quad (5.43)$$

This is an extension of the result given in equation (2.9) of Chapter 2 for the approximate variance of a function of a single random variable. To obtain the approximate variance of the function

$$g(\hat{\alpha}, \hat{\sigma}) = -\frac{\hat{\alpha}}{\hat{\sigma}},$$

the derivatives of  $g(\hat{\alpha}, \hat{\sigma})$  are required. We have that

$$\frac{\partial g}{\partial \hat{\alpha}} = -\frac{1}{\hat{\sigma}}, \quad \frac{\partial g}{\partial \hat{\sigma}} = \frac{\hat{\alpha}}{\hat{\sigma}^2},$$

and so using expression (5.43),

$$\text{var}\left(-\frac{\hat{\alpha}}{\hat{\sigma}}\right) \approx \left(-\frac{1}{\hat{\sigma}}\right)^2 \text{var}(\hat{\alpha}) + \left(\frac{\hat{\alpha}}{\hat{\sigma}^2}\right)^2 \text{var}(\hat{\sigma}) + 2\left(-\frac{1}{\hat{\sigma}}\right)\left(\frac{\hat{\alpha}}{\hat{\sigma}^2}\right) \text{cov}(\hat{\alpha}, \hat{\sigma}).$$

After some algebra, the approximate variance becomes

$$\frac{1}{\hat{\sigma}^4} \left\{ \hat{\sigma}^2 \text{var}(\hat{\alpha}) + \hat{\alpha}^2 \text{var}(\hat{\sigma}) - 2\hat{\alpha}\hat{\sigma} \text{cov}(\hat{\alpha}, \hat{\sigma}) \right\}, \quad (5.44)$$

and the square root of this is the standard error of  $\hat{\beta}$ .

#### Example 5.7 Prognosis for women with breast cancer

The computation of a standard error of a log-hazard ratio is now illustrated using the data on the survival times of two groups breast cancer patients. In this example, computer output from fitting the log-linear form of the Weibull proportional hazards model is used to illustrate the estimation of the hazard ratio, and calculation of the standard error of the estimate.

On fitting the model that contains the treatment effect, represented by a variable  $X$ , where  $X = 0$  for a woman with negative staining and  $X = 1$  for positive staining, we find that the estimated coefficient of  $X$  is  $\hat{\alpha} = -0.9967$ . Also, the estimates of  $\mu$  and  $\sigma$  are given by  $\hat{\mu} = 5.8544$  and  $\hat{\sigma} = 1.0668$ , respectively. The estimated log-hazard ratio for a woman with positive staining, ( $X = 1$ ) relative to a woman with negative staining ( $X = 0$ ) is

$$\hat{\beta} = -\frac{-0.9967}{1.0668} = 0.9343.$$

The corresponding hazard ratio is 2.55, as in Example 5.6.

The standard errors of  $\hat{\alpha}$  and  $\hat{\sigma}$  are generally included in standard computer output, and are 0.5441 and 0.1786, respectively. The estimated variances of  $\hat{\alpha}$  and  $\hat{\sigma}$  are therefore 0.2960 and 0.0319, respectively. The covariance between  $\hat{\alpha}$  and  $\hat{\sigma}$  can be found from computer software, although it is not usually part of the default output. It is found to be -0.0213.

Substituting for  $\hat{\alpha}$ ,  $\hat{\sigma}$ , and their variances and covariance in expression (5.44), we get

$$\text{var}(\hat{\beta}) \approx 0.2498,$$

and so the standard error of  $\hat{\beta}$  is given by  $\text{se}(\hat{\beta}) = 0.4998$ . This can be used in the construction of confidence intervals for the corresponding true hazard ratio.

### 5.5.3 Exploratory analyses

In Sections 5.2 and 5.4.1, we saw how a log-cumulative hazard plot could be used to assess whether survival data can be modelled by a Weibull dis-

tribution, and whether the proportional hazards assumption is valid. These procedures work perfectly well when we are faced with a single sample of survival data, or data where the number of groups is small and there is a reasonably large number of individuals in each group. But in situations where there are a small number of death times distributed over a relatively large number of groups, it may not be possible to estimate the survivor function, and hence the log-cumulative hazard function, for each group.

As an example, consider the data on the survival times of patients with hypernephroma, given in Table 3.3. Here, individuals are classified according to age group and whether or not a nephrectomy has been performed, giving six combinations of age group and nephrectomy status. To examine the assumption of a Weibull distribution for the survival times in each group, and the assumption of proportional hazards across the groups, a log-cumulative hazard plot would be required for each group. The number of patients in each age group who have not had a nephrectomy is so small that the survivor function cannot be properly estimated in these groups. If there were more individuals in the study who had died and not had a nephrectomy, it would be possible to construct a log-cumulative hazard plot. If this plot featured six parallel straight lines, the Weibull proportional hazards model is likely to be satisfactory.

When a model contains continuous variables, their values will first need to be grouped before a log-cumulative hazard plot can be obtained. This may also result in there being insufficient numbers of individuals in some groups to enable the log-cumulative hazard function to be estimated.

The only alternative to using each combination of factor levels in constructing a log-cumulative hazard plot is to ignore some of the factors. However, the resulting plot can be very misleading. For example, suppose that patients are classified according to the levels of two factors,  $A$  and  $B$ . The log-cumulative hazard plot obtained by grouping the individuals according to the levels of  $A$  ignoring  $B$ , or according to the levels of  $B$  ignoring  $A$ , may not give cause to doubt the Weibull or proportional hazards assumptions. However, if the log-cumulative hazard plot is obtained for individuals at each combination of levels of  $A$  and  $B$ , the plot may not feature a series of four parallel lines. By the same token, the log-cumulative hazard plot obtained when either  $A$  or  $B$  is ignored may not show sets of parallel straight lines, but when a plot is obtained for all combinations of  $A$  and  $B$ , parallel lines may result. This feature is illustrated in the following example, which is based on artificial data.

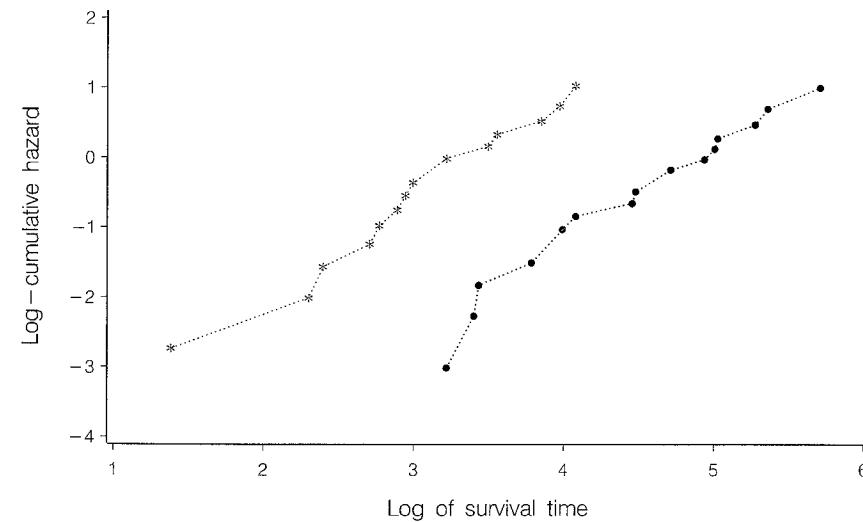
#### *Example 5.8 An artificial data set*

Suppose that a number of individuals are classified according to the levels of two factors,  $A$  and  $B$ , each with two levels, and that their survival times are as shown in Table 5.1. As usual, an asterisk denotes a censored observation.

The log-cumulative hazard plot shown in Figure 5.12 is derived from the individuals classified according to the two levels of  $A$ , ignoring the level of factor  $B$ . The plot in Figure 5.13 is from individuals classified according to the two levels of  $B$ , ignoring the level of factor  $A$ .

**Table 5.1** Artificial data on the survival times of 37 patients classified according to the levels of two factors,  $A$  and  $B$ .

$A = 1$		$A = 2$	
$B = 1$	$B = 2$	$B = 1$	$B = 2$
59	10	88	25*
20	4	70*	111
71	16	54	152
33	18	139	86
25	19	31	212
25	35	59	187*
15	11	111	54
53		149	357
47		30	301
		44	195
		25	



**Figure 5.12** Log-cumulative hazard plot for individuals for whom  $A = 1$  (\*) and  $A = 2$  (●).

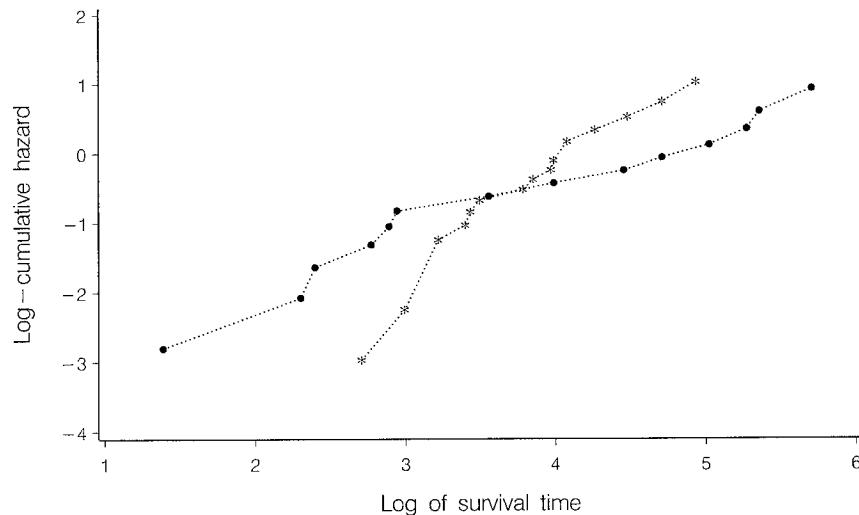


Figure 5.13 Log-cumulative hazard plot for individuals for whom  $B = 1$  (\*) and  $B = 2$  (●).

From Figure 5.12 there is no reason to doubt the assumption of a Weibull distribution for the survival times at the two levels of  $A$ , and the assumption of proportional hazards is clearly tenable. However, the crossed lines on the plot shown as Figure 5.13 strongly suggest that the hazards are not proportional when individuals are classified according to the levels of  $B$ . A different picture emerges when the 37 survival times are classified according to the levels of both  $A$  and  $B$ . The log-cumulative hazard plot based on the four groups is shown in Figure 5.14. The four parallel lines show that there is no doubt about the validity of the proportional hazards assumption across the groups.

In this example, the reason why the log-cumulative hazard plot for  $B$  ignoring  $A$  is misleading is that there is an interaction between  $A$  and  $B$ . An examination of the data reveals that, on average, the difference in the survival times of patients for whom  $B = 1$  and  $B = 2$  is greater when  $A = 2$  than when  $A = 1$ .

Even when a log-cumulative hazard plot gives no reason to doubt the assumption of a Weibull proportional hazards model, the validity of the adequacy of the fitted model will need to be examined using the methods to be described in Chapter 7. When it is not possible to use a log-cumulative hazard plot to explore whether a Weibull distribution provides a reasonable model for the survival times, a procedure based on the Cox regression model, described in Chapter 3, might be helpful. Essentially, a Cox regression model that includes all the relevant explanatory variables is fitted, and the baseline hazard function is estimated, using the procedure described in Section 3.8. A plot of this function may suggest whether or not the assumption of a Weibull

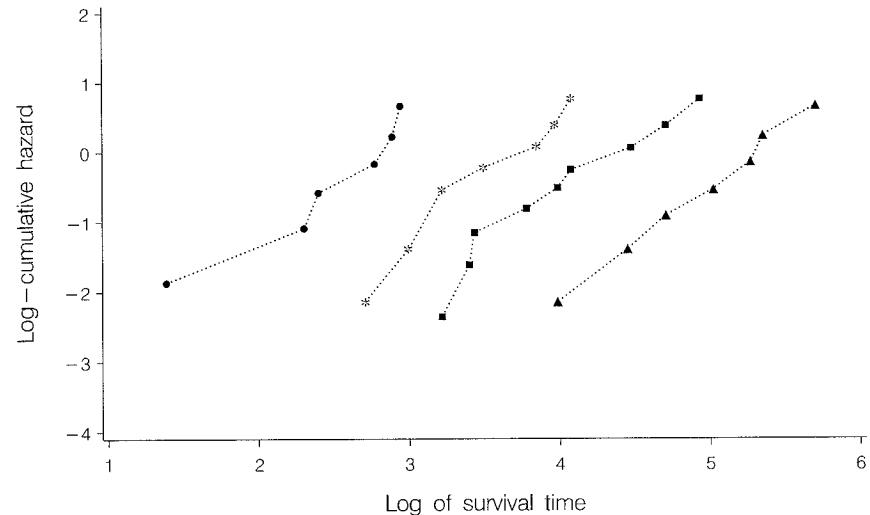


Figure 5.14 Log-cumulative hazard plot for individuals in the groups defined by the four combinations of levels of  $A$  and  $B$ .

distribution is tenable. In particular, if the estimated baseline hazard function in the Cox model is increasing or decreasing, the Weibull model may provide a more concise summary of the baseline hazard function than the Cox regression model. Because the estimated baseline hazard function for a fitted Cox model can be somewhat irregular, comparing the estimated baseline cumulative hazard or the baseline survivor function, under the fitted Cox regression model, with that of the Weibull model may be more fruitful.

## 5.6 Comparing alternative Weibull models

In order to ascertain which explanatory variables should be included in a Weibull proportional hazards model, alternative models need to be compared. Weibull models can be compared using methods analogous to those for the Cox regression model described in Section 3.5.

Suppose that one model contains a subset of the explanatory variables in another, so that the two models are nested. The two models can then be compared on the basis of the statistic  $-2 \log \hat{L}$ , where  $\hat{L}$  is the maximised value of the likelihood function under the fitted model. For a model that contains  $p$  explanatory variables, the sample likelihood is a function of  $p + 2$  unknown parameters,  $\beta_1, \beta_2, \dots, \beta_p, \lambda$  and  $\gamma$ . The maximised likelihood is then the value of this function when these parameters take their estimates,  $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p, \hat{\lambda}$  and  $\hat{\gamma}$ .

More specifically, if one model, Model (1), say, contains  $p$  explanatory variables, and another model, Model (2), contains an additional  $q$  explanatory

variables, the estimated hazard functions for these two models are as given below:

$$\text{Model (1): } h_i(t) = \exp\{\hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p\} \hat{\lambda} \hat{\gamma} t^{\hat{\gamma}-1};$$

$$\text{Model (2): } h_i(t) = \exp\{\hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_{p+q} x_{p+q}\} \hat{\lambda} \hat{\gamma} t^{\hat{\gamma}-1}.$$

The maximised likelihoods under Model (1) and Model (2) will be denoted by  $\hat{L}_1$  and  $\hat{L}_2$ , respectively. The difference between the values of  $-2 \log \hat{L}_1$  and  $-2 \log \hat{L}_2$ , that is,  $-2\{\log \hat{L}_1 - \log \hat{L}_2\}$ , then has an approximate chi-squared distribution with  $q$  degrees of freedom, under the null hypothesis that the coefficients of the additional  $q$  variates in Model (2) are all equal to zero. If the difference between the values of  $-2 \log \hat{L}$  for these two models is significantly large when compared with percentage points of the chi-squared distribution, we would deduce that the extra  $q$  terms are needed in the model, in addition to the  $p$  that are already included. Since differences between values of  $-2 \log \hat{L}$  are used in comparing models, it does not matter whether the maximised log-likelihood, used in computing the value of  $-2 \log \hat{L}$ , is based on expression (5.39) or (5.40).

The description of the modelling process in Section 3.5 applies equally well to models based on the Weibull proportional hazards model, and so will not be repeated here. However, the modelling process will be illustrated using two examples.

#### *Example 5.9 Treatment of hypernephroma*

Data on the survival times of 36 patients, classified according to their age group and whether or not they have had a nephrectomy, were introduced in Example 3.4 of Chapter 3. In that example, the data were analysed using the Cox proportional hazards model. Here, the analysis is repeated using the Weibull proportional hazards model. As in Example 3.4, the effect of the  $j$ th age group will be denoted by  $\alpha_j$ , and that associated with whether or not a nephrectomy was performed by  $\nu_k$ . There are then five possible models for the hazard function of the  $i$ th individual,  $h_i(t)$ , which are as follows:

$$\text{Model (1): } h_i(t) = h_0(t);$$

$$\text{Model (2): } h_i(t) = \exp\{\alpha_j\} h_0(t);$$

$$\text{Model (3): } h_i(t) = \exp\{\nu_k\} h_0(t);$$

$$\text{Model (4): } h_i(t) = \exp\{\alpha_j + \nu_k\} h_0(t);$$

$$\text{Model (5): } h_i(t) = \exp\{\alpha_j + \nu_k + (\alpha\nu)_{jk}\} h_0(t).$$

In these models,  $h_0(t) = \lambda \gamma t^{\gamma-1}$  is the baseline hazard function, and the

parameters  $\lambda$  and  $\gamma$  have to be estimated along with those in the linear component of the model.

These five models have the interpretations given in Example 3.4. They can be fitted by constructing indicator variables corresponding to the factors age group and nephrectomy status, as shown in Example 3.4, but some software packages will allow factors to be fitted directly.

Once a Weibull proportional hazards model has been fitted to the data, values of  $-2 \log \hat{L}$  can be found. These are given in Table 5.2 for the five models of interest.

**Table 5.2** Values of  $-2 \log \hat{L}$  on fitting five Weibull models to the data on hypernephroma.

Model	Terms in model	$-2 \log \hat{L}$
(1)	null model	104.886
(2)	$\alpha_j$	96.400
(3)	$\nu_k$	94.384
(4)	$\alpha_j + \nu_k$	87.758
(5)	$\alpha_j + \nu_k + (\alpha\nu)_{jk}$	83.064

The values of the  $-2 \log \hat{L}$  statistic in Table 5.2, and other examples in this book, have been computed using the log-likelihood in expression (5.40). Accordingly these values may differ from the values given by some computer software packages by an amount equal to  $2 \sum_{i=1}^n \delta_i \log t_i$ , which in this case has the value 136.3733.

The reduction in the value of  $-2 \log \hat{L}$  on adding the interaction term to Model (4) is 4.69 on two degrees of freedom. This reduction is just about significant at the 10% level ( $P = 0.096$ ) and so there is some suggestion of an interaction between age group and nephrectomy status. For comparison, note that when the Cox regression model was fitted in Example 3.4, the interaction was not significant ( $P = 0.220$ ).

The interaction can be investigated in greater detail by examining the hazard ratios under the model. Under Model (5), the estimated hazard function for the  $i$ th individual is

$$\hat{h}_i(t) = \exp\{\hat{\alpha}_j + \hat{\nu}_k + (\hat{\alpha}\hat{\nu})_{jk}\} \hat{h}_0(t),$$

where

$$\hat{h}_0(t) = \hat{\lambda} \hat{\gamma} t^{\hat{\gamma}-1}$$

is the estimated baseline hazard function. The logarithm of the hazard ratio for an individual in the  $j$ th age group,  $j = 1, 2, 3$ , and  $k$ th level of nephrectomy status,  $k = 1, 2$ , relative to an individual in the youngest age group who has not had a nephrectomy, is therefore

$$\hat{\alpha}_j + \hat{\nu}_k + (\hat{\alpha}\hat{\nu})_{jk} - \hat{\alpha}_1 - \hat{\nu}_1 - (\hat{\alpha}\hat{\nu})_{11}, \quad (5.45)$$

since the baseline hazard functions cancel out.

As in Example 3.4, models can be fitted to the data by defining indicator variables  $A_2$  and  $A_3$  for age group and  $N$  for nephrectomy status. As in that example,  $A_2$  is unity for an individual in the second age group and zero otherwise,  $A_3$  is unity for an individual in the third age group and zero otherwise, and  $N$  is unity if a nephrectomy has been performed and zero otherwise. Thus, fitting the term  $\alpha_j$  corresponds to fitting the variables  $A_2$  and  $A_3$ , fitting  $\nu_k$  corresponds to fitting  $N$ , and fitting the interaction term  $(\alpha\nu)_{jk}$  corresponds to fitting the products  $A_2N = A_2 \times N$  and  $A_3N = A_3 \times N$ . In particular, to fit Model (5), the five variables  $A_2, A_3, N, A_2N, A_3N$  are included in the model. With this choice of indicator variables,  $\hat{\alpha}_1 = 0$ ,  $\hat{\nu}_1 = 0$  and  $(\widehat{\alpha\nu})_{jk} = 0$  when either  $i$  or  $j$  is unity. The remaining values of  $\hat{\alpha}_j, \hat{\nu}_k$  and  $(\widehat{\alpha\nu})_{jk}$  are the coefficients of  $A_2, A_3, N, A_2N, A_3N$  and are given in Table 5.3.

**Table 5.3** Parameter estimates on fitting a Weibull model to the data on hypernephroma.

Parameter	Estimate
$\alpha_2$	-0.085
$\alpha_3$	0.115
$\nu_2$	-2.436
$(\alpha\nu)_{22}$	0.121
$(\alpha\nu)_{32}$	2.538

Many computer packages set up indicator variables internally, and so estimates such as those in the above table can be obtained directly from the output. However, to repeat an earlier warning, when packages are used to fit factors, the coding used to define the indicator variables must be known if the output is to be properly interpreted.

When the indicator variables specified above are used, the logarithm of the hazard ratio given in equation (5.45) reduces to

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

for  $j = 1, 2, 3, k = 1, 2$ . Table 5.4 gives the hazards for the individuals, relative to the baseline hazard. The baseline hazard corresponds to an individual in the youngest age group who has not had a nephrectomy, and so a hazard ratio of unity for these individuals is recorded in Table 5.4.

This table helps to explain the interaction between age group and nephrectomy status, in that the effect of a nephrectomy is not the same for individuals in each of the three age groups. For patients in the two youngest age groups, a nephrectomy substantially reduces the hazard of death at any given time. Performing a nephrectomy on patients aged over 70 does not have much effect

**Table 5.4** Hazard ratios for individuals classified by age group and nephrectomy status.

Age group	No nephrectomy	Nephrectomy
< 60	1.00	0.09
60–70	0.92	0.09
> 70	1.12	1.24

on the risk of death. We also see that for those patients who have not had a nephrectomy, age does not much affect the hazard of death.

Estimated median survival times can be found in a similar way. Using equation (5.42), the median survival time for a patient in the  $j$ th age group,  $j = 1, 2, 3$ , and the  $k$ th level of nephrectomy status,  $k = 1, 2$ , becomes

$$\hat{t}(50) = \left\{ \frac{\log 2}{\hat{\lambda} \exp\{\hat{\alpha}_j + \hat{\nu}_k + (\widehat{\alpha\nu})_{jk}\}} \right\}^{1/\hat{\gamma}}.$$

When the model containing the interaction term is fitted to the data, the estimated values of the parameters in the baseline hazard function are  $\hat{\lambda} = 0.0188$  and  $\hat{\gamma} = 1.5538$ . Table 5.5 gives the estimated median survival times, in months, for individuals with each combination of age group and nephrectomy status. This table shows that a nephrectomy leads to more than a fourfold

**Table 5.5** Median survival times for individuals classified by age group and nephrectomy status.

Age group	No nephrectomy	Nephrectomy
< 60	10.21	48.94
60–70	10.78	47.81
> 70	9.48	8.87

increase in the median survival time in patients aged up to 70 years. The median survival time of patients aged over 70 is not much affected by the performance of a nephrectomy.

We end this example with a note of caution. For some combinations of age group and nephrectomy status, particularly the groups of individuals who have not had a nephrectomy, the estimated hazard ratios and median survival times are based on small numbers of survival times. As a result, the standard errors of estimates of such quantities, which have not been given here, will be large.

#### Example 5.10 Chemotherapy in ovarian cancer patients

Following surgical treatment of ovarian cancer, patients may undergo a course of chemotherapy. In a study of two different forms of chemotherapy treat-

ment, Edmunson *et al.* (1979) compared the anti-tumour effects of cyclophosphamide alone and cyclophosphamide combined with adriamycin. The trial involved 26 women with minimal residual disease and who had experienced surgical excision of all tumour masses greater than 2 cm in diameter. Following surgery, the patients were further classified according to whether the residual disease was completely or partially excised. The age of the patient and their performance status were also recorded at the start of the trial. The response variable was the survival time in days following randomisation to one or other of the two chemotherapy treatments. The variables in the data set are therefore as follows:

- Time*: Survival time in days,
- Status*: Event indicator (0 = censored, 1 = uncensored),
- Treat*: Treatment (1 = single, 2 = combined),
- Age*: Age of patient in years,
- Rdisease*: Extent of residual disease (1 = incomplete, 2 = complete),
- Perf*: Performance status (1 = good, 2 = poor).

The data, which were obtained from Therneau (1986), are given in Table 5.6.

In modelling these data, the factors *Treat*, *Rdisease* and *Perf* each have two levels, and will be fitted as variates that take the values given in Table 5.6. This does of course mean that the baseline hazard function is not directly interpretable, since there can be no individual for whom the values of all these variates are zero. From both a computational and interpretive viewpoint, it is more convenient to relocate the values of the variables *Age*, *Rdisease*, *Perf* and *Treat*. If the variable *Age* - 50 is used in place of *Age*, and unity is subtracted from *Rdisease*, *Perf* and *Treat*, the baseline hazard then corresponds to the hazard for an individual of age 50 with incomplete residual disease, good performance status, and who has been allocated to the cyclophosphamide group. However, the original variables will be used in this example.

We begin by identifying which prognostic factors are associated with the survival times of the patients. The values of the statistic  $-2 \log \hat{L}$  on fitting a range of models to these data are given in Table 5.7.

When Weibull models that contain just one of *Age*, *Rdisease* and *Perf* are fitted, we find that both *Age* and *Rdisease* lead to reductions in the value of  $-2 \log \hat{L}$  that are significant at the 5% level. After fitting *Age*, the variables *Rdisease* and *Perf* further reduce  $-2 \log \hat{L}$  by 1.903 and 0.048, respectively, neither of which is significant at the 10% level. Also, when *Age* is added to the model that already includes *Rdisease*, the reduction in  $-2 \log \hat{L}$  is 13.719 on 1 d.f., which is highly significant ( $P < 0.001$ ). This leads us to the conclusion that *Age* is the only prognostic variable that needs to be incorporated in the model.

The term associated with the treatment effect is now added to the model. The value of  $-2 \log \hat{L}$  is then reduced by 2.440 on 1 d.f. This reduction of 2.440 is not quite large enough for it to be significant at the 10% level ( $P = 0.118$ ).

Table 5.6 *Survival times of ovarian cancer patients.*

Patient	Time	Status	Treat	Age	Rdisease	Perf
1	156	1	1	66	2	2
2	1040	0	1	38	2	2
3	59	1	1	72	2	1
4	421	0	2	53	2	1
5	329	1	1	43	2	1
6	769	0	2	59	2	2
7	365	1	2	64	2	1
8	770	0	2	57	2	1
9	1227	0	2	59	1	2
10	268	1	1	74	2	2
11	475	1	2	59	2	2
12	1129	0	2	53	1	1
13	464	1	2	56	2	2
14	1206	0	2	44	2	1
15	638	1	1	56	1	2
16	563	1	2	55	1	2
17	1106	0	1	44	1	1
18	431	1	1	50	2	1
19	855	0	1	43	1	2
20	803	0	1	39	1	1
21	115	1	1	74	2	1
22	744	0	2	50	1	1
23	477	0	1	64	2	1
24	448	0	1	56	1	2
25	353	1	2	63	1	2
26	377	0	2	58	1	1

Table 5.7 *Values of  $-2 \log \hat{L}$  on fitting models to the data in Table 5.6.*

Variables in model	$-2 \log \hat{L}$
none	59.534
<i>Age</i>	43.566
<i>Rdisease</i>	55.382
<i>Perf</i>	58.849
<i>Age</i> , <i>Rdisease</i>	41.663
<i>Age</i> , <i>Perf</i>	43.518
<i>Age</i> , <i>Treat</i>	41.126
<i>Age</i> , <i>Treat</i> , <i>Treat</i> $\times$ <i>Age</i>	39.708

There is therefore only very slight evidence of a difference in the effect of the two chemotherapy treatments on the hazard of death.

For comparison, when *Treat* alone is added to the null model, the value of  $-2 \log \hat{L}$  is reduced from 59.534 to 58.355. This reduction of 1.179 is certainly not significant when compared to percentage points of the chi-squared distribution on 1 d.f. Ignoring *Age* therefore leads to an underestimate of the magnitude of the treatment effect.

To explore whether the treatment difference is consistent over age, the interaction term formed as the product of *Age* and *Treat* is added to the model. On doing so,  $-2 \log \hat{L}$  is only reduced by 1.419. This reduction is nowhere near being significant and so there is no need to include an interaction term in the model.

The variable *Treat* will be retained in the model, since interest centres on the magnitude of the treatment effect. The fitted model for the hazard of death at time  $t$  for the  $i$ th individual is then found to be

$$\hat{h}_i(t) = \exp\{0.144 \text{Age}_i - 1.023 \text{Treat}_i\} \hat{\lambda} \hat{\gamma} t^{\hat{\gamma}-1},$$

where  $\hat{\lambda} = 5.645 \times 10^{-9}$  and  $\hat{\gamma} = 1.822$ . In this model, *Treat* = 1 for cyclophosphamide alone and *Treat* = 2 for the combination of cyclophosphamide with adriamycin. The hazard for a patient on the single treatment, relative to one on the combined treatment, is therefore estimated by

$$\hat{\psi} = \exp\{(-1.023 \times 1) - (-1.023 \times 2)\} = 2.78.$$

This means that a patient receiving the single chemotherapy treatment is nearly three times more likely to die at any given time than a patient on the combined treatment. Expressed in this way, the benefits of the combined chemotherapy treatment sound to be great. However, when account is taken of the inherent variability of the data on which these results are based, this relative hazard is only significantly greater than unity at the 12% level ( $P = 0.118$ ).

The median survival time can be estimated for patients of a given age on a given treatment from the equation

$$\hat{t}(50) = \left\{ \frac{\log 2}{\hat{\lambda} \exp(0.144 \text{Age} - 1.023 \text{Treat})} \right\}^{1/\hat{\gamma}}.$$

For example, a woman aged 60 (*Age* = 60) who is given cyclophosphamide alone (*Treat* = 1) has an estimated median survival time of 423 days, whereas someone of the same age on the combination of the two chemotherapy treatments has an estimated median survival time of 741 days. Confidence intervals for these estimates can be found using the method illustrated in Example 5.6.

## 5.7 The Gompertz proportional hazards model

Although the Weibull model is the most widely used parametric proportional hazards model, the Gompertz model has found application in demography and

the biological sciences. Indeed the distribution was introduced by Gompertz in 1825, as a model for human mortality.

The hazard function of the Gompertz distribution is given by

$$h(t) = \lambda e^{\theta t},$$

for  $0 \leq t < \infty$ , and  $\lambda > 0$ . In the particular case where  $\theta = 0$ , the hazard function has a constant value,  $\lambda$ , and the survival times then have an exponential distribution. The parameter  $\theta$  determines the shape of the hazard function, positive values leading to a hazard function that increases with time. The hazard function can also be expressed as  $h(t) = \exp(\alpha + \theta t)$ , which shows that the log-hazard function is linear in  $t$ . On the other hand, from equation (5.7), the Weibull log-hazard function is linear in  $\log t$ . Like the Weibull hazard function, the Gompertz hazard increases or decreases monotonically.

The survivor function of the Gompertz distribution is given by

$$S(t) = \exp \left\{ \frac{\lambda}{\theta} (1 - e^{\theta t}) \right\},$$

and the corresponding density function is

$$f(t) = \lambda e^{\theta t} \exp \left\{ \frac{\lambda}{\theta} (1 - e^{\theta t}) \right\}.$$

The  $p$ th percentile is such that

$$t(p) = \frac{1}{\theta} \log \left\{ 1 - \frac{\theta}{\lambda} \log \left( \frac{100 - p}{100} \right) \right\},$$

from which the median survival time is

$$t(50) = \frac{1}{\theta} \log \left\{ 1 + \frac{\theta}{\lambda} \log 2 \right\}.$$

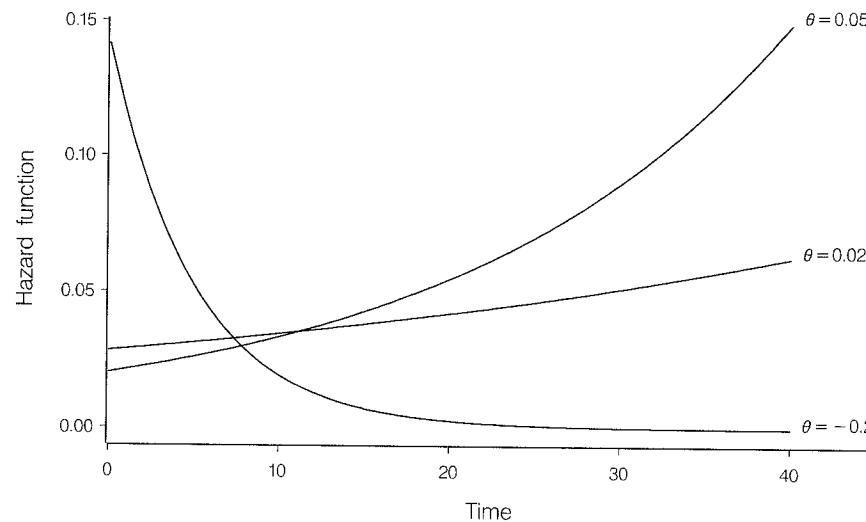
A plot of the Gompertz hazard function for distributions with a median of 20 and  $\theta = -0.2, 0.02$  and  $0.05$  is shown in Figure 5.15. The corresponding values of  $\lambda$  are 0.141, 0.028, and 0.020.

It is straightforward to see that the Gompertz distribution has the proportional hazards property, described in Section 5.4, since if we take  $h_0(t) = \lambda e^{\theta t}$ , then  $\psi h_0(t)$  is also a Gompertz hazard function with parameters  $\psi\lambda$  and  $\theta$ .

The general Gompertz proportional hazards model, for the hazard of death at time  $t$  for the  $i$ th of  $n$  individuals, is expressed as

$$h_i(t) = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}) \lambda e^{\theta t},$$

where  $x_{1i}, x_{2i}, \dots, x_{pi}$  are the values of  $p$  explanatory variables  $X_1, X_2, \dots, X_p$  for the  $i$ th individual,  $i = 1, 2, \dots, n$ , and the  $\beta$ 's,  $\lambda$  and  $\theta$  are unknown parameters. The model can be fitted by maximising the likelihood function given in expression (5.12) or (5.13). The  $\beta$ -coefficients are interpreted as log-hazard ratios, and alternative models are compared using the approach described in Section 5.6. No new principles are involved.



**Figure 5.15** Hazard functions for a Gompertz distribution with a median of 20 and  $\theta = -0.2$ , 0.02 and 0.05.

#### Example 5.11 Chemotherapy in ovarian cancer patients

In Example 5.10 on the survival times of ovarian cancer patients, a Weibull proportional hazards model that contained the variables *Age* and *Treat* was fitted. For comparison, a Gompertz proportional hazards model that contains these two variables is now fitted. Under this model, the fitted hazard function for the  $i$ th patient is

$$\hat{h}_i(t) = \exp\{0.122 \text{Age}_i - 0.848 \text{Treat}_i\} \hat{\lambda} \exp(\hat{\theta}t),$$

where  $\hat{\lambda} = 1.706 \times 10^{-6}$  and  $\hat{\theta} = 0.00138$ . The change in the value of  $-2 \log \hat{L}$  on adding *Treat* to the Gompertz proportional hazards model that contains *Age* alone is now 1.686 ( $P = 0.184$ ). The hazard ratio for the treatment effect, which is now  $\exp(0.848) = 2.34$ , is therefore smaller and less significant under this model than it was for the Weibull model.

#### 5.8 Model choice

One attraction of the proportional hazards model for survival data is that it is not necessary to adopt a specific probability distribution for the survival times. However, when a Weibull distribution is appropriate for the observed survival data, the parametric version of the proportional hazards model provides a more suitable basis for modelling the data.

Diagnostic plots based on the log-cumulative hazard function, described in Section 5.4.1, may throw light on whether the assumption of Weibull survival times is plausible, but as has already been pointed out, this technique is often not informative in the presence of explanatory variables that affect survival

times. In such circumstances, to help choose between the Cox and Weibull proportional hazards models, it can be useful to fit the Cox regression model and examine the shape of the baseline hazard function. The fitted Weibull baseline cumulative hazard function, or the fitted baseline survivor function, can also be compared with the corresponding estimates for the Cox regression model, as described in Section 5.5.3.

A suitable analysis of residuals, to be discussed in Chapter 7, can be used to investigate whether one model fits better than the other. However, it will only be in exceptional circumstances that model-checking diagnostics provide convincing evidence that one or other of the two models is more acceptable.

In general, discrimination between a Cox and a Weibull proportional hazards model will be difficult unless the sample data contain a large number of death times. In cases where there is little to choose between the two models in terms of goodness of fit, the standard errors of the estimated  $\beta$ -parameters in the linear component of the two models can be compared. If those for the Weibull model are substantially smaller than those for the Cox model, the Weibull model would be preferred on grounds of efficiency. On the other hand, if these standard errors are similar, the Cox model is likely to be the model of choice in view of its less restrictive assumptions.

#### 5.9 Further reading

The properties of the exponential, Weibull and Gompertz distributions are presented in Johnson and Kotz (1970). A thorough discussion of the theory of maximum likelihood estimation is included in Barnett (1999) and Cox and Hinkley (1974), and a useful summary of the main results is contained in Hinkley, Reid and Snell (1991). Numerical methods for obtaining maximum likelihood estimates, and the Newton-Raphson procedure in particular, are described by Everitt (1987) and Thisted (1988), for example; see also the description in Section 3.3.3 of Chapter 3. Byar (1982) presents a comparison of the Cox and Weibull proportional hazards models. One other distribution with the proportional hazards property is the *Pareto distribution*. This model is rarely used in practice, but see Davis and Feldstein (1979) for further details.

## Accelerated failure time and other parametric models

---

Although the proportional hazards model finds widespread applicability in the analysis of survival data, there are relatively few probability distributions for the survival times that can be used with this model. Moreover, the distributions that are available, principally the Weibull and Gompertz distributions, lead to hazard functions that increase or decrease monotonically. A model that encompasses a wider range of survival time distributions is the *accelerated failure time model*. In circumstances where the proportional hazards assumption is not tenable, models based on this general family may prove to be fruitful. Again, the Weibull distribution may be adopted for the distribution of survival times in the accelerated failure time model, but some other probability distributions are also available. This chapter therefore begins with a brief survey of alternative distributions for survival data, that may be used in conjunction with an accelerated failure time model. The model itself is then considered in detail in Sections 6.3 to 6.6.

One other general family of survival models, known as the *proportional odds model*, may be useful in some circumstances. This model is described in Section 6.7, and the chapter concludes with a brief reference to other parametric models that are sometimes used in practical applications.

### 6.1 Probability distributions for survival data

The Weibull distribution, described in Section 5.1.2, will not necessarily provide a satisfactory model for survival times in all circumstances, and so alternatives to this distribution need to be considered. Although any continuous distribution for non-negative random variables might be used, the properties of the log-logistic distribution make it a particularly attractive alternative to the Weibull distribution. The lognormal, gamma and inverse Gaussian distributions are sometimes used in accelerated failure time modelling, and so these distributions are also introduced in this section.

#### 6.1.1 *The log-logistic distribution*

One limitation of the Weibull hazard function is that it is a monotonic function of time. However, situations in which the hazard function changes direction can arise. For example, following a heart transplantation, a patient faces an increasing hazard of death over the first ten days or so after the transplant,

while the body adapts to the new organ. The hazard then decreases with time as the patient recovers. In situations such as this, a unimodal hazard function may be appropriate.

A particular form of unimodal hazard is the function

$$h(t) = \frac{e^\theta \kappa t^{\kappa-1}}{1 + e^\theta t^\kappa}, \quad (6.1)$$

for  $0 \leq t < \infty$ ,  $\kappa > 0$ . This hazard function decreases monotonically if  $\kappa \leq 1$ , but if  $\kappa > 1$ , the hazard has a single mode. The survivor function corresponding to the hazard function in equation (6.1) is given by

$$S(t) = \{1 + e^\theta t^\kappa\}^{-1}, \quad (6.2)$$

and the probability density function is

$$f(t) = \frac{e^\theta \kappa t^{\kappa-1}}{(1 + e^\theta t^\kappa)^2}.$$

This is the density of a random variable  $T$  that has a *log-logistic distribution*, with parameters  $\theta, \kappa$ . The distribution is so called because the variable  $\log T$  has a *logistic distribution*, a symmetric distribution whose probability density function is very similar to that of the normal distribution.

The  $p$ th percentile of the log-logistic distribution is

$$t(p) = \left( \frac{p e^{-\theta}}{100 - p} \right)^{1/\kappa},$$

and so the median of the distribution is

$$t(50) = e^{-\theta/\kappa}. \quad (6.3)$$

The hazard functions for log-logistic distributions with a median of 20 and  $\kappa = 0.5, 2.0$  and  $5.0$  are shown in Figure 6.1. The corresponding values of  $\theta$  for these distributions are  $-1.5, -6.0$  and  $-15.0$ , respectively.

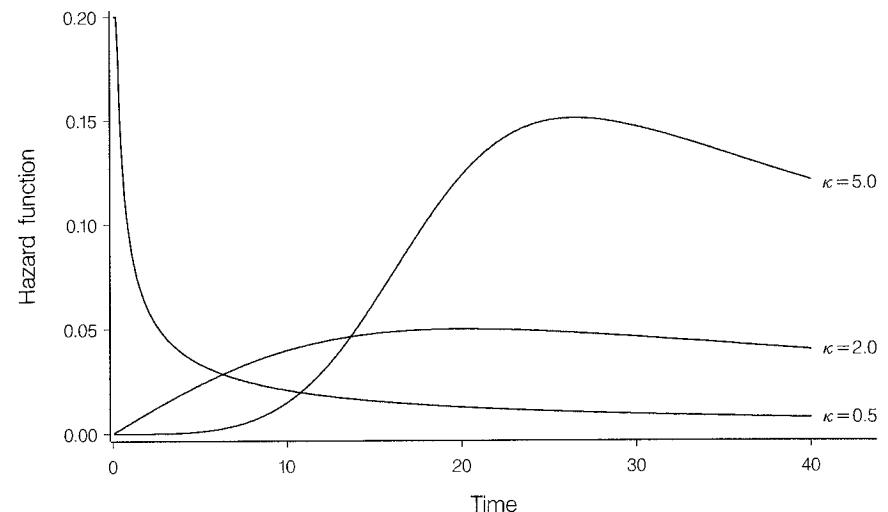
### 6.1.2 The lognormal distribution

The *lognormal distribution* is also defined for random variables that take positive values, and so may be used as a model for survival data. A random variable,  $T$ , is said to have a lognormal distribution, with parameters  $\mu$  and  $\sigma$ , if  $\log T$  has a normal distribution with mean  $\mu$  and variance  $\sigma^2$ . The probability density function of  $T$  is given by

$$f(t) = \frac{1}{\sigma \sqrt{2\pi}} t^{-1} \exp \{-(\log t - \mu)^2 / 2\sigma^2\},$$

for  $0 \leq t < \infty$ ,  $\sigma > 0$ , from which the survivor and hazard functions can be derived. The survivor function of the lognormal distribution is

$$S(t) = 1 - \Phi \left( \frac{\log t - \mu}{\sigma} \right), \quad (6.4)$$



**Figure 6.1** Hazard functions for a log-logistic distribution with a median of 20 and  $\kappa = 0.5, 2.0$  and  $5.0$ .

where  $\Phi(\cdot)$  is the standard normal distribution function, given by

$$\Phi(z) = \frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^z \exp(-u^2/2) du.$$

The  $p$ th percentile of the distribution is then

$$t(p) = \exp \{ \sigma \Phi^{-1}(p/100) + \mu \},$$

where  $\Phi^{-1}(p/100)$ , the  $p$ th percentile of the standard normal distribution, is often called the *probit* of  $p/100$ . In particular, the median survival time under this distribution is simply  $t(50) = e^\mu$ .

The hazard function can be found from the relation  $h(t) = f(t)/S(t)$ . This function is zero when  $t = 0$ , increases to a maximum and then decreases to zero as  $t$  tends to infinity. The fact that the survivor and hazard functions can only be expressed in terms of integrals limits the usefulness of this model. Moreover, in view of the similarity of the normal and logistic distributions, the lognormal model will tend to be very similar to the log-logistic model.

### 6.1.3\* The gamma distribution

The probability density function of a *gamma distribution* with mean  $\rho/\lambda$  and variance  $\rho/\lambda^2$  is such that

$$f(t) = \frac{\lambda^\rho t^{\rho-1} e^{-\lambda t}}{\Gamma(\rho)}, \quad (6.5)$$

for  $0 \leq t < \infty$ ,  $\lambda > 0$ , and  $\rho > 0$ . As for the lognormal distribution, the survivor function of the gamma distribution can only be expressed as an

integral, and we write

$$S(t) = 1 - \Gamma_{\lambda t}(\rho),$$

where  $\Gamma_{\lambda t}(\rho)$  is known as the *incomplete gamma function*, given by

$$\Gamma_{\lambda t}(\rho) = \frac{1}{\Gamma(\rho)} \int_0^{\lambda t} u^{\rho-1} e^{-u} du.$$

The hazard function for the gamma distribution is then  $h(t) = f(t)/S(t)$ . This hazard function increases monotonically if  $\rho > 1$  and decreases if  $\rho < 1$ , and tends to  $\lambda$  as  $t$  tends to  $\infty$ .

When  $\rho = 1$ , the gamma distribution reduces to the exponential distribution described in Section 5.1.1, and so this distribution, like the Weibull distribution, includes the exponential distribution as a special case. Indeed, the gamma distribution is quite similar to the Weibull, and inferences based on either model will often be very similar.

A generalisation of the gamma distribution is actually more useful than the gamma distribution itself, since it includes the Weibull and lognormal distributions as special cases. This model, known as the *generalised gamma distribution*, may therefore be used to discriminate between alternative parametric models for survival data.

The probability density function of the generalised gamma distribution is an extension of the gamma density in equation (6.5), that includes an additional parameter,  $\theta$ , where  $\theta > 0$ , and is defined by

$$f(t) = \frac{\theta \lambda^{\rho\theta} t^{\rho\theta-1} \exp\{-(\lambda t)^\theta\}}{\Gamma(\rho)},$$

for  $0 \leq t < \infty$ . The survivor function for this distribution is again defined in terms of the incomplete gamma function and is given by

$$S(t) = 1 - \Gamma_{(\lambda t)^\theta}(\rho),$$

and the hazard function is again found from  $h(t) = f(t)/S(t)$ . This distribution leads to a wide range of shapes for the hazard function, governed by the parameter  $\theta$ . This parameter is therefore termed the *shape parameter* of the distribution. When  $\rho = 1$ , the distribution becomes the Weibull, when  $\theta = 1$ , the gamma, and as  $\rho \rightarrow \infty$ , the lognormal.

#### 6.1.4\* The inverse Gaussian distribution

The *inverse Gaussian distribution* is a flexible model that has some important theoretical properties. The probability density function of the distribution which has mean  $\mu$  and scale parameter  $\lambda$  is given by

$$f(t) = \left( \frac{\lambda}{2\pi t^3} \right)^{\frac{1}{2}} \exp \left\{ \frac{\lambda(t-\mu^2)}{2\mu^2 t} \right\},$$

for  $0 \leq t < \infty$ , and  $\lambda > 0$ . The corresponding survivor function is

$$S(t) = \Phi \{ (1 - t\mu^{-1}) \vee (\lambda t^{-1}) \} - \exp(2\lambda/\mu) \Phi \{ -(1 + t\mu^{-1}) \vee (\lambda t^{-1}) \},$$

and the hazard function is found from the ratio of the density and survivor functions. However, the complicated form of the survivor function makes this distribution difficult to work with.

## 6.2 Exploratory analyses

When the number of observations in a single sample is reasonably large, an empirical estimate of the hazard function could be obtained using the method described in Section 2.3.1. A plot of the estimated hazard function may then suggest a suitable parametric form for the hazard function. For example, if the hazard plot is found to be unimodal, a log-logistic distribution could be used for the survival times. When the data base includes a number of explanatory variables, the form of the estimated baseline hazard, cumulative hazard, or survivor functions, for a fitted Cox regression model, may also indicate whether a particular parametric model is suitable, in the manner described in Section 5.5.3 of Chapter 5.

A method for exploring the adequacy of the Weibull model in describing a single sample of survival times was described in Section 5.2. A similar procedure can be used to assess the suitability of the log-logistic distribution. The basic idea is that a transformation of the survivor function is sought, which leads to a straight line plot. From equation (6.2), the odds of surviving beyond time  $t$  are

$$\frac{S(t)}{1 - S(t)} = e^{-\theta} t^{-\kappa},$$

and so the log-odds of survival beyond  $t$  can be expressed as

$$\log \left\{ \frac{S(t)}{1 - S(t)} \right\} = -\theta - \kappa \log t.$$

If the survivor function for the data is estimated using the Kaplan-Meier estimate, and the estimated log-odds of survival beyond  $t$  are plotted against  $\log t$ , a straight line plot will be obtained if a log-logistic model for the survival times is suitable. Estimates of the parameters of the log-logistic distribution,  $\theta$  and  $\kappa$ , can be obtained from the intercept and slope of the straight line plot.

The suitability of other parametric models can be investigated along similar lines. For example, from the survivor function of the lognormal distribution, given in equation (6.4),

$$\Phi^{-1}\{1 - S(t)\} = \frac{\log t - \mu}{\sigma},$$

and so a plot of  $\Phi^{-1}\{1 - \hat{S}(t)\}$  against  $\log t$  should give a straight line, if the lognormal model is appropriate. The slope and intercept of this line provide estimates of  $\sigma^{-1}$  and  $-\mu/\sigma$ , respectively.

#### Example 6.1 Time to discontinuation of the use of an IUD

In Example 5.1, a log-cumulative hazard plot was used to evaluate the fit of the Weibull distribution to the data on times to discontinuation of an IUD,

given in Example 1.1. We now consider whether the log-logistic distribution is appropriate. A plot of  $\log\{\hat{S}(t)/[1 - \hat{S}(t)]\}$  against  $\log t$  for the data on the times to discontinuation of an IUD is shown in Figure 6.2.

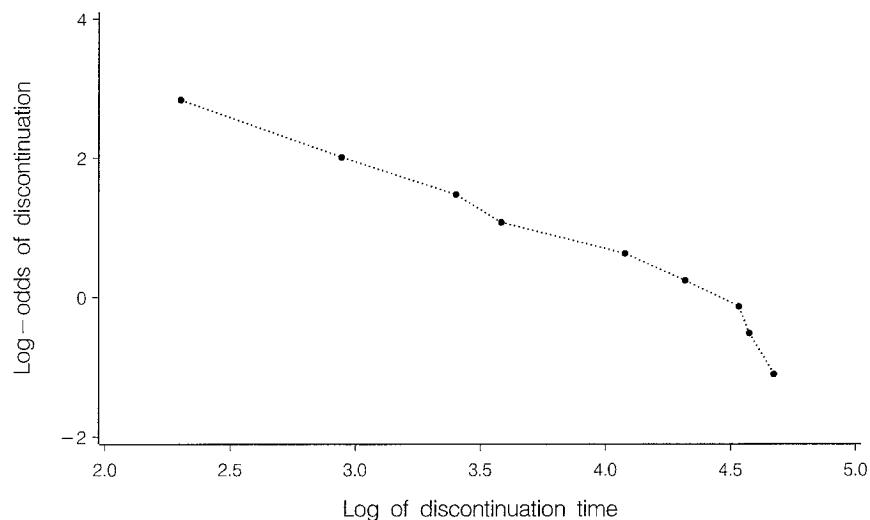


Figure 6.2 A plot of the estimated log-odds of discontinuation after  $t$  against  $\log t$  for the data from Example 1.1.

From this plot, it appears that the relationship between the estimated log-odds of discontinuing use of the contraceptive after time  $t$ , and  $\log t$ , is reasonably straight. This suggests that a log-logistic model could be used to model the observed data.

Notice that there is very little difference in the extent of departures from linearity in the plots in Figures 5.6 and 6.2. This means that either the Weibull distribution or the log-logistic distribution is likely to be satisfactory, even though the estimated hazard function under these two distributions may be quite different. Indeed, when survival data are obtained for a relatively small number of individuals, as in this example, there will often be little to choose between alternative distributional models for the data. The model that is the most convenient for the purpose in hand will then be adopted.

### 6.3 The accelerated failure time model for comparing two groups

The accelerated failure time model is a general model for survival data, in which explanatory variables measured on an individual are assumed to act multiplicatively on the time-scale, and so affect the rate at which an individual proceeds along the time axis. This means that the models can be interpreted in terms of the speed of progression of a disease, an interpretation that has immediate intuitive appeal. Before the general form of the model is presented

in Section 6.4, the model for comparing the survival times of two groups of patients is described in detail.

Suppose that patients are randomised to receive one of two treatments, a standard treatment,  $S$ , or a new treatment,  $N$ . Under an accelerated failure time model, the survival time of an individual on the new treatment is taken to be a multiple of the survival time for an individual on the standard treatment. Thus the effect of the new treatment is to “speed up” or “slow down” the passage of time. Under this assumption, the probability that an individual on the new treatment survives beyond time  $t$  is the probability that an individual on the standard treatment survives beyond time  $t/\phi$ , where  $\phi$  is an unknown positive constant.

Now let  $S_S(t)$  and  $S_N(t)$  be the survivor functions for individuals in the two treatment groups. Then, the accelerated failure time model specifies that

$$S_N(t) = S_S(t/\phi),$$

for any value of the survival time  $t$ . One interpretation of this model is that the lifetime of an individual on the new treatment is  $\phi$  times the lifetime that the individual would have experienced under the standard treatment. The parameter  $\phi$  therefore reflects the impact of the new treatment on the baseline time scale. When the end-point of concern is the death of a patient, values of  $\phi$  less than unity correspond to an acceleration in the time to death of an individual assigned to the new treatment, relative to an individual on the standard treatment. The standard treatment would then be the more suitable in terms of promoting longevity. On the other hand, when the end-point is the recovery from some disease state, values of  $\phi$  less than unity would be found when the effect of the new treatment is to speed up the recovery time. In these circumstances, the new treatment would be superior to the standard. The quantity  $\phi^{-1}$  is therefore termed the *acceleration factor*.

The acceleration factor can also be interpreted in terms of the median survival times of patients on the new and standard treatments,  $t_N(50)$  and  $t_S(50)$ , say. These values are such that  $S_N\{t_N(50)\} = S_S\{t_S(50)\} = 0.5$ . Now, under the accelerated failure time model,  $S_N\{t_N(50)\} = S_S\{t_N(50)/\phi\}$ , and so it follows that  $t_N(50) = \phi t_S(50)$ . In other words, under the accelerated failure time model, the median survival time of a patient on the new treatment is  $\phi$  times that of a patient on the standard treatment. In fact, the same argument can be used for any percentile of the survival time distribution. This means that the  $p$ th percentile of the survival time distribution for a patient on the new treatment,  $t_N(p)$ , is such that  $t_N(p) = \phi t_S(p)$ , where  $t_S(p)$  is the  $p$ th percentile for the standard treatment. This interpretation of the acceleration factor is particularly appealing to clinicians.

From the relationship between the survivor function, probability density function and hazard function given in equation (1.3), the relationship between the density and hazard functions for individuals in the two treatment groups is

$$f_N(t) = \phi^{-1} f_S(t/\phi),$$

and

$$h_N(t) = \phi^{-1} h_S(t/\phi).$$

Now let  $X$  be an indicator variable that takes the value zero for an individual in the group receiving the standard treatment, and unity for one who receives the new treatment. The hazard function for the  $i$ th individual can then be expressed as

$$h_i(t) = \phi^{-x_i} h_0(t/\phi^{x_i}), \quad (6.6)$$

where  $x_i$  is the value of  $X$  for the  $i$ th individual in the study. Putting  $x_i = 0$  in this expression shows that the function  $h_0(t)$  is the hazard function for an individual on the standard treatment. This is again referred to as the baseline hazard function. The hazard function for an individual on the new treatment is then  $\phi^{-1} h_0(t/\phi)$ .

The parameter  $\phi$  must be non-negative, and so it is convenient to set  $\phi = e^\alpha$ . The accelerated failure time model in equation (6.6) then becomes

$$h_i(t) = e^{-\alpha x_i} h_0(t/e^{\alpha x_i}), \quad (6.7)$$

so that the hazard function for an individual on the new treatment is now  $e^{-\alpha} h_0(t/e^\alpha)$ .

### 6.3.1\* Comparison with the proportional hazards model

To illustrate the difference between a proportional hazards model and the accelerated failure time model, again suppose that the survival times of individuals in two groups, Group I and Group II, say, are to be modelled. Further suppose that for the individuals in Group I, the hazard function is given by

$$h_0(t) = \begin{cases} 0.5 & \text{if } t \leq 1, \\ 1.0 & \text{if } t > 1, \end{cases}$$

where the time-scale is measured in months. This type of hazard function arises from a *piecewise exponential model*, since a constant hazard in each time interval implies exponentially distributed survival times, with different means, in each interval. This model provides a simple way of representing a variable hazard function, and may be appropriate in situations where there is a constant short-term risk that increases abruptly after a threshold time.

Now let  $h_P(t)$  and  $h_A(t)$  denote the hazard functions for individuals in Group II under a proportional hazards model and an accelerated failure time model, respectively. Consequently, we may write

$$h_P(t) = \psi h_0(t),$$

and

$$h_A(t) = \phi^{-1} h_0(t/\phi),$$

for the two hazard functions. Using the result  $S(t) = \exp\{-\int_0^t h(u) du\}$ , the baseline survivor function is

$$S_0(t) = \begin{cases} e^{-0.5t} & \text{if } t \leq 1, \\ e^{-0.5-(t-1)} & \text{if } t > 1. \end{cases}$$

Since  $S_0(t) > 0.61$  if  $t < 1$ , the median occurs in the second part of the survivor function and is when  $\exp\{-0.5 - (t-1)\} = 0.5$ . The median survival time for those in Group I is therefore 1.19 months.

The survivor functions for the individuals in Group II under the two models are

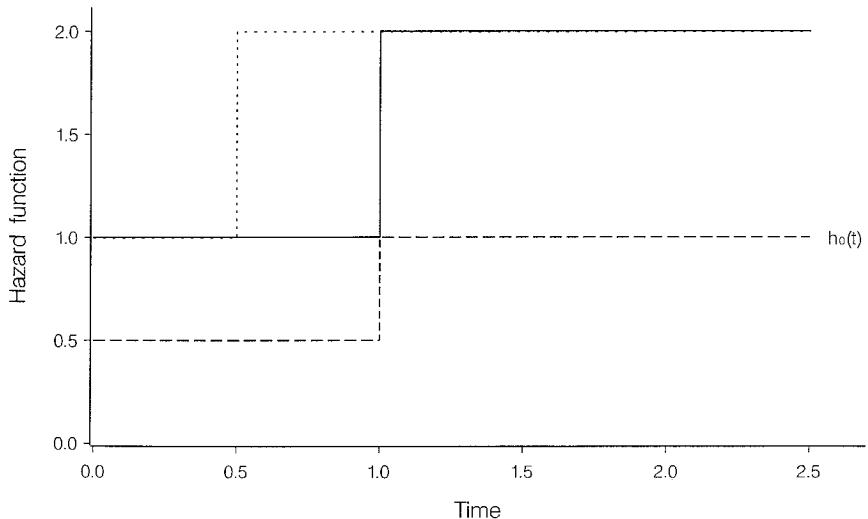
$$S_P(t) = [S_0(t)]^\psi,$$

and

$$S_A(t) = S_0(t/\phi),$$

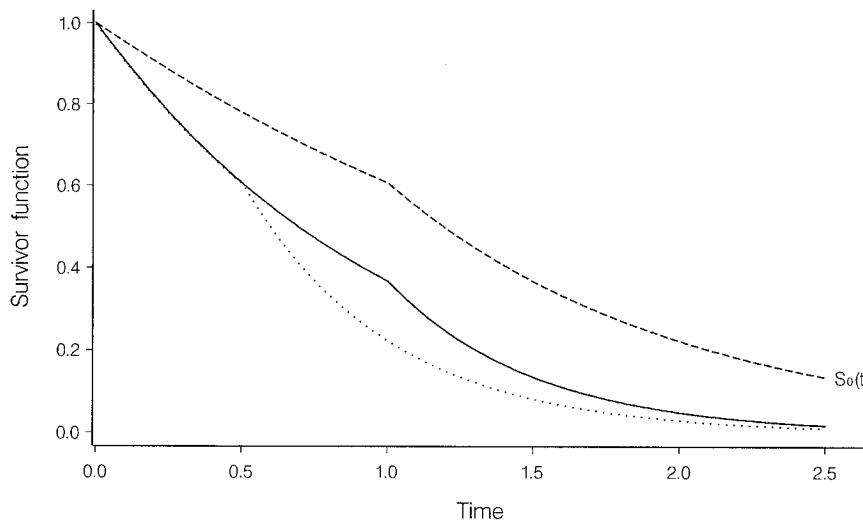
respectively.

To illustrate the difference between the hazard functions under proportional hazards and accelerated failure time models, consider the particular case where  $\psi = \phi^{-1} = 2.0$ . The median survival time for individuals in Group II is 0.69 months under the proportional hazards model, and 0.60 months under the accelerated failure time model. The hazard functions for the two groups under both models are shown in Figure 6.3 and the corresponding survivor functions are shown in Figure 6.4.



**Figure 6.3** The hazard functions for individuals in Group I,  $h_0(t)$ , and in Group II under (a) a proportional hazards model (—) and (b) an accelerated failure time model (···).

Under the accelerated failure time model, the increase in the hazard for Group II from 1.0 to 2.0 occurs sooner than under the proportional hazards model. The “kink” in the survivor function also occurs earlier under the accelerated failure time model.



**Figure 6.4** The survivor functions for individuals in Group I,  $S_0(t)$ , and in Group II, under (a) a proportional hazards model (—) and (b) an accelerated failure time model (···).

### 6.3.2 The percentile-percentile plot

The *percentile-percentile plot*, also known as the *quantile-quantile plot* or the *Q-Q plot*, provides an exploratory method for assessing the validity of an accelerated failure time model for two groups of survival data. Recall that the  $p$ th percentile of a distribution is the value  $t(p)$ , which is such that the estimated survivor function at time  $t(p)$  is  $1 - (p/100)$ , for any value of  $p$  in the interval  $(0, 100)$ . The  $p$ th percentile is therefore such that

$$t(p) = S^{-1} \left( \frac{100-p}{100} \right).$$

Now let  $t_0(p)$  and  $t_1(p)$  be the  $p$ th percentiles estimated from the survivor functions of the two groups of survival data. The values of  $p$  might be taken to be  $10, 20, \dots, 90$ , so long as the number of observations in each of the two groups is not too small. The percentiles of the two groups may therefore be expressed as

$$t_0(p) = S_0^{-1} \left( \frac{100-p}{100} \right), \quad t_1(p) = S_1^{-1} \left( \frac{100-p}{100} \right),$$

where  $S_0(t)$  and  $S_1(t)$  are the survivor functions for the two groups. It then follows that

$$S_1 \{t_1(p)\} = S_0 \{t_0(p)\}, \quad (6.8)$$

for any given value of  $p$ .

Under the accelerated failure time model,  $S_1(t) = S_0(t/\phi)$ , and so the  $p$ th

percentile for the second group,  $t_1(p)$ , is such that

$$S_1 \{t_1(p)\} = S_0 \{t_1(p)/\phi\}.$$

Using equation (6.8),

$$S_0 \{t_0(p)\} = S_0 \{t_1(p)/\phi\},$$

and hence

$$t_0(p) = \phi^{-1} t_1(p).$$

Now let  $\hat{t}_0(p)$ ,  $\hat{t}_1(p)$  be the estimated percentiles in the two groups, so that

$$\hat{t}_0(p) = \hat{S}_0^{-1} \left( \frac{100-p}{100} \right), \quad \hat{t}_1(p) = \hat{S}_1^{-1} \left( \frac{100-p}{100} \right).$$

A plot of the quantity  $\hat{t}_0(p)$  against  $\hat{t}_1(p)$ , for suitably chosen values of  $p$ , should give a straight line through the origin if the accelerated failure time model is appropriate. The slope of this line will be an estimate of the acceleration factor,  $\phi^{-1}$ . This plot may therefore be used in an exploratory assessment of the adequacy of the accelerated failure time model. In this sense, it is an analogue of the log-cumulative hazard plot, used in Section 4.4.1 to examine the validity of the proportional hazards model.

### Example 6.2 Prognosis for women with breast cancer

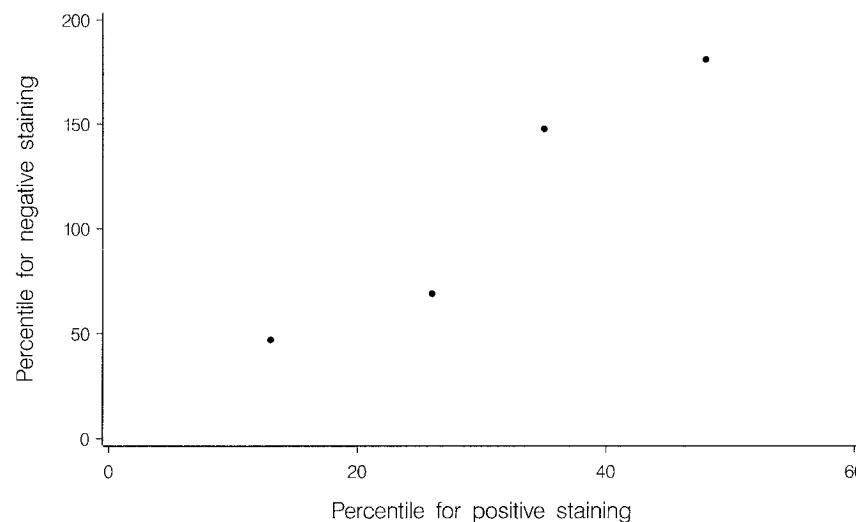
In this example, the data on the survival times of women with breast tumours that were negatively or positively stained, originally given as Example 1.2 in Chapter 1, is used to illustrate the percentile-percentile plot. The percentiles of the distribution of the survival times in each of the two groups can be estimated from the Kaplan-Meier estimate of the respective survivor functions. These are given in Table 6.1.

**Table 6.1** Estimated percentiles of the distributions of survival times for women with tumours that were positively or negatively stained.

Percentile	Negative staining	Positive staining
10	47	13
20	69	26
30	148	35
40	181	48
50	—	61
60	—	113
70	—	143
80	—	—
90	—	—

The relatively small numbers of death times, and the censoring pattern in the data from the two groups of women, mean that not all of the percentiles can be estimated. The percentile-percentile plot will therefore have just four

pairs of points. For illustration, this is shown in Figure 6.5. The points fall on a line that is reasonably straight, suggesting that the accelerated failure time model would not be inappropriate. However, this conclusion must be regarded with some caution in view of the limited number of points in the graph.



**Figure 6.5** Percentile-percentage plot for the data on the survival times of breast cancer patients.

The slope of a straight line drawn through the points in Figure 6.5 is approximately equal to 3, which is a rough estimate of the acceleration factor. The interpretation of this is that for women whose tumours were positively stained, the disease process is speeded up by a factor of three, relative to those whose tumours were negatively stained. We can also say that the median survival time for women with negatively stained tumours is estimated to be three times that of women with positively stained tumours.

#### 6.4 The general accelerated failure time model

The accelerated failure time model in equation (6.7) can be generalised to the situation where the values of  $p$  explanatory variables have been recorded for each individual in a study. According to the general accelerated failure time model, the hazard function of the  $i$ th individual at time  $t$ ,  $h_i(t)$ , is then such that

$$h_i(t) = e^{-\eta_i} h_0(t/e^{\eta_i}), \quad (6.9)$$

where

$$\eta_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}$$

is the linear component of the model, in which  $x_{ji}$  is the value of the  $j$ th explanatory variable,  $X_j$ ,  $j = 1, 2, \dots, p$ , for the  $i$ th individual,  $i = 1, 2, \dots, n$ . As in the proportional hazards model, the baseline hazard function,  $h_0(t)$ ,

is the hazard of death at time  $t$  for an individual for whom the values of the  $p$  explanatory variables are all equal to zero. The corresponding survivor function for the  $i$ th individual is

$$S_i(t) = S_0\{t/\exp(\eta_i)\},$$

where  $S_0(t)$  is the baseline survivor function.

Parametric accelerated failure time models are unified by the adoption of a log-linear representation of the model, described in the sequel. This representation shows that the accelerated failure time model for survival data is closely related to the general linear model used in regression analysis. Moreover, this form of the model is adopted by most computer software packages for accelerated failure time modelling.

##### 6.4.1\* Log-linear form of the accelerated failure time model

Consider a *log-linear model* for the random variable  $T_i$ , associated with the lifetime of the  $i$ th individual in a survival study, according to which

$$\log T_i = \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi} + \sigma \epsilon_i. \quad (6.10)$$

In this model,  $\alpha_1, \alpha_2, \dots, \alpha_p$  are the unknown coefficients of the values of  $p$  explanatory variables,  $X_1, X_2, \dots, X_p$ , and  $\mu, \sigma$  are two further parameters, known as the *intercept* and *scale parameter*, respectively. The quantity  $\epsilon_i$  is a random variable used to model the deviation of the values of  $\log T_i$  from the linear part of the model, and  $\epsilon_i$  is assumed to have a particular probability distribution. In this formulation of the model, the  $\alpha$ -parameters reflect the effect that each explanatory variable has on the survival times; positive values suggest that the survival time increases with increasing values of the explanatory variable, and vice versa.

To show the relationship between this representation of the model and that in equation (6.9), consider the survivor function of  $T_i$ , the random variable associated with the survival time of the  $i$ th individual. Using equation (6.10), this is given by

$$S_i(t) = P(T_i \geq t) = P\{\exp(\mu + \alpha' x_i + \sigma \epsilon_i) \geq t\},$$

where  $\alpha' x_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}$ .

Now,  $S_i(t)$  can be written in the form

$$S_i(t) = P\{\exp(\mu + \sigma \epsilon_i) \geq t / \exp(\alpha' x_i)\},$$

and the baseline survivor function,  $S_0(t)$ , the survivor function of an individual for whom  $x = \mathbf{0}$ , is

$$S_0(t) = P\{\exp(\mu + \sigma \epsilon_i) \geq t\}.$$

It then follows that

$$S_i(t) = S_0\{t / \exp(\alpha' x_i)\}, \quad (6.11)$$

which is the general form of the survivor function for the  $i$ th individual in an accelerated failure time model. In this version of the model, the acceleration

factor is  $\exp(-\boldsymbol{\alpha}' \mathbf{x}_i)$  for the  $i$ th individual. The corresponding relationship between the hazard functions is obtained using equation (1.4) of Chapter 1. Specifically, taking logarithms of both sides of equation (6.11), multiplying by  $-1$ , and differentiating with respect to  $t$ , leads to

$$h_i(t) = \exp(-\boldsymbol{\alpha}' \mathbf{x}_i) h_0\{t/\exp(\boldsymbol{\alpha}' \mathbf{x}_i)\},$$

which is the model in equation (6.9) with  $\eta_i = \boldsymbol{\alpha}' \mathbf{x}_i$ .

The log-linear formulation of the model can also be used to give a general form of the survivor function for the  $i$ th individual, which is

$$S_i(t) = P(T_i \geq t) = P(\log T_i \geq \log t).$$

From equation (6.10),

$$\begin{aligned} S_i(t) &= P(\mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi} + \sigma \epsilon_i \geq \log t), \\ &= P\left(\epsilon_i \geq \frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma}\right). \end{aligned} \quad (6.12)$$

If we now write  $S_{\epsilon_i}(\epsilon)$  for the survivor function of the random variable  $\epsilon_i$  in the log-linear model of equation (6.10), the survivor function of the  $i$ th individual can, from equation (6.12), be expressed as

$$S_i(t) = S_{\epsilon_i}\left(\frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma}\right). \quad (6.13)$$

This result shows how the survivor function for  $T_i$  can be found from the survivor function of the distribution of  $\epsilon_i$ . The result also demonstrates that an accelerated failure time model can be derived from many probability distributions for  $\epsilon_i$ , although some are more tractable than others.

A general expression for the  $p$ th percentile of the distribution of survival times also follows from the results in this section. The  $p$ th percentile for the  $i$ th individual,  $t_i(p)$ , is given by

$$S_i\{t_i(p)\} = \frac{100-p}{100},$$

and using equation (6.12),

$$P\left(\epsilon_i \geq \frac{\log t_i(p) - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma}\right) = \frac{100-p}{100}.$$

If  $\epsilon_i(p)$  is used to denote the  $p$ th percentile of the distribution of  $\epsilon_i$ , then

$$S_{\epsilon_i}\{\epsilon_i(p)\} = P\{\epsilon_i \geq \epsilon_i(p)\} = \frac{100-p}{100}.$$

Consequently,

$$\epsilon_i(p) = \frac{\log t_i(p) - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma},$$

and so

$$t_i(p) = \exp\{\sigma \epsilon_i(p) + \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}\} \quad (6.14)$$

is the  $p$ th percentile of the distribution of survival times for the  $i$ th individual. Note that the percentile in equation (6.14) can be written in the form

$$t_i(p) = \exp(\alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}) t_0(p),$$

where  $t_0(p)$  is the  $p$ th percentile for a baseline individual for whom all explanatory variables take the value zero. This confirms that the  $\alpha$ -coefficients can be interpreted in terms of the effect of the explanatory variables on a particular percentile of the distribution of survival times.

The cumulative hazard function of the distribution of  $T_i$  is given by  $H_i(t) = -\log S_i(t)$ , and from equation (6.13),

$$\begin{aligned} H_i(t) &= -\log S_{\epsilon_i}\left(\frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma}\right), \\ &= H_{\epsilon_i}\left(\frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma}\right), \end{aligned} \quad (6.15)$$

where  $H_{\epsilon_i}(\epsilon) = -\log S_{\epsilon_i}(\epsilon)$  is the cumulative hazard function of  $\epsilon_i$ . The corresponding hazard function, found by differentiating  $H_i(t)$  in equation (6.15) with respect to  $t$ , is

$$h_i(t) = \frac{1}{\sigma t} h_{\epsilon_i}\left(\frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma}\right), \quad (6.16)$$

where  $h_{\epsilon_i}(\epsilon)$  is the hazard function of the distribution of  $\epsilon_i$ .

The distributions of  $\epsilon_i$  that are most often used in accelerated failure time modelling are such that their percentiles,  $\epsilon_i(p)$ , have a simple form. Models based on such distributions are described in the following section.

## 6.5 Parametric accelerated failure time models

Particular choices for the distribution of  $\epsilon_i$  in the log-linear formulation of the accelerated failure time model, described in Section 6.4.1, lead to distributions for the random variable associated with the survival time of the  $i$ th individual. But the representation of the model in equation (6.10) invariably leads to different parameterisations of the models from those given in Sections 5.1 and 6.1. Parametric accelerated failure time models based on the Weibull, logistic and lognormal distributions for the survival times are most commonly used in practice, and so these models are described in detail, and summarised in Section 6.5.4.

### 6.5.1 The Weibull accelerated failure time model

Suppose that survival times are assumed to have a Weibull distribution with scale parameter  $\lambda$  and shape parameter  $\gamma$ , written  $W(\lambda, \gamma)$ , so that the baseline hazard function is

$$h_0(t) = \lambda \gamma t^{\gamma-1}.$$

The hazard function for the  $i$ th individual is then, from equation (6.9), given by

$$h_i(t) = e^{-\eta_i} \lambda \gamma (e^{-\eta_i} t)^{\gamma-1} = (e^{-\eta_i})^\gamma \lambda \gamma t^{\gamma-1},$$

so that the survival time of this individual has a  $W(\lambda e^{-\gamma\eta_i}, \gamma)$  distribution. The Weibull distribution is therefore said to possess the *accelerated failure time property*. Indeed, this is the only probability distribution that has both the proportional hazards and accelerated failure time properties.

Because the Weibull distribution has both the proportional hazards property and the accelerated failure time property, there is a direct correspondence between the parameters under the two models. If the baseline hazard function is the hazard function of a  $W(\lambda, \gamma)$  distribution, the survival times under the general proportional hazards model in equation (5.35) of Chapter 5 have a  $W(\lambda \exp(\beta' \mathbf{x}_i), \gamma)$  distribution, while those under the accelerated failure time model have a  $W(\lambda \exp(-\gamma \alpha' \mathbf{x}_i), \gamma)$  distribution. It then follows that when the coefficients of the explanatory variables in the linear component of the accelerated failure time model are multiplied by  $-\gamma$ , we get the corresponding  $\beta$ -coefficients in the proportional hazards model. In the particular case of comparing two groups, an acceleration factor of  $\phi^{-1} = e^{-\alpha}$  under the accelerated failure time model corresponds to a hazard ratio of  $\phi^{-\gamma} = e^{-\gamma\alpha}$  in a proportional hazards model.

In terms of the log-linear representation of the model in equation (6.10), if  $T_i$  has a Weibull distribution, then  $\epsilon_i$  does in fact have a type of extreme value distribution known as the *Gumbel distribution*. This is an asymmetric distribution with survivor function given by

$$S_{\epsilon_i}(\epsilon) = \exp(-e^\epsilon),$$

for  $-\infty < \epsilon < \infty$ . The cumulative hazard and hazard functions of this distribution are given by  $H_{\epsilon_i}(\epsilon) = e^\epsilon$ , and  $h_{\epsilon_i}(\epsilon) = e^\epsilon$ , respectively.

To show that the random variable  $T_i = \exp(\mu + \alpha' \mathbf{x}_i + \sigma \epsilon_i)$  has a Weibull distribution, from equation (6.13), the survivor function of  $T_i$  is given by

$$S_i(t) = \exp \left\{ -\exp \left( \frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma} \right) \right\}. \quad (6.17)$$

This can be expressed in the form

$$S_i(t) = \exp \left( -\lambda_i t^{1/\sigma} \right),$$

where

$$\lambda_i = \exp \{-(\mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi})/\sigma\},$$

which, from equation (5.8) of Chapter 5, is the survivor function of a Weibull distribution with scale parameter  $\lambda_i$ , and shape parameter  $\sigma^{-1}$ . Consequently, equation (6.17) is the accelerated failure time representation of the survivor function of the Weibull model described in Section 5.5 of Chapter 5.

The cumulative hazard and hazard functions for the Weibull accelerated failure time model can be found directly from the survivor function in equation (6.17), or from  $H_{\epsilon_i}(\epsilon)$  and  $h_{\epsilon_i}(\epsilon)$ , using the general results in equa-

tions (6.15) and (6.16). We find that the cumulative hazard function is

$$H_i(t) = -\log S_i(t) = \exp \left( \frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma} \right),$$

which can also be expressed as  $\lambda_i t^{1/\sigma}$ , and the hazard function is given by

$$h_i(t) = \frac{1}{\sigma t} \exp \left( \frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma} \right), \quad (6.18)$$

$$\text{or } h_i(t) = \lambda_i \sigma^{-1} t^{\sigma^{-1}-1}.$$

We now reconcile this form of the model with that for the Weibull proportional hazards model. From equation (5.37) of Chapter 5, the survivor function for the  $i$ th individual is

$$S_i(t) = \exp \{ -\exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}) \lambda t^\gamma \}, \quad (6.19)$$

in which  $\lambda$  and  $\gamma$  are the parameters of the Weibull baseline hazard function. There is a direct correspondence between equation (6.17) and equation (6.19), in the sense that

$$\lambda = \exp(-\mu/\sigma), \quad \gamma = \sigma^{-1}, \quad \beta_j = -\alpha_j/\sigma,$$

for  $j = 1, 2, \dots, p$ . We therefore deduce that the log-linear model where

$$\log T_i = \frac{1}{\gamma} \{ -\log \lambda - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_p x_{pi} + \epsilon_i \},$$

and in which  $\epsilon_i$  has a Gumbel distribution, provides an alternative representation of the Weibull proportional hazards model.

In this form of the model, the  $p$ th percentile of the survival time distribution for the  $i$ th individual is the value  $t_i(p)$ , which is such that  $S_i\{t_i(p)\} = 1 - (p/100)$ , where  $S_i(t)$  is as given in equation (6.17). Straightforward algebra leads to the result that

$$t_i(p) = \exp \left[ \sigma \log \left\{ -\log \left( \frac{100-p}{100} \right) \right\} + \mu + \alpha' \mathbf{x}_i \right] \quad (6.20)$$

for that individual. Equivalently, the  $p$ th percentile of the distribution of  $\epsilon_i$ ,  $\epsilon_i(p)$ , is such that

$$\exp \left\{ -e^{\epsilon_i(p)} \right\} = \frac{100-p}{100},$$

so that

$$\epsilon_i(p) = \log \left\{ -\log \left( \frac{100-p}{100} \right) \right\},$$

and the general result in equation (6.14) leads directly to equation (6.20).

The survivor function and hazard function of the Weibull model follow from equations (6.17) and (6.18), and equation (6.20) enables percentiles to be estimated directly.

### 6.5.2 The log-logistic accelerated failure time model

Now suppose that the survival times have a log-logistic distribution. If the baseline hazard function in the general accelerated failure time model in equation (6.9) is derived from a log-logistic distribution with parameters  $\theta, \kappa$ , this function is given by

$$h_0(t) = \frac{e^\theta \kappa t^{\kappa-1}}{1 + e^\theta t^\kappa}.$$

Under the accelerated failure time model, the hazard of death at time  $t$  for the  $i$ th individual is

$$h_i(t) = e^{-\eta_i} h_0(e^{-\eta_i} t),$$

where  $\eta_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}$  is a linear combination of the values of  $p$  explanatory variables for the  $i$ th individual. Consequently,

$$h_i(t) = \frac{e^{-\eta_i} e^\theta \kappa (e^{-\eta_i} t)^{\kappa-1}}{1 + e^\theta (e^{-\eta_i} t)^\kappa},$$

that is,

$$h_i(t) = \frac{e^{\theta - \kappa \eta_i} \kappa t^{\kappa-1}}{1 + e^{\theta - \kappa \eta_i} t^\kappa}.$$

It then follows that the survival time for the  $i$ th individual also has a log-logistic distribution with parameters  $\theta - \kappa \eta_i$  and  $\kappa$ . The log-logistic distribution therefore has the accelerated failure time property. However, this distribution does not have the proportional hazards property.

The log-linear form of the accelerated failure time model in equation (6.10) also provides a representation of the log-logistic distribution. Suppose that in this formulation,  $\epsilon_i$  now has a logistic distribution with zero mean and variance  $\pi^2/3$ , so that the survivor function of  $\epsilon_i$  is

$$S_{\epsilon_i}(\epsilon) = \frac{1}{1 + e^\epsilon}.$$

Using equation (6.13), the survivor function of  $T_i$  is then

$$S_i(t) = \left\{ 1 + \exp \left( \frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma} \right) \right\}^{-1}. \quad (6.21)$$

From equation (6.2), the survivor function of  $T_i$ , when  $T_i$  has a log-logistic distribution with parameters  $\theta - \kappa \eta_i, \kappa$ , where  $\eta_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}$ , is

$$S_i(t) = \frac{1}{1 + e^{\theta - \kappa \eta_i} t^\kappa}.$$

On comparing this expression with that for the survivor function in equation (6.21), we see that the parameters  $\theta$  and  $\kappa$  can be expressed in terms of  $\mu$  and  $\sigma$ . Specifically,

$$\theta = -\mu/\sigma, \quad \kappa = \sigma^{-1},$$

and this shows that the accelerated failure time model with log-logistic survival times can also be formulated in terms of a log-linear model. This is the form of the model that is usually adopted by computer software, and so

computer-based parameter estimates are usually estimates of  $\mu$  and  $\sigma$ , rather than  $\theta$  and  $\kappa$ .

The cumulative hazard and hazard functions of the distribution of  $\epsilon_i$  are such that

$$H_{\epsilon_i}(\epsilon) = \log(1 + e^\epsilon),$$

and

$$h_{\epsilon_i}(\epsilon) = (1 + e^{-\epsilon})^{-1},$$

respectively. Equations (6.15) and (6.16) may then be used to obtain the cumulative hazard, and hazard function, of  $T_i$ . In particular, the hazard function for the  $i$ th individual is

$$h_i(t) = \frac{1}{\sigma t} \left\{ 1 + \exp \left[ - \left( \frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma} \right) \right] \right\}^{-1}. \quad (6.22)$$

Estimates of quantities such as the acceleration factor, or the median survival time, can be obtained directly from the estimates of  $\mu, \sigma$  and the  $\alpha_j$ 's. For example, the acceleration factor for the  $i$ th individual is  $\exp\{-(\alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi})\}$  and the  $p$ th percentile of the survival time distribution, from equation (6.21), or the general result in equation (6.14), is

$$t_i(p) = \exp \left\{ \sigma \log \left( \frac{p}{100-p} \right) + \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi} \right\}.$$

The median survival time is simply

$$t_i(50) = \exp \{ \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi} \}, \quad (6.23)$$

and so an estimate of the median can be straightforwardly obtained from the estimated values of the parameters in the model.

### 6.5.3 The lognormal accelerated failure time model

If the survival times are assumed to have a lognormal distribution, the baseline survivor function is given by

$$S_0(t) = 1 - \Phi \left( \frac{\log t - \mu}{\sigma} \right),$$

where  $\mu$  and  $\sigma$  are two unknown parameters. Under the accelerated failure time model, the survivor function for the  $i$ th individual, is then

$$S_i(t) = S_0(e^{-\eta_i} t),$$

where  $\eta_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}$  is a linear combination of the values of  $p$  explanatory variables for the  $i$ th individual. Therefore,

$$S_i(t) = 1 - \Phi \left( \frac{\log t - \eta_i - \mu}{\sigma} \right), \quad (6.24)$$

which is the survivor function for an individual whose survival times have a lognormal distribution with parameters  $\mu + \eta_i$  and  $\sigma$ . The lognormal distribution therefore has the accelerated failure time property.

In the log-linear formulation of the model, the random variable associated with the survival time of the  $i$ th individual has a lognormal distribution if  $\log T_i$  is normally distributed. We therefore take  $\epsilon_i$  in equation (6.10) to have a standard normal distribution, so that the survivor function of  $\epsilon_i$  is

$$S_{\epsilon_i}(\epsilon) = 1 - \Phi(\epsilon).$$

The cumulative hazard, and hazard function, of  $\epsilon_i$  are

$$H_{\epsilon_i}(\epsilon) = -\log \{1 - \Phi(\epsilon)\},$$

and

$$h_{\epsilon_i}(\epsilon) = \frac{f_{\epsilon_i}(\epsilon)}{S_{\epsilon_i}(\epsilon)},$$

respectively, where  $f_{\epsilon_i}(\epsilon)$  is the density function of a standard normal random variable, given by

$$f_{\epsilon_i}(\epsilon) = \frac{1}{\sqrt{(2\pi)}} \exp(-\epsilon^2/2).$$

The random variable  $T_i$ , in the general accelerated failure time model, then has a lognormal distribution with parameters  $\mu + \boldsymbol{\alpha}' \mathbf{x}_i$  and  $\sigma$ . The survivor function of  $T_i$  is as given in equation (6.24), and the hazard function is found from equation (6.16).

The  $p$ th percentile of the distribution of  $T_i$ , from equation (6.14), is

$$t_i(p) = \exp \{ \sigma \Phi^{-1}(p/100) + \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi} \},$$

and, in particular,  $t(50) = \exp(\mu + \boldsymbol{\alpha}' \mathbf{x}_i)$  is the median survival time for the  $i$ th individual.

#### 6.5.4 Summary

It is convenient to summarise the models and results that have been described in this section, so that the different parameterisations of the distributions used in accelerated failure time models can clearly be seen.

The general accelerated failure time model for the survival time of the  $i$ th of  $n$  individuals, for whom  $x_{1i}, x_{2i}, \dots, x_{pi}$  are the values of  $p$  explanatory variables,  $X_1, X_2, \dots, X_p$ , is such that the random variable associated with the survival time,  $T_i$ , can be expressed in the form

$$\log T_i = \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi} + \sigma \epsilon_i.$$

Particular distributions for  $T_i$  are derived from assumptions about the distribution of  $\epsilon_i$  in this model. The survivor function and hazard function of the distributions of  $\epsilon_i$ , that lead to commonly used accelerated failure time models for the survival times, are summarised in Table 6.2.

The cumulative hazard function of  $\epsilon_i$  is found from  $H_{\epsilon_i}(\epsilon) = -\log S_{\epsilon_i}(\epsilon)$ , and, if desired, the density function of  $\epsilon_i$  is  $f_{\epsilon_i}(\epsilon) = h_{\epsilon_i}(\epsilon) S_{\epsilon_i}(\epsilon)$ . From the survivor and hazard function of  $\epsilon_i$ , the survivor and hazard function of  $T_i$  can

**Table 6.2** Summary of parametric accelerated failure time models.

Distribution of $T_i$	$S_{\epsilon_i}(\epsilon)$	$h_{\epsilon_i}(\epsilon)$	Percentile, $\epsilon_i(p)$
Exponential	$\exp(-e^\epsilon)$	$e^\epsilon$	$\log \left\{ -\log \left( \frac{100-p}{100} \right) \right\}$
Weibull	$\exp(-e^\epsilon)$	$e^\epsilon$	$\log \left\{ -\log \left( \frac{100-p}{100} \right) \right\}$
Log-logistic	$(1 + e^\epsilon)^{-1}$	$(1 + e^{-\epsilon})^{-1}$	$\log \left( \frac{p}{100-p} \right)$
Lognormal	$1 - \Phi(\epsilon)$	$\frac{\exp(-\epsilon^2/2)}{\{1 - \Phi(\epsilon)\}\sqrt{(2\pi)}}$	$\Phi^{-1}(p/100)$

be found from

$$S_i(t) = S_{\epsilon_i} \left( \frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma} \right),$$

and

$$h_i(t) = \frac{1}{\sigma t} h_{\epsilon_i} \left( \frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma} \right),$$

results that were first given in equations (6.13) and (6.16), respectively.

The  $p$ th percentile of the distribution of  $\epsilon_i$  is also given in Table 6.2, from which  $t_i(p)$ , the  $p$ th percentile of the survival times for the  $i$ th individual, can be found from

$$t_i(p) = \exp \{ \sigma \epsilon_i(p) + \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi} \},$$

given as equation (6.14).

The log-linear representation of the Weibull and log-logistic models leads to parameterisations of the survival time distributions that differ from those used in Sections 5.1 and 6.1, when the distributions were first presented. The link between the two sets of parameters is summarised in Table 6.3, which includes the number of the equation that gives the survivor function of each distribution in terms of the original parameters.

**Table 6.3** Summary of the parameterisation of accelerated failure time models.

Distribution of $T_i$	Equation number	Parameterisation of survivor function
Exponential	(5.5)	$\lambda = e^{-\mu}, \gamma = 1 (\sigma = 1)$
Weibull	(5.8)	$\lambda = e^{-\mu/\sigma}, \gamma = 1/\sigma,$
Log-logistic	(6.2)	$\theta = -\mu/\sigma, \kappa = 1/\sigma$

For the proportional hazards representation of the Weibull model, where

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

the correspondence between  $\beta_j$ , and  $\alpha_j$  in the accelerated time model, is such that  $\beta_j = -\alpha_j/\sigma$ ,  $j = 1, 2, \dots, p$ . This shows how the  $\beta$ -parameters in a

Weibull proportional hazards model, which represent log-hazard ratios, can be obtained from the fitted  $\alpha$ -parameters in an accelerated failure time model.

## 6.6 Fitting and comparing accelerated failure time models

Accelerated failure time models are fitted using the method of maximum likelihood. The likelihood function is best derived from the log-linear representation of the model, after which iterative methods are used to obtain the estimates. The likelihood of the  $n$  observed survival times,  $t_1, t_2, \dots, t_n$ , is, from expression (5.12) in Chapter 5, given by

$$L(\boldsymbol{\alpha}, \mu, \sigma) = \prod_{i=1}^n \{f_i(t_i)\}^{\delta_i} \{S_i(t_i)\}^{1-\delta_i},$$

where  $f_i(t_i)$  and  $S_i(t_i)$  are the density and survivor functions for the  $i$ th individual at  $t_i$ , and  $\delta_i$  is the event indicator for the  $i$ th observation, so that  $\delta_i$  is unity if the  $i$ th observation is an event and zero if it is censored. Now, from equation (6.13),

$$S_i(t_i) = S_{\epsilon_i}(z_i),$$

where  $z_i = (\log t_i - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \dots - \alpha_p x_{pi})/\sigma$ , and differentiation with respect to  $t$  gives

$$f_i(t_i) = \frac{1}{\sigma t_i} f_{\epsilon_i}(z_i).$$

The likelihood function can then be expressed in terms of the survivor and density functions of  $\epsilon_i$ , giving

$$L(\boldsymbol{\alpha}, \mu, \sigma) = \prod_{i=1}^n (\sigma t_i)^{-\delta_i} \{f_{\epsilon_i}(z_i)\}^{\delta_i} \{S_{\epsilon_i}(z_i)\}^{1-\delta_i}.$$

The log-likelihood function is then

$$\log L(\boldsymbol{\alpha}, \mu, \sigma) = \sum_{i=1}^n \{-\delta_i \log(\sigma t_i) + \delta_i \log f_{\epsilon_i}(z_i) + (1 - \delta_i) \log S_{\epsilon_i}(z_i)\}, \quad (6.25)$$

and the maximum likelihood estimates of the  $p + 2$  unknown parameters,  $\mu$ ,  $\sigma$  and  $\alpha_1, \alpha_2, \dots, \alpha_p$ , are found by maximising this function using the Newton-Raphson procedure, described in Section 3.3.3.

Note that the expression for the log-likelihood function in equation (6.25) includes the term  $-\sum_{i=1}^n \delta_i \log t_i$ , which does not involve any unknown parameters. This term may therefore be omitted from the log-likelihood function, as noted in Section 5.5.1 of Chapter 5, in the context of the Weibull proportional hazards model. Indeed, log-likelihood values given by most computer software for accelerated failure time modelling do not include the value of  $-\sum_{i=1}^n \delta_i \log t_i$ .

After fitting a model, the value of the statistic  $-2 \log \hat{L}$  can be computed, and used in making comparisons between nested models, just as for the proportional hazards model. Specifically, to compare two nested models, the dif-

ference in the values of the statistic  $-2 \log \hat{L}$  for the two models is calculated, and compared with percentage points of the chi-squared distribution, with degrees of freedom equal to the difference in the number of  $\alpha$ -parameters included in the linear component of the model.

Once a suitable model has been identified, estimates of the survivor and hazard functions may be obtained and plotted. The fitted model can be interpreted in terms of the estimated value of the acceleration factor for particular individuals, or in terms of the median and other percentiles of the distribution of survival times. In particular, the estimated  $p$ th percentile of the distribution of survival times, for an individual whose vector of values of the explanatory variables is  $\mathbf{x}_i$ , is, from equation (6.14), given by

$$\hat{t}_i(p) = \exp\{\hat{\sigma}\epsilon_i(p) + \hat{\mu} + \hat{\alpha}_1 x_{1i} + \hat{\alpha}_2 x_{2i} + \dots + \hat{\alpha}_p x_{pi}\}.$$

The standard error of the estimated  $p$ th percentile,  $\hat{t}_i(p)$ , is generally obtained from the standard error of  $\log \hat{t}_i(p)$ , which is

$$\log \hat{t}_i(p) = \hat{\sigma}\epsilon_i(p) + \hat{\mu} + \hat{\alpha}_1 x_{1i} + \hat{\alpha}_2 x_{2i} + \dots + \hat{\alpha}_p x_{pi}.$$

Using the results in Appendix C,  $\text{se}\{\log \hat{t}_i(p)\} = \mathbf{d}' \mathbf{V} \mathbf{d}$ , where  $\mathbf{d}$  is a vector with components  $1, x_{1i}, \dots, x_{pi}, \epsilon_i(p)$ , and  $\mathbf{V}$  is the variance covariance matrix of the parameter estimates in the order  $\hat{\mu}, \hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_p, \hat{\sigma}$ . Then, the standard error of the estimated percentile itself is obtained from  $\text{se}\{\hat{t}_i(p)\} = \hat{t}_i(p) \text{se}\{\log \hat{t}_i(p)\}$ .

### Example 6.3 Prognosis for women with breast cancer

In this example, accelerated failure time models are fitted to the data on the survival times of women with breast cancer. The Weibull accelerated failure time model is first considered. A log-linear model for the random variable associated with the survival time of the  $i$ th woman,  $T_i$ , is such that

$$\log T_i = \mu + \alpha x_i + \sigma \epsilon_i,$$

where  $\epsilon_i$  has a Gumbel distribution,  $\mu$ ,  $\sigma$  and  $\alpha$  are unknown parameters, and  $x_i$  is the value of an explanatory variable,  $X$ , associated with staining, such that  $x_i = 0$  if the  $i$ th woman is negatively stained and  $x_i = 1$  if positively stained. When this model is fitted, we find that  $\hat{\mu} = 5.854$ ,  $\hat{\sigma} = 1.067$ , and  $\hat{\alpha} = -0.997$ .

The acceleration factor,  $e^{-\alpha x_i}$ , is estimated by  $e^{0.997} = 2.71$  for a woman with positive staining. The time to death of a woman with a positively stained tumour is therefore accelerated by a factor of about 2.7 under this model. This is in broad agreement with the estimated slope of the percentile-percentile plot for this data set, found in Example 6.2.

The estimated survivor function for the  $i$ th woman is given by

$$\hat{S}_i(t) = \exp\left\{-\exp\left(\frac{\log t - \hat{\mu} - \hat{\alpha}x_i}{\hat{\sigma}}\right)\right\},$$

from equation (6.17), and may be plotted against  $t$  for the two possible values of  $x_i$ . The median survival time of the  $i$ th woman under the Weibull acceler-

ated failure time model, using equation (6.20), is

$$t_i(50) = \exp \{ \sigma \log (\log 2) + \mu + \alpha x_i \}.$$

The estimated median survival time for a woman with negative staining ( $x_i = 0$ ) is 236 days, while that for women with positive staining ( $x_i = 1$ ) is 87 days, as in Example 5.6. Note that the ratio of the two medians is 2.71, which is the acceleration factor. The median survival time for women with positively stained tumours is therefore about one third that of those whose tumours were negatively stained.

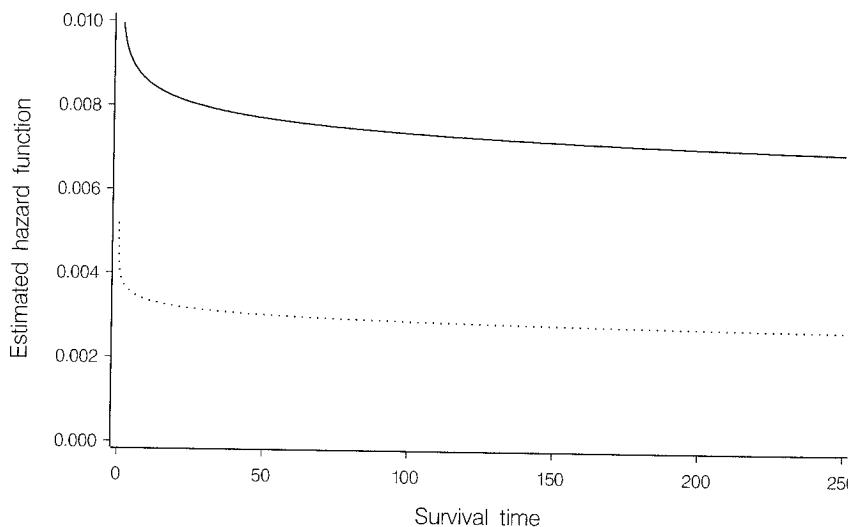
The estimated hazard function for the  $i$ th woman is, from equation (6.18), given by

$$\hat{h}_i(t) = \hat{\sigma}^{-1} t^{\hat{\sigma}-1-1} \exp \left( \frac{-\hat{\mu} - \hat{\alpha} x_i}{\hat{\sigma}} \right),$$

that is,

$$\hat{h}_i(t) = 0.937 t^{-0.063} \exp(-5.486 + 0.934 x_i).$$

A plot of this function for the two groups of women is shown in Figure 6.6.



**Figure 6.6** Estimated hazard functions under the Weibull accelerated failure time model for women with positively stained (—) and negatively stained (···) tumours.

In the proportional hazards representation of the Weibull model, given in Section 5.5 of Chapter 5, the hazard of death at time  $t$  for the  $i$ th woman is

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

where  $x_i$  takes the value zero if the  $i$ th woman had a negatively stained tumour, and unity if the tumour was positively stained. For the Weibull distribution, the baseline hazard function is

$$h_0(t) = \lambda \gamma t^{\gamma-1},$$

which is the hazard function for women with negatively stained tumours, and hence

$$h_i(t) = e^{\beta x_i} \lambda \gamma t^{\gamma-1}.$$

The corresponding estimated values of the parameters  $\lambda$ ,  $\gamma$  and  $\beta$  are given by  $\hat{\lambda} = \exp(-\hat{\mu}/\hat{\lambda}) = 0.00414$ ,  $\hat{\gamma} = 1/\hat{\sigma} = 0.937$  and  $\hat{\beta} = -\hat{\alpha}/\hat{\sigma} = 0.997$ . The correspondence between the Weibull accelerated failure time model and the Weibull proportional hazards model means that the hazard ratio under the latter model is  $e^{-\alpha/\sigma} = e^\beta$ , which is estimated to be 2.55. This is in agreement with the value found in Example 5.6.

We now fit the log-logistic accelerated failure time model to the same data set. The log-linear form of the model now leads to  $\hat{\mu} = 5.461$ ,  $\hat{\sigma} = 0.805$  and  $\hat{\alpha} = -1.149$ . The acceleration factor is  $e^{-\alpha}$ , which is estimated by 3.16. This is slightly greater than that found under the Weibull accelerated failure time model.

The median survival time for the  $i$ th woman under this model is, from equation (6.23), given by

$$\exp(\mu + \alpha x_i),$$

from which the estimated median survival time for a women with negative staining is 235 days, while that for women with positive staining is 75 days. These values are very close to those obtained under the Weibull accelerated failure time model.

The estimated hazard function for the  $i$ th woman is now

$$\hat{h}_i(t) = \frac{1}{\hat{\sigma} t} \left\{ 1 + \exp \left[ - \left( \frac{\log t - \hat{\mu} - \hat{\alpha} x_i}{\hat{\sigma}} \right) \right] \right\}^{-1},$$

from equation (6.22), which is

$$\hat{h}_i(t) = 1.243 t^{-1} \left\{ 1 + t^{-1.243} \exp(6.787 - 1.428 x_i) \right\}^{-1}.$$

A graph of this function for the two groups of women is shown in Figure 6.7. This can be compared with the graph in Figure 6.6.

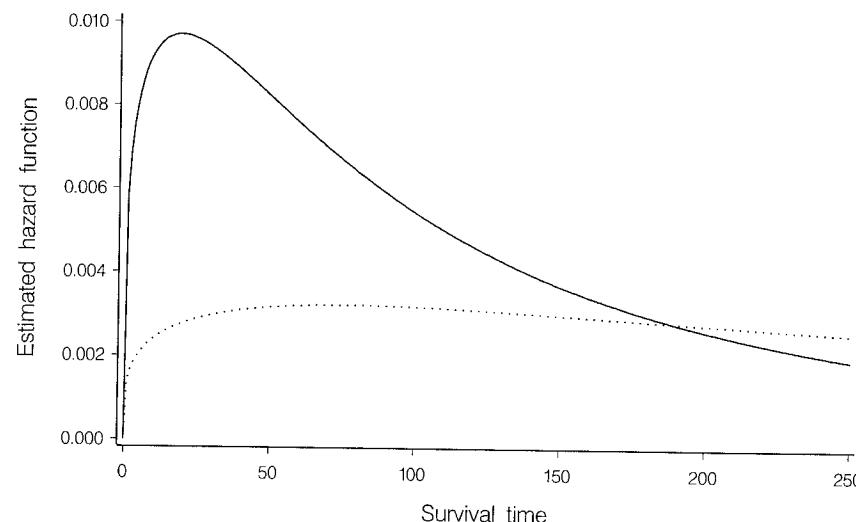
The hazard functions for those with negative staining are quite similar under the two models. However, the hazard function for those with positive staining under the log-logistic model is different from that under the Weibull model. The values of the statistic  $-2 \log \hat{L}$  for the fitted Weibull and log-logistic models are 121.77 and 118.495. On this basis, the log-logistic model is a slightly better fit. An analysis of residuals, to be discussed in Chapter 7, may help in choosing between these two models, although with this small data set, such an analysis is unlikely to be very informative.

Finally, in terms of the parameterisation of the model given in Section 6.1.1, the baseline hazard function is

$$h_0(t) = \frac{e^\theta \kappa t^{\kappa-1}}{1 + e^\theta t^\kappa},$$

and so the hazard function for the  $i$ th woman in the study is

$$h_i(t) = \frac{e^{\theta - \kappa \alpha x_i} \kappa t^{\kappa-1}}{1 + e^{\theta - \kappa \alpha x_i} t^\kappa}.$$



**Figure 6.7** Estimated hazard functions under the log-logistic accelerated failure time model for women with positively stained (—) and negatively stained (···) tumours.

The corresponding estimated values of  $\theta$  and  $\kappa$  are given by  $\hat{\theta} = -\hat{\mu}/\hat{\sigma} = -6.787$ , and  $\hat{\kappa} = 1/\hat{\sigma} = 1.243$ .

#### Example 6.4 Comparison of two treatments for prostatic cancer

In a further illustration of modelling survival data using the log-logistic accelerated failure time model, the data from a clinical trial to compare two treatments for prostatic cancer are considered. These data were first given in Example 1.4, and analysed using a Cox regression model in Examples 3.6 and 3.10.

To identify the terms that should be in the linear component of the log-logistic accelerated failure time model, the procedure described in Example 3.6 can again be followed. The values of the statistic  $-2 \log \hat{L}$  on fitting models with all combinations of the four prognostic variables, *Age*, *Shb*, *Size* and *Index*, are shown in Table 6.4. As in Example 3.6, the variables *Size* and *Index* are the ones that are needed in the model. When either of these variables is omitted, the corresponding increase in the value of  $-2 \log \hat{L}$  is significant, and neither *Age* nor *Shb* reduce  $-2 \log \hat{L}$  by a significant amount when they are added to the model.

When the term corresponding to the treatment effect, *Treat*, is added to the model that contains *Size* and *Index*,  $-2 \log \hat{L}$  decreases to 21.245. When this reduction of 1.867 is compared with percentage points of a chi-squared distribution on 1 d.f., the reduction is not significant at the 10% level ( $P = 0.172$ ). There is no evidence of any interaction between *Treat* and the prognostic variables *Size* and *Index*, and so the conclusion is that there is no statistically significant treatment effect.

**Table 6.4** Values of  $-2 \log \hat{L}$  for models fitted to the data from Example 1.4.

Variables in model	$-2 \log \hat{L}$
none	35.806
<i>Age</i>	35.752
<i>Shb</i>	35.700
<i>Size</i>	27.754
<i>Index</i>	27.965
<i>Age</i> + <i>Shb</i>	35.657
<i>Age</i> + <i>Size</i>	27.652
<i>Age</i> + <i>Index</i>	27.859
<i>Shb</i> + <i>Size</i>	27.722
<i>Shb</i> + <i>Index</i>	26.873
<i>Size</i> + <i>Index</i>	23.112
<i>Age</i> + <i>Shb</i> + <i>Size</i>	27.631
<i>Age</i> + <i>Shb</i> + <i>Index</i>	26.870
<i>Age</i> + <i>Size</i> + <i>Index</i>	23.002
<i>Shb</i> + <i>Size</i> + <i>Index</i>	22.895
<i>Age</i> + <i>Shb</i> + <i>Size</i> + <i>Index</i>	22.727

The magnitude of the treatment effect can be assessed by calculating the acceleration factor. According to the log-linear form of the model, the random variable associated with the survival time of the  $i$ th patient,  $T_i$ , is such that

$$\log T_i = \mu + \alpha_1 \text{Size}_i + \alpha_2 \text{Index}_i + \alpha_3 \text{Treat}_i + \sigma \epsilon_i,$$

in which  $\epsilon_i$  has a logistic distribution,  $\text{Size}_i$  and  $\text{Index}_i$  are the values of tumour size and Gleason index for the  $i$ th individual, and  $\text{Treat}_i$  is zero if the  $i$ th individual is in the placebo group and unity if in the treated group. The maximum likelihood estimates of the unknown parameters in this model are given by  $\hat{\mu} = 7.661$ ,  $\hat{\sigma} = 0.338$ ,  $\hat{\alpha}_1 = -0.029$ ,  $\hat{\alpha}_2 = -0.293$ , and  $\hat{\alpha}_3 = 0.573$ . The values of the  $\alpha$ 's suggests that the survival time tends to be shorter for larger values of the tumour size and tumour index, and longer for individuals assigned to the active treatment.

Using equation (6.21), the fitted survivor function for the  $i$ th patient is

$$\hat{S}_i(t) = \left[ 1 + \exp \left\{ \frac{\log t - \hat{\mu} - \hat{\alpha}_1 \text{Size}_i - \hat{\alpha}_2 \text{Index}_i - \hat{\alpha}_3 \text{Treat}_i}{\hat{\sigma}} \right\} \right]^{-1},$$

which can be written in the form

$$\hat{S}_i(t) = \left\{ 1 + t^{1/\hat{\sigma}} \exp \left( \hat{\zeta}_i \right) \right\}^{-1},$$

where

$$\hat{\zeta}_i = \frac{1}{\hat{\sigma}} \{ -\hat{\mu} - \hat{\alpha}_1 \text{Size}_i - \hat{\alpha}_2 \text{Index}_i - \hat{\alpha}_3 \text{Treat}_i \},$$

that is,

$$\hat{\zeta}_i = -22.645 + 0.085 \text{Size}_i + 0.865 \text{Index}_i - 1.693 \text{Treat}_i.$$

The corresponding estimated hazard function can be found by differentiating the estimated cumulative hazard function,  $\hat{H}_i(t) = -\log \hat{S}_i(t)$ , with respect to  $t$ . This gives

$$\hat{h}_i(t) = \frac{1}{\hat{\sigma}t} \left\{ 1 + t^{-1/\hat{\sigma}} \exp(-\hat{\zeta}_i) \right\}^{-1},$$

a result that can also be obtained directly from the hazard function given in equation (6.22).

The estimated acceleration factor for an individual in the treated group, relative to an individual in the control group, is  $e^{-0.573} = 0.56$ . The interpretation of this result is that after allowing for the size and index of the tumour, the effect of the treatment with DES is to slow down the progression of the cancer by a factor of about 2. This effect might be of clinical importance, even though it is not statistically significant. However, before accepting this interpretation, the adequacy of the fitted model should be checked using suitable diagnostics.

A confidence interval for the acceleration factor is found by exponentiating the confidence limits for the logarithm of the acceleration factor. In this example, the logarithm of the acceleration factor for the treatment effect is the estimated coefficient of *Treat* in the model for the hazard function, multiplied by  $-1$ , which is  $-0.573$ , and the standard error of this estimate is 0.473. Thus a 95% confidence interval for the acceleration factor has limits of  $\exp\{-0.573 \pm 1.96 \times 0.473\}$ , and the required interval is from 0.70 to 4.48. Notice that this interval estimate includes unity, which is consistent with the earlier finding of a non-significant treatment difference.

Finally, in terms of the parameterisation of the model in Section 6.1.1, the fitted hazard function for the  $i$ th patient,  $i = 1, 2, \dots, 38$ , is given by

$$\hat{h}_i(t) = e^{-\hat{\eta}_i} \hat{h}_0(e^{-\hat{\eta}_i} t),$$

where

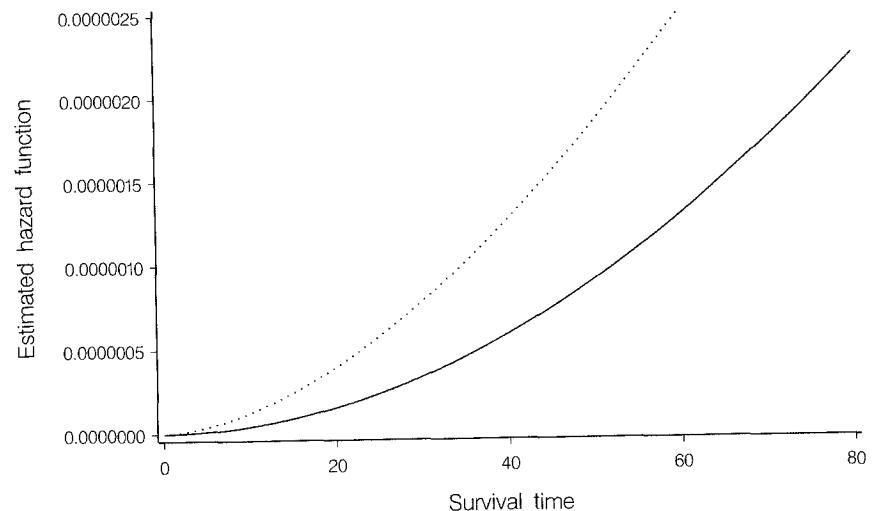
$$\hat{\eta}_i = -0.029 \text{Size}_i - 0.293 \text{Index}_i + 0.573 \text{Treat}_i,$$

and from equation (6.1),

$$\hat{h}_0(t) = \frac{e^{\hat{\theta}} \hat{\kappa} t^{\hat{\kappa}-1}}{1 + e^{\hat{\theta}} t^{\hat{\kappa}}}.$$

The estimated parameters in this form of the estimated baseline hazard function,  $\hat{h}_0(t)$ , are given by  $\hat{\theta} = -22.644$  and  $\hat{\kappa} = 2.956$ . A graph of this function is shown in Figure 6.8.

This figure indicates that the baseline hazard is increasing over time. Comparison with the baseline hazard function for a fitted Weibull model, also shown in this figure, indicates that under the log-logistic model, the estimated baseline hazard function does not increase quite so rapidly.



**Figure 6.8** Estimated baseline hazard function for the fitted log-logistic model (—) and a fitted Weibull model (···).

### 6.7\* The proportional odds model

In this general model, the odds of an individual surviving beyond some time  $t$  are expressed as

$$\frac{S_i(t)}{1 - S_i(t)} = e^{\eta_i} \frac{S_0(t)}{1 - S_0(t)}, \quad (6.26)$$

where

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

is a linear combination of the values of  $p$  explanatory variables,  $X_1, X_2, \dots, X_p$ , measured on the  $i$ th individual, and  $S_0(t)$ , the baseline survivor function, is the survivor function for an individual whose explanatory variables all take the value zero.

In this model, the explanatory variables act multiplicatively on the odds of survival beyond  $t$ . The logarithm of the ratio of the odds of survival beyond  $t$  for the  $i$ th individual, relative to an individual for whom the explanatory variables are all equal to zero, is therefore just  $\eta_i$ . The model is therefore a linear model for the log-odds ratio.

Now consider the particular case of a two-group study, in which individuals receive either a standard treatment or new treatment. Let the single indicator variable  $X$  take the value zero if an individual is on the standard treatment and unity if on the new. The odds of the  $i$ th individual surviving beyond time  $t$  is then

$$\frac{S_i(t)}{1 - S_i(t)} = e^{\beta x_i} \frac{S_0(t)}{1 - S_0(t)},$$

where  $x_i$  is the value of  $X$  for the  $i$ th individual,  $i = 1, 2, \dots, n$ . Thus if  $S_N(t)$

and  $S_N(t)$  are the survivor functions for individuals on the new and standard treatments, respectively,

$$\frac{S_N(t)}{1 - S_N(t)} = e^\beta \frac{S_S(t)}{1 - S_S(t)},$$

and the log-odds ratio is simply  $\beta$ . The parameters in the linear component of the model therefore have an immediate interpretation.

As for the proportional hazards model, a non-parametric estimate of the baseline hazard function can be obtained. The model is then fitted by estimating the  $\beta$ -parameters in the linear component of the model, and the baseline survivor function, from the data. A method for accomplishing this has been described by Bennett (1983a), but details will not be included here. Fully parametric versions of the proportional odds model can be derived by using a specific probability distribution for the survival times. One such model is described below in Section 6.7.1.

One particularly important property of the proportional odds model concerns the ratio of the hazard function for the  $i$ th individual to the baseline hazard,  $h_i(t)/h_0(t)$ . It can be shown that this ratio converges from the value  $e^{-\eta_i}$  at time  $t = 0$ , to unity at  $t = \infty$ . To show this, the model in equation (6.26) can be rearranged to give

$$S_i(t) = S_0(t) \{e^{-\eta_i} + (1 - e^{-\eta_i})S_0(t)\}^{-1},$$

and taking logarithms, we get

$$\log S_i(t) = \log S_0(t) - \log \{e^{-\eta_i} + (1 - e^{-\eta_i})S_0(t)\}. \quad (6.27)$$

Using the general result from equation (1.4), the hazard function is

$$h_i(t) = -\frac{d}{dt} \log S_i(t),$$

and so

$$h_i(t) = h_0(t) - \frac{(1 - e^{-\eta_i})f_0(t)}{e^{-\eta_i} + (1 - e^{-\eta_i})S_0(t)},$$

after differentiating both sides of equation (6.27) with respect to  $t$ , where  $f_0(t)$  is the baseline probability density function. After some rearrangement, this equation becomes

$$h_i(t) = h_0(t) - \frac{f_0(t)}{(e^{\eta_i} - 1)^{-1} + S_0(t)}. \quad (6.28)$$

From equation (1.3), we also have that  $h_0(t) = f_0(t)/S_0(t)$  and substituting for  $f_0(t)$  in equation (6.28) gives

$$h_i(t) = h_0(t) \left\{ 1 - \frac{S_0(t)}{(e^{\eta_i} - 1)^{-1} + S_0(t)} \right\}.$$

Finally, after further rearrangement, the hazard ratio is given by

$$\frac{h_i(t)}{h_0(t)} = \{1 + (e^{\eta_i} - 1)S_0(t)\}^{-1}.$$

## THE PROPORTIONAL ODDS MODEL

As  $t$  increases from 0 to  $\infty$ , the baseline survivor function decreases monotonically from 1 to 0. When  $S_0(t) = 1$ , the hazard ratio is  $e^{-\eta_i}$  and as  $t$  increases to  $\infty$ , the hazard ratio converges to unity.

In practical applications, it is common for the hazard functions obtained for patients in two or more groups to converge with time. For example, in a follow-up study of patients in a clinical trial, the effect on survival of the treatment, or the initial stage of disease, may wear off. Similarly, in studies where a group of patients with some disease are being compared with a control group of disease-free individuals, an effective cure of the disease would lead to the survival experience of each group becoming more similar over time. This suggests that the proportional odds model, with its property of convergent hazard functions, might be of considerable value. However, there are two reasons why this general model has not been widely used in practice. The first of these is that computer software for fitting the model is not generally available. The second is that the model is likely to give similar results to a Cox regression model that includes a time-dependent variable to produce non-proportional hazards. This particular approach to modelling survival data with non-proportional hazards was described in Section 4.4.3, and is considered more fully in Chapter 8.

### 6.7.1 The log-logistic proportional odds model

If survival times for individuals are assumed to have a log-logistic distribution, the baseline survivor function is

$$S_0(t) = \{1 + e^{\theta t^\kappa}\}^{-1},$$

where  $\theta$  and  $\kappa$  are unknown parameters. The baseline odds of survival beyond time  $t$  are then given by

$$\frac{S_0(t)}{1 - S_0(t)} = e^{-\theta t^{-\kappa}}.$$

The odds of the  $i$ th individual surviving beyond time  $t$  are therefore

$$\frac{S_i(t)}{1 - S_i(t)} = e^{\eta_i - \theta t^{-\kappa}},$$

and so the survival time of the  $i$ th individual has a log-logistic distribution with parameters  $\theta - \eta_i$  and  $\kappa$ . The log-logistic distribution therefore has the *proportional odds property*, and the distribution is the natural one to use in conjunction with the proportional odds model. In fact, it is the only distribution to share both the accelerated failure time property and the proportional odds property.

This result means that the  $\beta$ -parameters under the proportional odds model can be obtained from those under the accelerated failure time model, and vice versa. In particular, the estimated coefficients of the explanatory variable in the linear component of the proportional odds model are obtained by multiplying those in the accelerated failure time model by  $\hat{\sigma}^{-1}$ , where  $\hat{\sigma}$  is the

estimated value of the parameter  $\sigma$  in the accelerated failure time model. This enables the result of the survival analysis to be interpreted in terms of an acceleration factor or the ratio of the odds of survival beyond some time, whichever is the more convenient.

As for other models for survival data, the proportional odds model can be fitted using the method of maximum likelihood. Alternative models may then be compared on the basis of the statistic  $-2 \log \hat{L}$ .

In a two-group study, a preliminary examination of the likely suitability of the model can easily be undertaken. The log-odds of the  $i$ th individual surviving beyond time  $t$  are

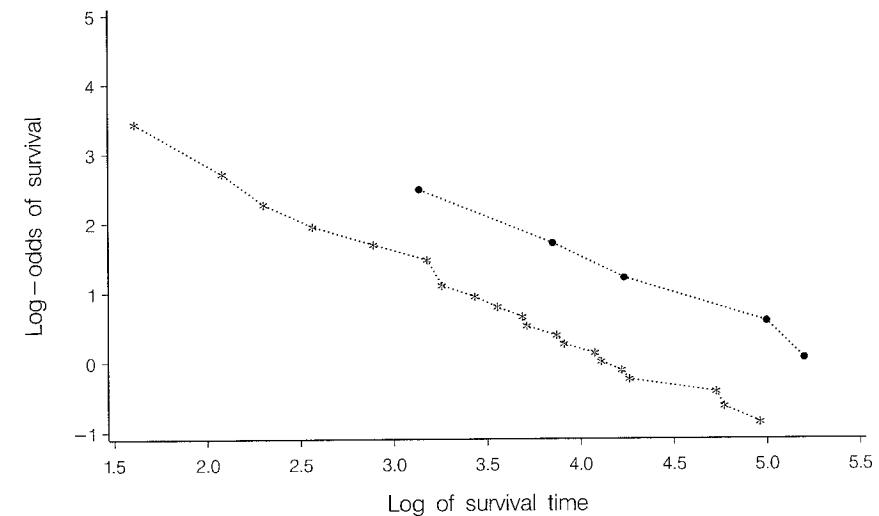
$$\log \left\{ \frac{S_i(t)}{1 - S_i(t)} \right\} = \beta x_i - \theta - \kappa \log t,$$

where  $x_i$  is the value of an indicator variable that takes the value zero if an individual is in one group and unity if in the other. The Kaplan-Meier estimate of the survivor function is then obtained for the individuals in each group and the estimated log-odds of survival beyond time  $t$ ,  $\log \left\{ \hat{S}_i(t)/[1 - \hat{S}_i(t)] \right\}$ , are plotted against  $\log t$ . If the plot shows two parallel straight lines, this would indicate that the log-logistic model was appropriate. If the lines were straight but not parallel, this would suggest that the parameter  $\kappa$  in the model was not the same for each treatment group. Parallel curves in this plot suggest that although the proportional odds assumption is valid, the survival times cannot be taken to have a log-logistic distribution.

#### *Example 6.5 Prognosis for women with breast cancer*

In this example to illustrate the use of the proportional odds model, the model is fitted to the data on the survival times of breast cancer patients. In order to assess the likely suitability of the proportional odds model, the Kaplan-Meier estimate of the survivor function for the negatively and positively stained women is computed. For the two groups of women, the log-odds of survival beyond time  $t$  are estimated and plotted against  $\log t$ . The resulting graph is shown in Figure 6.9. The lines are reasonably straight and parallel, and so we go on to use the log-logistic proportional odds model to summarise these data.

The model can be fitted using software for fitting the log-logistic accelerated failure time model. In Example 6.3, this latter model was fitted to the data on the survival of breast cancer patients. The estimated value of  $\kappa$  and  $\theta$  in the proportional odds model are 1.243 and  $-6.787$ , the same as those in the accelerated failure time model. However, the estimated value of  $\beta$  in the linear component of the proportional odds model is  $\hat{\beta} = -1.149 \times 1.243 = -1.428$ . This is an estimate of the logarithm of the ratio of the odds of a positively stained woman surviving beyond time  $t$ , relatively to one who is negatively stained. The corresponding odds ratio is  $e^{-1.428} = 0.24$ , so that the odds of a woman surviving beyond  $t$  are about four times greater if that woman has a negatively stained tumour.



**Figure 6.9** Estimated values of the log-odds of survival beyond  $t$  plotted against  $\log t$  for women with positively stained (\*) and negatively stained (●) tumours.

#### 6.8\* Some other distributions for survival data

Although a number of distributions for survival data have already been considered in some detail, there are others that can be useful in specific circumstances. Some of these are mentioned in this section.

When the hazard of death is expected to increase or decrease with time in the short term, and to then become constant, a hazard function that follows a *general exponential curve* or *Mitscherlich curve* may be appropriate. We would then take the hazard function to be

$$h(t) = \theta - \beta e^{-\gamma t},$$

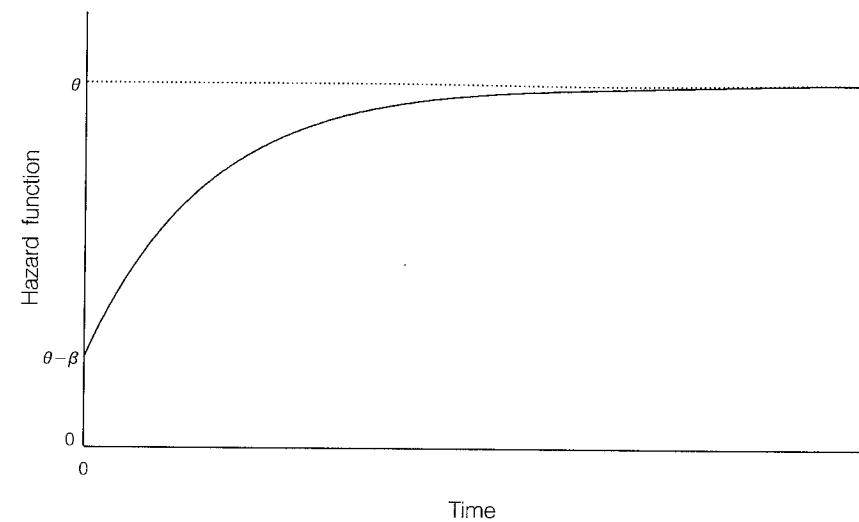
where  $\theta > 0$ ,  $\beta > 0$  and  $\gamma > 0$ . This is essentially a Gompertz hazard function, defined in Section 5.7, with an additional constant. The general shape of this function is depicted in Figure 6.10. This function has a value of  $\theta - \beta$  when  $t = 0$  and increases to a horizontal asymptote at a hazard of  $\theta$ . Similarly the function

$$h(t) = \theta + \beta e^{-\gamma t},$$

where  $\theta > 0$ ,  $\beta > 0$  and  $\gamma > 0$ , could be used to model a hazard which decreases from  $\theta + \beta$  to a horizontal asymptote at  $\theta$ .

Using equation (1.5), the corresponding survivor function can be found, from which the probability density function can be obtained. The probability distribution corresponding to this specification of the hazard function is known as the *Gompertz-Makeham distribution*.

To model a hazard function that decreases and then increases symmetrically about the minimum value, a quadratic hazard function might be suitable.



**Figure 6.10** An asymptotic hazard function, where  $h(t) = \theta - \beta e^{-\gamma t}$ .

Thus if

$$h(t) = \theta + \beta t + \gamma t^2,$$

for values of  $\theta, \beta$  and  $\gamma$  which give the required shape of hazard and ensure that  $h(t) \geq 0$ , explicit forms for the survivor function and probability density function can be obtained.

Another form of hazard function that decreases to a single minimum and increases thereafter is the ‘bathtub’ hazard. The model with

$$h(t) = \alpha t + \frac{\beta}{1 + \gamma t}$$

provides a straightforward representation of this form of hazard, and corresponding expressions for the survivor and density functions can be found.

Each of the models described in this section can be fitted by constructing a log-likelihood function, using the result in expression (5.38) of Chapter 5, and maximising this with respect to the unknown model parameters. In principle, the unknown parameters in the hazard function can also depend on the values of explanatory variables. Non-linear optimisation routines can then be used to maximise the log-likelihood.

## 6.9 Further reading

The properties of random variables that have probability distributions such as the logistic, lognormal and gamma, are presented in Johnson and Kotz (1970). Chhikara and Folks (1989) give a detailed study of the inverse Gaussian distribution.

A description of the log-linear model for survival data is contained in many

of the major textbooks on survival analysis; see in particular Cox and Oakes (1984), Kalbfleisch and Prentice (2002), Klein and Moeschberger (1997) or Lawless (2002). Descriptions of this form of the model are also presented in computer manuals that accompany computer software packages for accelerated failure time modelling.

Cox and Oakes (1984) show that the Weibull distribution is the only one to have both the proportional hazards property and the accelerated failure time property. They also demonstrate that the log-logistic distribution is the only one that shares the accelerated failure time property and the proportional odds property.

A non-parametric version of the accelerated failure time model, which does not require the specification of a probability distribution for the survival data, has been introduced by Wei (1992). This paper, and the published discussion, Fisher (1992), includes comments on whether the accelerated failure time model should be used more widely in the analysis of survival data.

The application of the accelerated failure time and proportional odds models to the analysis of reliability data is described by Crowder *et al.* (1991). The general proportional odds model for survival data was introduced by Bennett (1983a). Bennett (1983c) describes the log-logistic proportional odds model, and GLIM macros for fitting the model are described in Bennett and Whitehead (1981). The model has been further developed by Yang and Prentice (1999).

The piecewise exponential model, mentioned in Section 6.3.1, in which hazards are constant over particular time intervals, was introduced by Breslow (1974). Breslow also points out that the Cox regression model is equivalent to a piecewise exponential model with constant hazards between each death time. The piecewise exponential model and the use of the normal, lognormal, logistic and log-logistic distributions for modelling survival times are described in Aitkin *et al.* (1989).

Use of the quadratic hazard function was discussed by Gaver and Acar (1979) and the bathtub hazard function was proposed by Hjorth (1980).

A more general way of modelling survival data is to use a general family of distributions for survival times, which includes the Weibull and log-logistic as special cases. The choice between alternative distributions can then be made within a likelihood framework. In particular, the exponential, Weibull, log-logistic, lognormal and gamma distributions are special cases of the generalised  $F$ -distribution described by Kalbfleisch and Prentice (2002). However, this methodology will only tend to be informative in the analysis of data sets in which the number of death times is relatively large.

In another approach to modelling survival data, the baseline hazard function is represented by regression splines. A flexible class of models, based on proportional hazards and proportional odds assumptions, has been proposed by Royston and Parmar (2002). The implementation of this approach, through the statistical package Stata, has been described by Royston (2001).

## Model checking in parametric models

Diagnostic procedures for the assessment of model adequacy are as important in parametric modelling as they are when the Cox regression model is used in the analysis of survival data. Procedures based on residuals are particularly relevant, and so we begin this chapter by defining residuals for parametric models, some of which stem from those developed for the Cox model, described in Chapter 4. This is followed by a summary of graphical procedures for assessing the suitability of models fitted to data that are assumed to have a Weibull, log-logistic or lognormal distribution. Other ways of examining the fit of a parametric regression model are then considered, along with methods for the detection of influential observations. We conclude with a summary of how the assumption of proportional hazards can be examined after fitting the Weibull proportional hazards model.

### 7.1 Residuals for parametric models

Suppose that  $T_i$  is the random variable associated with the survival time of the  $i$ th individual,  $i = 1, 2, \dots, n$ , and that  $x_{1i}, x_{2i}, \dots, x_{pi}$  are the values of  $p$  explanatory variables,  $X_1, X_2, \dots, X_p$ , for this individual. Assuming an accelerated failure time model for  $T_i$ , we have that

$$\log T_i = \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \dots + \alpha_p x_{pi} + \sigma \epsilon_i,$$

where  $\epsilon_i$  is a random variable with a probability distribution that depends on the distribution adopted for  $T_i$ , and  $\mu$ ,  $\sigma$  and  $\alpha_j$ ,  $j = 1, 2, \dots, p$ , are unknown parameters. If the observed survival time of the  $i$ th individual is censored, the corresponding residual will also be censored, complicating the interpretation of these quantities.

#### 7.1.1 Standardised residuals

A natural form of residual to adopt in accelerated failure time modelling is the *standardised residual* defined by

$$r_{Si} = \{\log t_i - \hat{\mu} - \hat{\alpha}_1 x_{1i} - \hat{\alpha}_2 x_{2i} - \dots - \hat{\alpha}_p x_{pi}\} / \hat{\sigma}, \quad (7.1)$$

where  $t_i$  is the observed survival time of the  $i$ th individual, and  $\hat{\mu}$ ,  $\hat{\sigma}$ ,  $\hat{\alpha}_j$ ,  $j = 1, 2, \dots, p$ , are the estimated parameters in the fitted accelerated failure time model. This residual has the appearance of a quantity of the form “observation – fitted value”, and would be expected to have the same distribution as that of  $\epsilon_i$  in the accelerated failure time model, if the model were

correct. For example, if a Weibull distribution is adopted for  $T_i$ , the  $r_{Si}$  would be expected to behave as if they were a possibly censored sample from a Gumbel distribution, if the fitted model is correct. The estimated survivor function of the residuals would then be similar to the survivor function of  $\epsilon_i$ , that is,  $S_{\epsilon_i}(\epsilon)$ . Using the general result in Section 4.1.1 of Chapter 4,  $-\log S_{\epsilon_i}(\epsilon)$  has a unit exponential distribution, and so it follows that  $-\log S_{\epsilon_i}(r_{Si})$  will have an approximate unit exponential distribution, if the fitted model is appropriate. This provides the basis for a diagnostic plot that may be used in the assessment of model adequacy, described in Section 7.2.4.

### 7.1.2 Cox-Snell residuals

The Cox-Snell residuals that were defined for the Cox regression model in Section 4.1.1 of Chapter 4 are essentially the estimated values of the cumulative hazard function for the  $i$ th observation, at the corresponding event time,  $t_i$ . Residuals that have a similar form may also be used in assessing the adequacy of parametric models. The main difference is that now the survivor and hazard functions are parametric functions that depend on the distribution adopted for the survival times. In particular, the estimated survivor function for the  $i$ th individual, on fitting an accelerated failure time model, from equation (6.13), is given by

$$\hat{S}_i(t) = S_{\epsilon_i} \left( \frac{\log t - \hat{\mu} - \hat{\alpha}_1 x_{1i} - \hat{\alpha}_2 x_{2i} - \cdots - \hat{\alpha}_p x_{pi}}{\hat{\sigma}} \right), \quad (7.2)$$

where  $S_{\epsilon_i}(\epsilon)$  is the survivor function of  $\epsilon_i$  in the accelerated failure time model,  $\hat{\alpha}_j$  is the estimated coefficient of  $x_{ji}$ ,  $j = 1, 2, \dots, p$ , and  $\hat{\mu}, \hat{\sigma}$  are the estimated values of  $\mu$  and  $\sigma$ . The form of  $S_{\epsilon_i}(\epsilon)$  for some commonly used distributions for  $T_i$  was summarised in Table 6.2 of Chapter 6.

The Cox-Snell residuals for a parametric model are defined by

$$r_{Ci} = \hat{H}_i(t_i) = -\log \hat{S}_i(t_i), \quad (7.3)$$

where  $\hat{H}_i(t_i)$  is the estimated cumulative hazard function, and  $\hat{S}_i(t_i)$  is the estimated survivor function in equation (7.2), evaluated at  $t_i$ . As in the context of the Cox regression model, these residuals can be taken to have a unit exponential distribution when the correct model has been fitted, with censored observations leading to censored residuals; see Section 4.1.1 for details.

The Cox-Snell residuals in equation (7.3) are very closely related to the standardised residuals in equation (7.1), since from equation (7.2), we see that  $r_{Ci} = -\log S_{\epsilon_i}(r_{Si})$ . Assessment of whether the standardised residuals have a particular distribution is therefore equivalent to assessing whether the corresponding Cox-Snell residuals have a unit exponential distribution.

### 7.1.3 Martingale residuals

The martingale residuals provide a measure of the difference between the observed number of deaths in the interval  $(0, t_i]$ , which is either 0 or 1, and the

number predicted by the model. Observations with unusually large martingale residuals are not well fitted by the model. The analogue of the martingale residual, defined for the Cox regression model in equation (4.6) of Chapter 4, is such that

$$r_{Mi} = \delta_i - r_{Ci}, \quad (7.4)$$

where  $\delta_i$  is the event indicator for the  $i$ th observation, so that  $\delta_i$  is unity if that observation is an event and zero if censored, and now  $r_{Ci}$  is the Cox-Snell residual given in equation (7.3). For reasons given in Section 7.1.5, the martingale residuals for a parametric accelerated failure time model sum to zero, but are not symmetrically distributed about zero. Strictly speaking, it is no longer appropriate to refer to these residuals as martingale residuals since the derivation of them, based on martingale methods, does not carry over to the accelerated failure time model. However, for semantic convenience, we will continue to refer to the quantities in equation (7.4) as martingale residuals.

### 7.1.4 Deviance residuals

The deviance residuals, which were first presented in equation (4.7) of Chapter 4, can be regarded as an attempt to make the martingale residuals symmetrically distributed about zero, and are defined by

$$r_{Di} = \text{sgn}(r_{Mi}) [-2 \{r_{Mi} + \delta_i \log(\delta_i - r_{Mi})\}]^{\frac{1}{2}}. \quad (7.5)$$

It is important to note that these quantities are not components of the deviance for the fitted parametric model, but nonetheless it will be convenient to continue to refer to them as deviance residuals.

### 7.1.5\* Score residuals

Score residuals, which parallel the score residuals, or Schoenfeld residuals, used in connection with the Cox regression model, can be defined for any parametric model. The score residuals are the components of the derivatives of the log-likelihood function, with respect to the unknown parameters,  $\mu, \sigma$  and  $\alpha_j$ ,  $j = 1, 2, \dots, p$ , and evaluated at the maximum likelihood estimates of these parameters,  $\hat{\mu}, \hat{\sigma}$  and  $\hat{\alpha}_j$ . From equation (6.25) of Chapter 6, the log-likelihood function for  $n$  observations is

$$\log L(\boldsymbol{\alpha}, \mu, \sigma) = \sum_{i=1}^n \{-\delta_i \log(\sigma t_i) + \delta_i \log f_{\epsilon_i}(z_i) + (1 - \delta_i) \log S_{\epsilon_i}(z_i)\},$$

where  $z_i = (\log t_i - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi})/\sigma$ ,  $f_{\epsilon_i}(\epsilon)$  and  $S_{\epsilon_i}(\epsilon)$  are the density and survivor functions of  $\epsilon_i$ , and  $\delta_i$  is the event indicator.

Differentiating this log-likelihood function with respect to the parameters  $\mu, \sigma$ , and  $\alpha_j$ , for  $j = 1, 2, \dots, p$  gives the following derivatives:

$$\frac{\partial \log L}{\partial \mu} = \sigma^{-1} \sum_{i=1}^n g(z_i),$$

$$\begin{aligned}\frac{\partial \log L}{\partial \sigma} &= \sigma^{-1} \sum_{i=1}^n \{z_i g(z_i) - \delta_i\}, \\ \frac{\partial \log L}{\partial \alpha_j} &= \sigma^{-1} \sum_{i=1}^n x_{jig}(z_i),\end{aligned}$$

where the function  $g(z_i)$  is given by

$$g(z_i) = \frac{(1 - \delta_i)f_{\epsilon_i}(z_i)}{S_{\epsilon_i}(z_i)} - \frac{\delta_i f'_{\epsilon_i}(z_i)}{f_{\epsilon_i}(z_i)},$$

and  $f'_{\epsilon_i}(z_i)$  is the derivative of  $f_{\epsilon_i}(z_i)$  with respect to  $z_i$ .

The  $i$ th component of each derivative, evaluated at the maximum likelihood estimates of the unknown parameters, is then the score residual for the corresponding term. Consequently, from the definition of the standardised residual in equation (7.1), the  $i$ th score residual for  $\mu$  is

$$\hat{\sigma}^{-1} g(r_{Si}),$$

that for the scale parameter,  $\sigma$ , is

$$\hat{\sigma}^{-1} \{r_{Si} g(r_{Si}) - \delta_i\},$$

and that for the  $j$ th explanatory variable in the model,  $X_j$ , is

$$\hat{\sigma}^{-1} x_{ji} g(r_{Si}).$$

Of these, the score residuals for  $X_j$  are the most important, and so specific expressions are only given for this residual when particular models are considered in the sequel. Because the sums of score residuals are the derivatives of the log-likelihood function at its maximum, these residuals must sum to zero.

## 7.2 Residuals for particular parametric models

In this section, the form of the residuals for parametric models based on Weibull, log-logistic and lognormal distributions for the survival times are described.

### 7.2.1 Weibull distribution

The residuals described in Section 7.1 may be used in conjunction with either the proportional hazards or the accelerated failure time representations of the Weibull model. We begin with the proportional hazards model described in Chapter 5, according to which the hazard of death at time  $t$  for the  $i$ th individual is

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

where  $h_0(t) = \lambda \gamma t^{\gamma-1}$  is the baseline hazard function. The corresponding estimate of the cumulative hazard function is

$$\hat{H}_i(t) = \exp(\hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_p x_{pi}) \hat{\lambda} t^{\hat{\gamma}}$$

which are the Cox-Snell residuals, as defined in equation (7.3).

In the accelerated failure time form of the model,  $\epsilon_i$  has a Gumbel distribution, with survivor function

$$S_{\epsilon_i}(\epsilon) = \exp(-e^\epsilon). \quad (7.6)$$

The standardised residuals are then as given in equation (7.1), and if an appropriate model has been fitted, these will be expected to behave as a possibly censored sample from a Gumbel distribution. This is equivalent to assessing whether the Cox-Snell residuals, defined below, have a unit exponential distribution.

The Cox-Snell residuals,  $r_{Ci} = -\log S_{\epsilon_i}(r_{Si})$ , are, from equation (7.6), simply the exponentiated standardised residuals, that is  $r_{Ci} = \exp(r_{Si})$ . These residuals lead immediately to the martingale and deviance residuals for the Weibull model, using equations (7.4) and (7.5).

The score residuals for the Weibull model are found from the general results in Section 7.1.5. In particular, the  $i$ th score residual for the  $j$ th explanatory variable in the model,  $X_j$ , is

$$\hat{\sigma}^{-1} x_{ji} (e^{r_{Si}} - \delta_i),$$

where  $r_{Si}$  is the  $i$ th standardised residual. We also note that the  $i$ th score residual for  $\mu$  is  $\hat{\sigma}^{-1}(e^{r_{Si}} - \delta_i)$ , which is  $\hat{\sigma}^{-1}(r_{Ci} - \delta_i)$ . Since these score residuals sum to zero, it follows that the sum of the martingale residuals, defined in equation (7.4), must be zero in the Weibull model.

### 7.2.2 Log-logistic distribution

In the log-logistic accelerated failure time model, the random variable  $\epsilon_i$  has a logistic distribution with survivor function

$$S_{\epsilon_i}(\epsilon) = (1 + e^\epsilon)^{-1}.$$

Accordingly, the standardised residuals, obtained from equation (7.1), should behave as a sample from a logistic distribution, if the fitted model is correct. Equivalently, the Cox-Snell residuals for the log-logistic accelerated failure time model are given by

$$r_{Ci} = -\log S_{\epsilon_i}(r_{Si}),$$

that is,

$$r_{Ci} = \log \{1 + \exp(r_{Si})\},$$

where  $r_{Si}$  is the  $i$ th standardised residual. The score residuals are found from the general results in Section 7.1.5, and we find that the  $i$ th score residual for the  $j$ th explanatory variable in the model is

$$\hat{\sigma}^{-1} x_{ji} \left\{ \frac{\exp(r_{Si}) - \delta_i}{1 + \exp(r_{Si})} \right\}.$$

### 7.2.3 Lognormal distribution

If the survival times are assumed to have a lognormal distribution, then  $\epsilon_i$  in the log-linear formulation of the accelerated failure time model is normally distributed. The estimated survivor function for the  $i$ th individual, from equation (6.24), is

$$\hat{S}_i(t) = 1 - \Phi\left(\frac{\log t - \hat{\mu} - \hat{\alpha}_1 x_{1i} - \hat{\alpha}_2 x_{2i} - \cdots - \hat{\alpha}_p x_{pi}}{\hat{\sigma}}\right),$$

and so the Cox-Snell residuals become

$$r_{Ci} = -\log\{1 - \Phi(r_{Si})\},$$

where, as usual,  $r_{Si}$  is the  $i$ th standardised residual in equation (7.1). Again the martingale and deviance residuals are obtained from these, and the score residuals are obtained from the results in Section 7.1.5. Specifically, the  $i$ th score residual for  $X_j$ , is

$$\hat{\sigma}^{-1} \left\{ \frac{(1 - \delta_i) f_{\epsilon_i}(r_{Si})}{1 - \Phi(r_{Si})} + \delta_i r_{Si} \right\},$$

where  $f_{\epsilon_i}(r_{Si})$  is the standard normal density function at  $r_{Si}$ , and  $\Phi(r_{Si})$  is the corresponding distribution function.

### 7.2.4 Analysis of residuals

In the analysis of residuals after fitting parametric model to survival data, one of the most useful plots is based on comparing the distribution of the Cox-Snell residuals with the unit exponential distribution. As noted in Section 7.1.1, this is equivalent to comparing the distribution of the standardised residuals with that of the random variable  $\epsilon_i$  in the log-linear form of the accelerated failure time model. This comparison is made using a cumulative hazard, or log-cumulative hazard, plot of the residuals, as shown in Section 4.2.1 of Chapter 4, where the use of this plot in connection with residuals after fitting the Cox regression model was described. In summary, the Kaplan-Meier estimate of the survivor function of the Cox-Snell residuals, denoted  $\hat{S}(r_{Ci})$ , is obtained, and  $-\log \hat{S}(r_{Ci})$  is plotted against  $r_{Ci}$ . A straight line with unit slope and zero intercept should be obtained if the fitted model is appropriate. Alternatively, a log-cumulative hazard plot of the residuals is obtained by plotting  $\log\{-\log \hat{S}(r_{Ci})\}$  against  $\log r_{Ci}$ , which will also give a straight line with unit slope and passing through the origin, if the fitted survival model is satisfactory.

In Section 4.2.1, substantial criticisms were levied against the use of this plot. However, these criticisms do not have as much force for residuals derived from parametric models. The reason for this is that the non-parametric estimate of the baseline cumulative hazard function, used in the Cox regression model, is now replaced by an estimate of a parametric function. This function usually depends on just two parameters,  $\mu$  and  $\sigma$ , and so fewer parameters are

being estimated when an accelerated failure time model is fitted to survival data. The Cox-Snell residuals for a parametric model are therefore much more likely to be approximated by a unit exponential distribution, when the correct model has been fitted.

Other residual plots that are useful include index plots of martingale or deviance residuals, which can be used to identify observations not well fitted by the model. A plot of martingale or deviance residuals against the survival times, the rank order of the times, or explanatory variables, shows whether there are particular times, or particular values of explanatory variables, for which the model is not a good fit. Plots of martingale or deviance residuals against the estimated acceleration factor,  $\exp(-\hat{\alpha}' x_i)$ , or simply the estimated linear component of the accelerated failure time model,  $\hat{\alpha}' x_i$ , also provide information about the relationship between the residuals and the likely survival time of an individual. Those with large values of the estimated acceleration factor will tend to have shorter survival times. Index plots of score residuals, or plots of these residuals against the survival times, or the rank order of the survival times, might be examined in a more comprehensive assessment of model adequacy.

#### Example 7.1 Chemotherapy in ovarian cancer patients

In Example 5.10 of Chapter 5, data on the survival times of patients with ovarian cancer were presented. The data were analysed using a Weibull proportional hazards model, and the model chosen contained variables corresponding to the age of the woman,  $Age$ , and the treatment group to which the woman was assigned,  $Treat$ . In the accelerated failure time representation of the model, the estimated survivor function for the  $i$ th woman is

$$\hat{S}_i(t) = S_{\epsilon_i}\left(\frac{\log t - \hat{\mu} - \hat{\alpha}_1 Age_i - \hat{\alpha}_2 Treat_i}{\hat{\sigma}}\right),$$

where  $S_{\epsilon_i}(\epsilon) = \exp(-e^{\epsilon})$ , so that

$$\hat{S}_i(t) = \exp\left\{-\exp\left(\frac{\log t - 10.4254 + 0.0790 Age_i - 0.5615 Treat_i}{0.5489}\right)\right\}.$$

The standardised residuals are the values of

$$r_{Si} = (\log t_i - 10.4254 + 0.0790 Age_i - 0.5615 Treat_i)/0.5489,$$

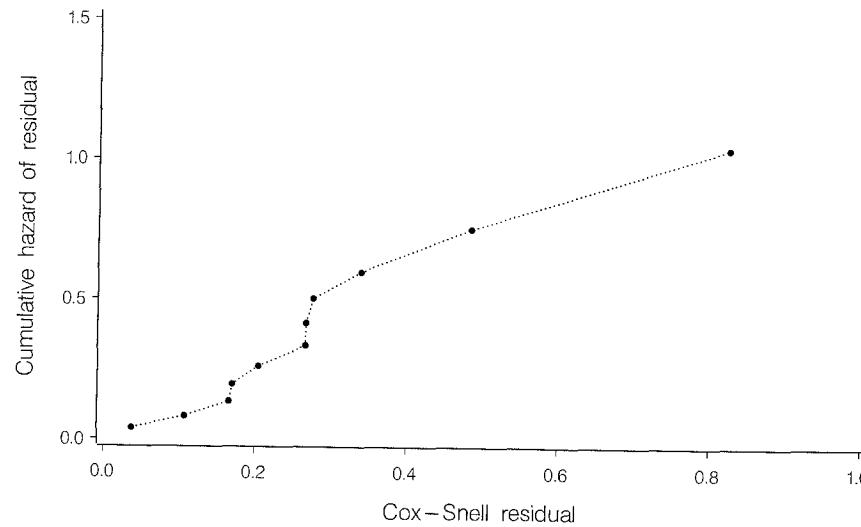
for  $i = 1, 2, \dots, 26$ , and these are given in Table 7.1. Also given are the values of the Cox Snell residuals, which for the Weibull model, are such that  $r_{Ci} = \exp(r_{Si})$ .

A cumulative hazard plot of the Cox-Snell residuals is given in Figure 7.1. In this plot, the plotted points lie on a line that has an intercept and slope close to zero and unity, respectively. However, there is some evidence of a systematic deviation from the straight line, giving some cause for concern about the adequacy of the fitted model.

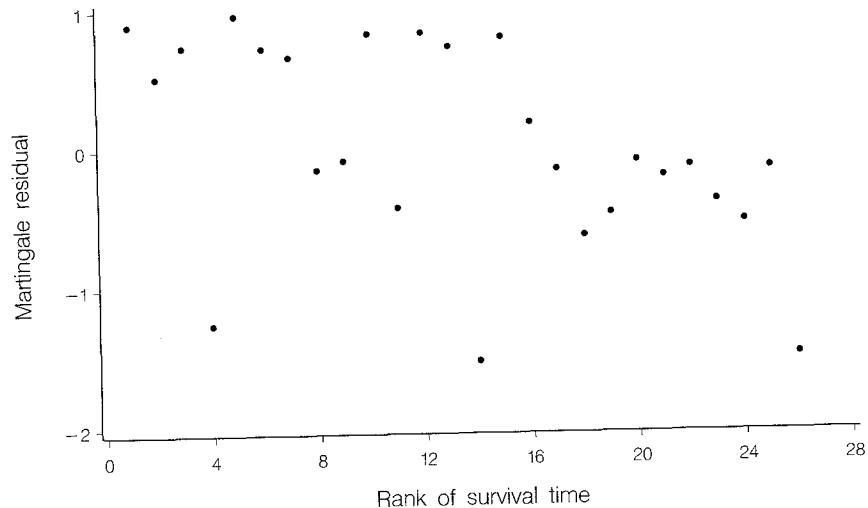
Plots of the martingale and deviance residuals against the rank order of the survival times are shown in Figures 7.2 and 7.3, respectively. Both of these

**Table 7.1** Values of the standardised and Cox-Snell residuals for 26 ovarian cancer patients.

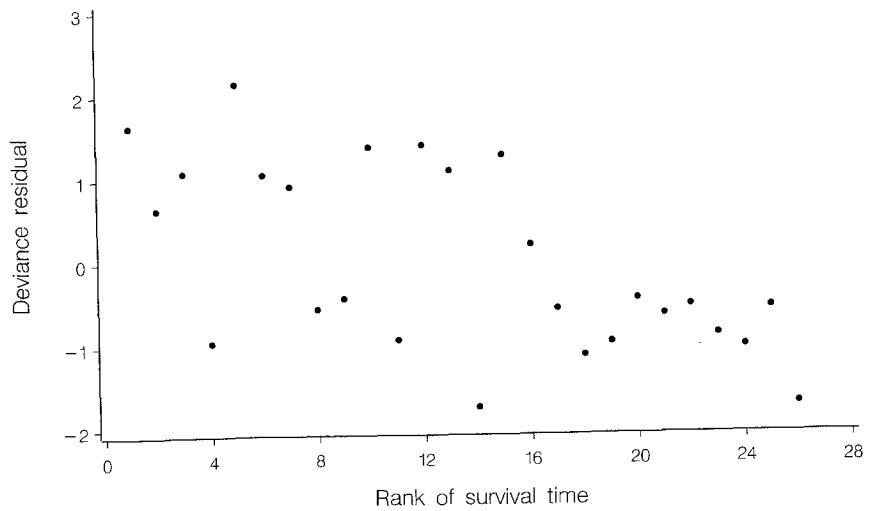
Patient	$r_{Si}$	$r_{Ci}$	Patient	$r_{Si}$	$r_{Ci}$
1	-1.320	0.267	14	-1.782	0.168
2	-1.892	0.151	15	-0.193	0.825
3	-2.228	0.108	16	-1.587	0.204
4	-2.404	0.090	17	-0.917	0.400
5	-3.270	0.038	18	-1.771	0.170
6	-0.444	0.642	19	-1.530	0.217
7	-1.082	0.339	20	-2.220	0.109
8	-0.729	0.482	21	-0.724	0.485
9	0.407	1.503	22	-1.799	0.165
10	0.817	2.264	23	0.429	1.535
11	-1.321	0.267	24	-0.837	0.433
12	-0.607	0.545	25	-1.287	0.276
13	-1.796	0.166	26	-1.886	0.152



**Figure 7.1** Cumulative hazard plot of the Cox-Snell residuals.



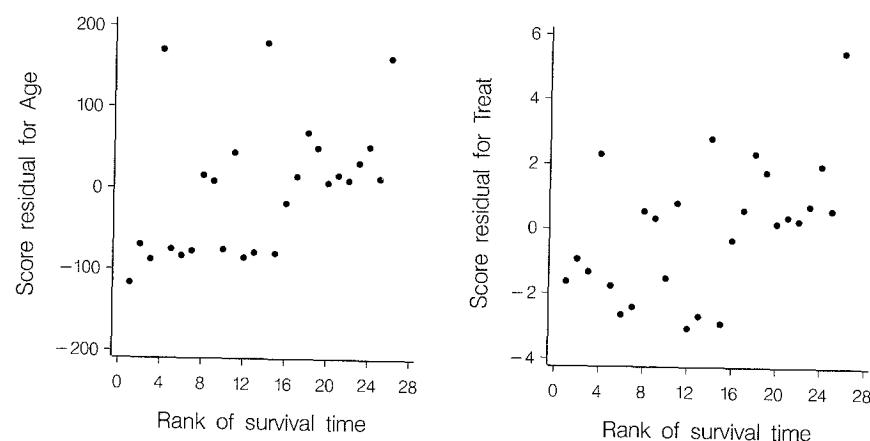
**Figure 7.2** Plot of the martingale residuals against rank order of survival time.



**Figure 7.3** Plot of the deviance residuals against rank order of survival time.

plots show a slight tendency for observations with longer survival times to have smaller residuals, but these are also the observations that are censored.

The graphs in Figure 7.4 show the score residuals for the two variables in the model, *Age* and *Treat*, plotted against the rank order of the survival times.



**Figure 7.4** Score residuals plotted against rank order of survival time for *Age* and *Treat*.

The plot of the score residuals for *Age* shows that there are three observations with relatively large residuals. These correspond to patients 14, 4 and 26 in the original data set given in Table 5.6. However, there does not appear to be anything unusual about these observations. The score residual for *Treat* for patient 26 is also somewhat larger than the others. This points to the fact that the model is not a good fit to the data from patients 14, 4 and 26.

### 7.3 Comparing observed and fitted survivor functions

In parametric modelling, the estimated survivor function is a continuous function of the survival time,  $t$ , and so this function can be plotted for particular values of the explanatory variables included in the model. When there is just a single sample of survival data, with no explanatory variables, the fitted survivor function can be compared directly with the Kaplan-Meier estimate of the survivor function, described in Section 2.1.2 of Chapter 2. If the fitted survivor function is close to the Kaplan-Meier estimate, which is a step function, the fitted model is an appropriate summary of the data. Similarly, suppose that the model incorporates one or two factors that classify individuals according to treatment group, or provide a cross-classification of treatment group and gender, say. For each group of individuals defined by the combinations of levels of the factors in the model, the fitted survivor function can then be compared with the corresponding Kaplan-Meier estimate of the survivor function.

In situations where the values of a number of explanatory variables are

recorded, groups of individuals are formed from the values of the estimated linear component of the fitted model. For the  $i$ th individual, whose values of the explanatory variables in the model are  $\mathbf{x}_i$ , this is just the risk score in a proportional hazards model,  $\hat{\beta}' \mathbf{x}_i$ , or the value of  $\hat{\alpha}' \mathbf{x}_i$  in an accelerated failure time model. The following discussion is based on the risk score, where large positive values correspond to a greater hazard, but it could equally well be based on the linear component of an accelerated failure time model, or the value of the acceleration factor.

The values of the risk score for each individual are arranged in increasing order, and these values are used to divide the individuals into a number of groups. For example, if three groups were used, there would be individuals with low, medium and high values of the risk score. The actual number of groups formed in this way will depend on the size of the data base. For larger data bases, five or even seven groups might be constructed. In particular, with five groups, there would be 20% of the individuals in each group; those with the lowest and highest values of the risk score would be at low and high risk, respectively, while the middle 20% would be of medium risk.

The next step is to compare the observed and fitted survivor functions in each of the groups. Suppose that  $\hat{S}_{ij}(t)$  is the model-based estimate of the survivor function for the  $i$ th individual in the  $j$ th group. The average fitted survivor function is then obtained for each group, or just the groups with the smallest, middle and highest risk scores, from

$$\bar{S}_j(t) = \frac{1}{n_i} \sum_{i=1}^{n_i} \hat{S}_{ij}(t),$$

where  $n_i$  is the number of observations in the  $j$ th group. The value of  $\bar{S}_j(t)$  would be obtained for a range of  $t$  values, so that a plot of the values of  $\bar{S}_j(t)$  against  $t$ , for each value of  $j$ , yields a smooth curve. The corresponding observed survivor function for a particular group is the Kaplan-Meier estimate of the survivor function for the individuals in that group. Superimposing these two sets of estimates gives a visual representation of the agreement between the observed and fitted survivor functions.

Using this approach, it is often easier to detect departures from the fitted model, than from plots based on residuals. However, the procedure can be criticised for using the same fitted model to define the groups, and to obtain the estimated survivor function for each group. If the data base is sufficiently large, the survivor function could be estimated from half of the data, and the fit of the model evaluated on the remaining half. Also, since the method is based on the values of the risk score, no account is taken of differences between individuals who have different sets of values of the explanatory variables, but just happen to have the same value of the risk score.

#### Example 7.2 Chemotherapy in ovarian cancer patients

In this example, we examine the fit of a Weibull proportional hazards model to the data on the survival times of 26 women, following treatment for ovarian

cancer. A Weibull model that contains the variables *Age* and *Treat* is fitted, as in Example 5.10, so that the fitted survivor function for the  $i$ th individual is

$$\hat{S}_i(t) = \exp \left\{ -e^{\hat{\eta}_i} \hat{\lambda} t^{\hat{\gamma}} \right\}, \quad (7.7)$$

where  $\hat{\eta}_i = 0.144 \text{Age}_i - 1.023 \text{Treat}_i$  is the risk score,  $i = 1, 2, \dots, 26$ . This is equivalent to the accelerated failure time representation of the model, used in Example 7.1.

The values of  $\hat{\eta}_i$  are then arranged in ascending order and divided into three groups, as shown in Table 7.2.

**Table 7.2** Values of the risk score, with the patient number in parentheses, for three groups of ovarian cancer patients.

Group	Risk score				
1 (low risk)	4.29 (14)	4.45 (2)	4.59 (20)	5.15 (22)	5.17 (5)
	5.17 (19)	5.31 (17)	5.59 (4)	5.59 (12)	
2 (medium risk)	5.87 (16)	6.02 (13)	6.16 (8)	6.18 (18)	6.31 (26)
	6.45 (6)	6.45 (9)	6.45 (11)	7.03 (25)	
3 (high risk)	7.04 (15)	7.04 (24)	7.17 (7)	8.19 (23)	
	8.48 (1)	9.34 (3)	9.63 (10)	9.63 (21)	

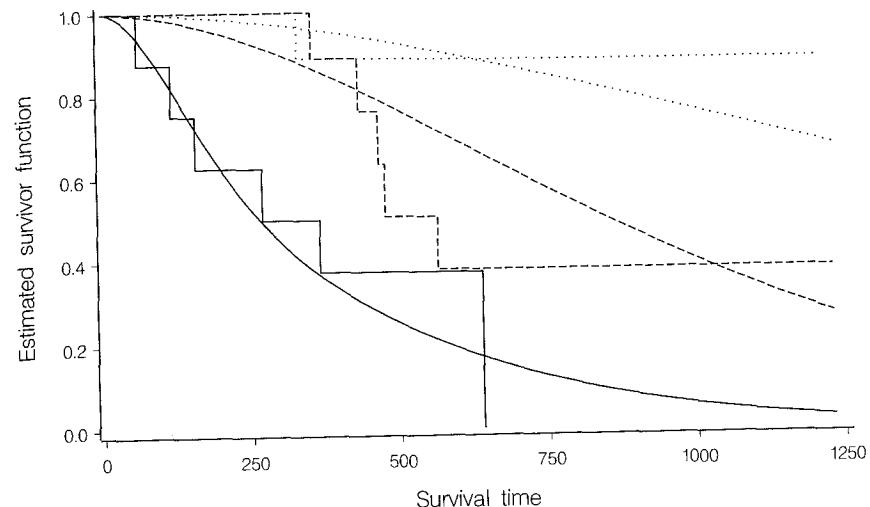
The next step is to obtain the average survivor function for each group by averaging the values of the estimated survivor function, in equation (7.7), for the patients in the three groups. This is done for  $t = 0, 1, \dots, 1230$ , and the three average survivor functions are shown in Figure 7.5. The Kaplan-Meier estimate of the survivor function for the individuals in each of the three groups shown in Table 7.2 is then calculated, and this is also shown in Figure 7.5.

From this plot, we see that the model is a good fit to the patients in the high-risk group. For those in the middle group, the agreement between the observed and fitted survivor functions is not that good, as the fitted model leads to estimates of the survivor function that are a little too high. In fact, the patients in this group have the largest values of the martingale residuals, which also indicates that the death times of these individuals are not adequately summarised by the fitted model. There is only one death among the individuals in the low-risk group, and so little can be said about the fit of the model to this set of patients.

#### 7.4\* Identification of influential observations

As when fitting the Cox regression model, it will be important to identify observations that exert an undue influence on particular parameter estimates, or on the complete set of parameter estimates. These two aspects of influence are considered in turn in this section.

A number of influence diagnostics for the Weibull proportional hazards



**Figure 7.5** Plot of the observed and fitted survivor functions for patients of low (dotted), medium (dashed) and high (—) risk. The observed survivor function is the step function.

model have been proposed by Hall *et al.* (1982), derived from the accelerated failure time representation of the model. However, they may also be used with other parametric models. These diagnostics are computed from the estimates of all  $p + 2$  parameters in the model, and their variance-covariance matrix. For convenience, the vector of  $p + 2$  parameters will be denoted by  $\boldsymbol{\theta}$ , so that  $\boldsymbol{\theta}' = (\mu, \alpha_1, \alpha_2, \dots, \alpha_p, \sigma)$ . The vector  $\hat{\boldsymbol{\theta}}'$  will be used to denote the corresponding vector of estimates of the parameters.

##### 7.4.1 Influence of observations on a parameter estimate

An approximation to the change in the estimated value of  $\theta_j$ , the  $j$ th component of the vector  $\boldsymbol{\theta}$ , on omitting the  $i$ th observation,  $\Delta_i \hat{\theta}_j$ , is the  $j$ th component of the  $(p + 2) \times 1$  vector

$$\mathbf{V}(\hat{\boldsymbol{\theta}}) \mathbf{s}_i. \quad (7.8)$$

In expression (7.8),  $\mathbf{V}(\hat{\boldsymbol{\theta}})$  is the estimated variance-covariance matrix of the parameters in  $\boldsymbol{\theta}$ , and  $\mathbf{s}_i$  is the  $(p + 2) \times 1$  vector of values of the first partial derivatives of the log-likelihood for the  $i$ th observation, with respect to the  $p + 2$  parameters in  $\boldsymbol{\theta}$ , evaluated at  $\hat{\boldsymbol{\theta}}$ . The vector  $\mathbf{s}_i$  is therefore the vector of values of the score residuals for the  $i$ th observation, defined in Section 7.1.5.

The quantities  $\Delta_i \hat{\alpha}_j$  are components 2 to  $p - 1$  of the vector in expression (7.8), which we will continue to refer to as delta-betas rather than as delta-alphas. These values may be standardised by dividing them by the standard error of  $\hat{\alpha}_j$ , leading to standardised delta-betas. Index plots or plots of

the standardised or unstandardised values of  $\Delta_i \hat{\alpha}_j$  provide informative summaries of this aspect of influence.

#### 7.4.2 Influence of observations on the set of parameter estimates

Two summary measures of the influence of the  $i$ th observation on the set of parameters that make up the vector  $\boldsymbol{\theta}$  have been proposed by Hall *et al.* (1982). These are the statistics  $F_i$  and  $C_i$ . The quantity  $F_i$  is given by

$$F_i = \frac{s'_i \mathbf{R}^{-1} s_i}{(p+2)\{1 - s'_i \mathbf{R}^{-1} s_i\}}, \quad (7.9)$$

where the  $(p+2) \times (p+2)$  matrix  $\mathbf{R}$  is the cross-product matrix of score residuals, that is,  $\mathbf{R} = \sum_{i=1}^n s_i s_i'$ . Equivalently,  $\mathbf{R} = \mathbf{S}' \mathbf{S}$ , where  $\mathbf{S}$  is the  $n \times (p+2)$  matrix whose  $i$ th row is the transpose of the vector of score residuals,  $s_i'$ . An alternative measure of the influence of the  $i$ th observation on the set of parameter estimates is the statistic

$$C_i = \frac{s'_i \mathbf{V}(\hat{\boldsymbol{\theta}}) s_i}{\{1 - s'_i \mathbf{V}(\hat{\boldsymbol{\theta}}) s_i\}^2}. \quad (7.10)$$

The statistics  $F_i$  and  $C_i$  will typically have values that are quite different from each other. However, in each case a relatively large value of the statistic will indicate that the corresponding observation is influential. Exactly how such observations influence the estimates would need to be investigated by omitting that observation from the data set and refitting the model.

#### Example 7.3 Chemotherapy in ovarian cancer patients

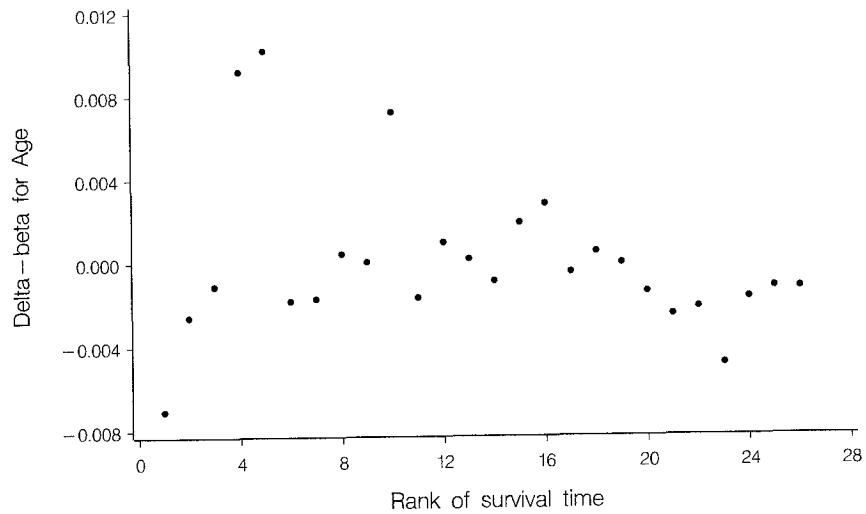
We now go on to investigate whether there are any influential observations in the data on the survival times following chemotherapy treatment for ovarian cancer.

The unstandardised delta-betas for *Age* and *Treat*, plotted against the rank order of the survival times, are shown in Figures 7.6 and 7.7. In Figure 7.6, two observations have relatively large values of the delta-beta for *Age*. These occur for patients 4 and 5 in the original data set. Both women have short survival times, and in addition one is relatively old at 74 years and the other relatively young at 43 years. The delta-betas for *Treat* displayed in Figure 7.7 show no unusual features.

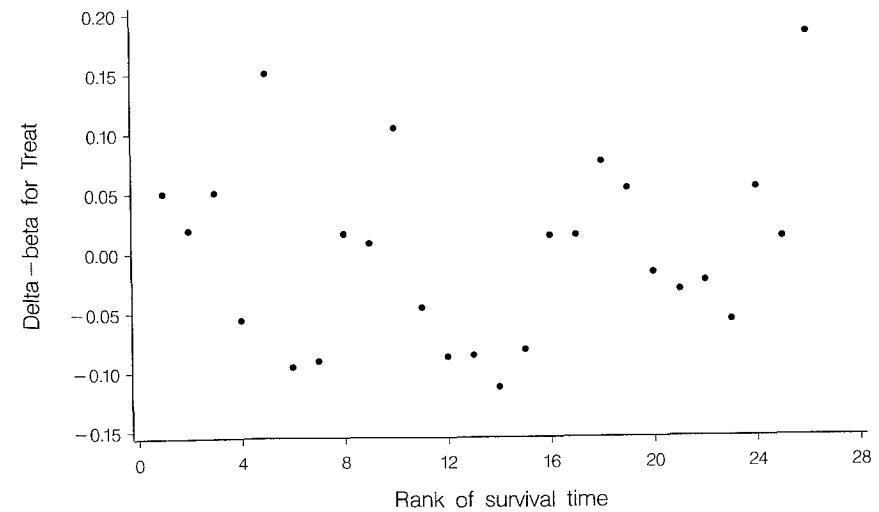
We next investigate the influence of each observation on the set of parameter estimates. The values of  $F_i$  and  $C_i$ , defined in equations (7.9) and (7.10), have been calculated using a SAS macro described in Section 12.3.5 of Chapter 12. Plots of the values of the  $F$ -statistic and the  $C$ -statistic, against the rank order of the survival times, are shown in Figures 7.8 and 7.9.

Figure 7.8 clearly shows that the observation corresponding to patient 5 is influential, and that the influence of patients 1, 4, 14 and 26 should be investigated in greater detail. Figure 7.9 strongly suggests that the data from patients 5 and 26 is influential.

The linear component of the fitted hazard function in the model fitted to



**Figure 7.6** Plot of the delta-betas for Age against rank order of survival time.



**Figure 7.7** Plot of the delta-betas for Treat against rank order of survival time.

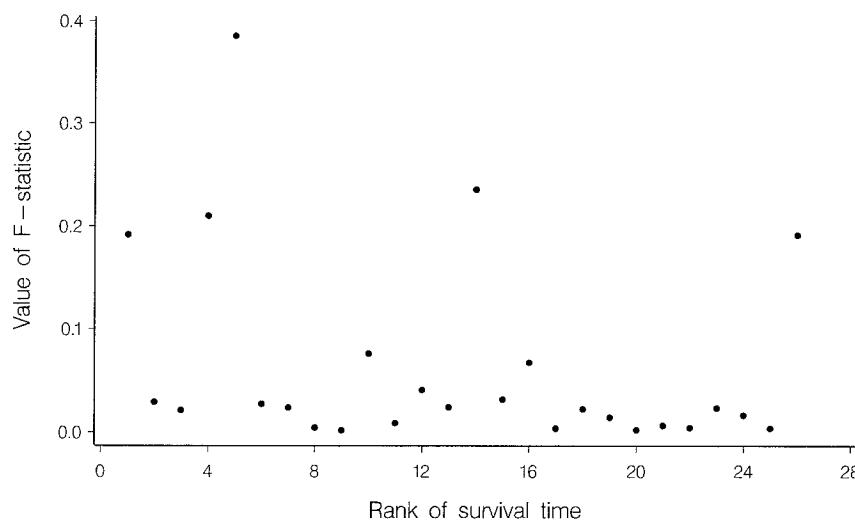


Figure 7.8 Plot of the  $F$ -statistic against rank order of survival time.

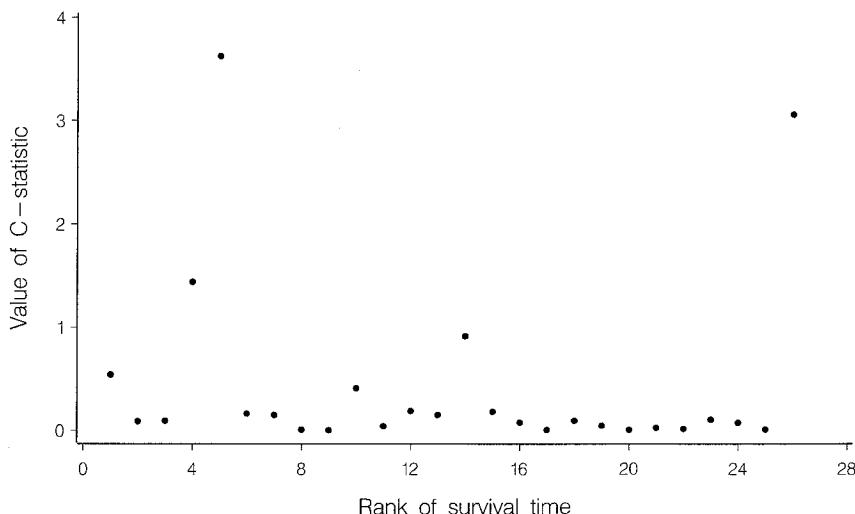


Figure 7.9 Plot of the  $C$ -statistic against rank order of survival time.

all 26 patients is

$$0.144 \text{Age}_i - 1.023 \text{Treat}_i,$$

while that on omitting each of observations 1, 4, 5, 14 and 26 in turn is as follows:

$$\text{Omitting patient number 1: } 0.142 \text{Age}_i - 1.016 \text{Treat}_i,$$

$$\text{Omitting patient number 4: } 0.175 \text{Age}_i - 1.190 \text{Treat}_i,$$

$$\text{Omitting patient number 5: } 0.177 \text{Age}_i - 0.710 \text{Treat}_i,$$

$$\text{Omitting patient number 14: } 0.149 \text{Age}_i - 1.318 \text{Treat}_i,$$

$$\text{Omitting patient number 26: } 0.159 \text{Age}_i - 0.697 \text{Treat}_i.$$

These results show that the effect of omitting the data from patient 1 on the parameter estimates is small. When the data from patient 4 are omitted, the estimated coefficient of  $\text{Age}$  is most affected, whereas when the data from patient 14 are omitted, the coefficient of  $\text{Treat}$  is changed the most. On leaving out the data from patients 5 and 26, both estimates are considerably affected.

The hazard ratio for a patient on the combined treatment ( $\text{Treat} = 2$ ), relative to one on the single treatment ( $\text{Treat} = 1$ ), is estimated by  $e^{-1.023} = 0.36$ , when the model is fitted to all 26 patients. When the observations from patients 1, 4, 5, 14 and 26 are omitted in turn, the estimated age-adjusted hazard ratios are 0.36, 0.30, 0.49, 0.27 and 0.50, respectively. The data from patients 5 and 26 clearly has the greatest effect on the estimated hazard ratio; in each case the estimate is increased, and the magnitude of the treatment effect is diminished. Omission of the data from patients 4 or 14 decreases the estimated hazard ratio, thereby increasing the estimated treatment difference.

## 7.5 Testing proportional hazards in the Weibull model

The Weibull model is most commonly used as a parametric proportional hazards model, and so it will be important to test that the proportional hazards assumption is tenable. In Section 4.4.3 of Chapter 4, it was shown how a time-dependent variable can be used in testing proportionality of hazards in the Cox regression model. Parametric models containing time-dependent variables are more complicated, and because software for fitting such models is not widely available, further details on this approach will not be given here.

In the Weibull model, the assumption of proportional hazards across a number of groups,  $g$ , say, corresponds to the assumption that the shape parameter  $\gamma$  in the baseline hazard function is the same in each group. One way of testing this assumption is to fit a separate Weibull model to each of the  $g$  groups, where the linear component of the model is the same in each case. The models fitted to the data from each group will then have different shape parameters as well as different scale parameters. The values of the statistic  $-2 \log \hat{L}$  for

each of these  $g$  separate models are then summed to give a value of  $-2 \log \hat{L}$  for a model that has different shape parameter for each group. Denote this by  $-2 \log \hat{L}_1$ . We then combine the  $g$  sets of data and fit a Weibull proportional hazards model that includes parameters associated with the group effect. This model then corresponds to there being a common shape parameter for each group. The inclusion of group effects in the model leads to there being different scale parameters for each group. The value of  $-2 \log \hat{L}$  for this model,  $-2 \log \hat{L}_0$ , say, is then compared with  $-2 \log \hat{L}_1$ . The difference between the values of these two statistics is the change in  $-2 \log \hat{L}$  due to constraining the Weibull shape parameters to be equal, and can be compared with a chi-squared distribution on  $g - 1$  degrees of freedom. If the difference is not significant, the assumption of proportional hazards is justified.

Some alternatives to the proportional hazards model are described in Chapter 6, and further comments on how to deal with situations in which the hazards are not proportional are given in Section 11.1 of Chapter 11. An example on the use of this method of testing the proportional hazards assumption is given below.

#### *Example 7.4 Chemotherapy in ovarian cancer patients*

Data from the study of survival following treatment for ovarian cancer, given in Example 5.10 of Chapter 5, are now used to illustrate the procedure for testing the assumption that the Weibull shape parameter is the same for the patients in each of the two treatment groups. The first step is to fit a Weibull proportional hazards model that contains *Age* alone to the data from the women in each treatment group. When such a model is fitted to the data from those on the single chemotherapy treatment, the value of the statistic  $-2 \log \hat{L}$  is 22.851, while that for the women on the combined treatment is 16.757. The sum of the two values of  $-2 \log \hat{L}$  is 39.608, which is the value of the statistic for a Weibull model with different shape parameters for the two treatment groups. The value of  $-2 \log \hat{L}$  for the model that constrains the shape parameters to be equal is 41.126. The change in  $-2 \log \hat{L}$  on constraining the shape parameters to be equal is therefore 1.52, which is not significant when compared with a chi-squared distribution on one degree of freedom. The two shape parameters may therefore be taken to be equal.

## 7.6 Further reading

There have been relatively few publications on model checking in parametric survival models, compared to the literature on model checking in the Cox regression model. Residuals and influence measures for the Weibull proportional hazards model are described by Hall *et al.* (1982). Hollander and Proschan (1979) show how to assess whether a sample of censored observations is drawn from a particular probability distribution. Weissfeld and Schneider (1994) describe and illustrate a number of residuals that can be used in conjunction with parametric models for survival data. Cohen and Barnett (1995) describe how the interpretation of cumulative hazard plots of residuals can be helped

by the use of simulated envelopes for the plots. Influence diagnostics for use in parametric survival modelling are given by Weissfeld and Schneider (1990) and Escobar and Meeker (1992). A SAS macro for evaluating influence diagnostics for the Weibull proportional hazards model is described by Escobar and Meeker (1988). These papers involve arguments based on local influence, a topic that is explored in general terms by Cook (1986), and reviewed in Rancel and Sierra (2001). A method for testing the assumptions of proportional hazards and accelerated failure times against a general model for the hazard function is presented by Ciampi and Etezadi-Amoli (1985). An interesting application of parametric modelling, based on data on times to reoffending by prisoners released on parole, which incorporates elements of model checking, is given by Copas and Heydari (1997).

## Time-dependent variables

In earlier chapters, we have seen how the dependence of the hazard function for an individual on the values of certain explanatory variables can be modelled. When explanatory variables are incorporated in a model for survival data, the values taken by such variables are those recorded at the time origin of the study. For example, consider the study to compare two treatments for prostatic cancer first described in Example 1.4 of Chapter 1. Here, the age of a patient, serum haemoglobin level, size of the tumour, value of the Gleason index, and of course the treatment group, were all recorded at the time when a patient was entered into the study. The impact of these variables on the hazard of death is then evaluated.

In many studies that generate survival data, individuals are monitored for the duration of the study. During this period, the values of certain explanatory variables may be recorded on a regular basis. Thus, in the example on prostatic cancer, the size of the tumour, and other variables, may be recorded at frequent intervals. If account can be taken of the values of explanatory variables as they evolve, a more satisfactory model for the hazard of death at any given time would be obtained. For example, in connection with the prostatic cancer study, more recent values of the size of the tumour may provide a better indication of future life expectancy than the value at the time origin.

Variables whose values change over time are known as *time-dependent variables*, and in this chapter we see how such variables can be incorporated in models used in the analysis of survival data. In this process, the most recent value of a time-dependent variable is used at each specific time in the modelling procedure.

### 8.1 Types of time-dependent variables

It is useful to consider two types of variables that change over time, which may be referred to as *internal variables* and *external variables*.

Internal variables relate to a particular individual in a study, and can only be measured while a patient is alive. Such data arises when repeated measurements of certain characteristics are made on a patient over time, and examples include measures of lung function such as vital capacity and peak flow rate, white blood cell count, systolic blood pressure and serum cholesterol level. Variables that describe changes in the status of a patient are also of this type. For example, following a bone marrow transplant, a patient may be susceptible to the development of graft versus host disease. A binary explanatory

variable, that reflects whether the patient is suffering from this life-threatening side effect at any given time, is a further example of an internal variable. In each case, such variables reflect the condition of the patient and their values may well be associated with the survival time of the patient.

On the other hand, external variables are time-dependent variables that do not necessarily require the survival of a patient for their existence. One type of external variable is a variable that changes in such a way that its value will be known in advance at any future time. The most obvious example is the age of a patient, in that once the age at the time origin is known, that patient's age at any future time will be known exactly. However, there are other examples, such as the dose of a drug that is to be varied in a predetermined manner during the course of a study, or planned changes to the type of immunosuppressant to be used following organ transplantation. Another type of external variable is one that exists totally independently of any particular individual, such as the level of atmospheric sulphur dioxide, or air temperature. Changes in the values of such quantities may well have an effect on the lifetime of individuals, as in studies concerning the management of patients with certain types of respiratory disease.

Time-dependent variables also arise in situations where the coefficient of a time-constant explanatory variable is a function of time. In Section 3.7 of Chapter 3, it was explained that the coefficient of an explanatory variable in the Cox proportional hazards model is a log-hazard ratio, and so under this model, the hazard ratio is constant over time. If this ratio were in fact a function of time, then the coefficient of the explanatory variable that varies with time is referred to as a *time-varying coefficient*. In this case, the log-hazard ratio is not constant and so we no longer have a proportional hazards model. More formally, suppose that the coefficient of an explanatory variable,  $X$ , is a linear function of time,  $t$ , so that we may write the term as  $\beta t X$ . This means that the corresponding log-hazard ratio is a linear function of time. This was precisely the sort of term introduced into the model in order to test the assumption of proportional hazards in Section 4.4.3 of Chapter 4. This term can also be written as  $\beta X(t)$ , where  $X(t) = Xt$  is a time-dependent variable. In general, suppose that a model includes the explanatory variable,  $X$ , with a time-varying coefficient of the form  $\beta(t)$ . The corresponding term in the model would be  $\beta(t)X$ , which can be expressed as  $\beta X(t)$ . In other words, a term that involves a time-varying coefficient can be expressed as a time-dependent variable with a constant coefficient. However, if  $\beta(t)$  is a non-linear function of one or more unknown parameters, for example  $\beta_0 \exp(\beta_1 t)$ , the term is not easily fitted in a model.

These different types of time-dependent variables can be introduced into the Cox proportional hazards model. The resulting model will simply be referred to as the Cox regression model, and is described in the following section.

## 8.2 A model with time-dependent variables

According to the Cox proportional hazards model described in Chapter 3, the

hazard of death at time  $t$  for the  $i$ th of  $n$  individuals in a study can be written in the form

$$h_i(t) = \exp \left\{ \sum_{j=1}^p \beta_j x_{ji} \right\} h_0(t),$$

where  $x_{ji}$  is the baseline value of the  $j$ th explanatory variable,  $X_j$ ,  $j = 1, 2, \dots, p$ , for the  $i$ th individual,  $i = 1, 2, \dots, n$ , and  $h_0(t)$  is the baseline hazard function. Generalising this model to the situation in which some of the explanatory variables are time-dependent, we write  $x_{ji}(t)$  for the value of the  $j$ th explanatory variable at time  $t$ , in the  $i$ th individual. The Cox regression model then becomes

$$h_i(t) = \exp \left\{ \sum_{j=1}^p \beta_j x_{ji}(t) \right\} h_0(t). \quad (8.1)$$

In this model, the baseline hazard function  $h_0(t)$  is interpreted as the hazard function for an individual for whom all the variables are zero at the time origin, and remain at this same value through time.

It is important to note that in the model given in equation (8.1), the values of the variables  $x_{ji}(t)$  depend on the time  $t$ , and so the relative hazard  $h_i(t)/h_0(t)$  is also time-dependent. This means that the hazard of death at time  $t$  is no longer proportional to the baseline hazard, and the model is no longer a proportional hazards model.

To provide an interpretation of the  $\beta$ -parameters in this model, consider the ratio of the hazard functions at time  $t$  for two individuals, the  $r$ th and  $s$ th, say. This is given by

$$\frac{h_r(t)}{h_s(t)} = \exp [\beta_1 \{x_{r1}(t) - x_{s1}(t)\} + \dots + \beta_p \{x_{rp}(t) - x_{sp}(t)\}].$$

The coefficient  $\beta_j$ ,  $j = 1, 2, \dots, p$ , can therefore be interpreted as the log-hazard ratio for two individuals whose value of the  $j$ th explanatory variable at any time  $t$  differs by one unit, with the two individuals having the same values of all the other  $p - 1$  variables at that time.

### 8.2.1\* Fitting the Cox model

When the Cox regression model is extended to incorporate time-dependent variables, the partial log-likelihood function, from equation (3.5) in Chapter 3, can be generalised to

$$\sum_{i=1}^n \delta_i \left\{ \sum_{j=1}^p \beta_j x_{ji}(t_i) - \log \sum_{l \in R(t_i)} \exp \left( \sum_{j=1}^p \beta_j x_{jl}(t_i) \right) \right\}, \quad (8.2)$$

in which  $R(t_i)$  is the risk set at time  $t_i$ , the death time of the  $i$ th individual in the study,  $i = 1, 2, \dots, n$ , and  $\delta_i$  is an event indicator that is zero if the survival time of the  $i$ th individual is censored and unity otherwise. This expression can then be maximised to give estimates of the  $\beta$ -parameters.

In order to use equation (8.1) in this maximisation process, the values of each of the variables in the model must be known at each death time for all individuals in the risk set at time  $t_i$ . This is no problem for external variables whose values are preordained, but it may be a problem for external variables that exist independently of the individuals in a study, and certainly for internal variables.

To illustrate the problem, consider a trial of two maintenance therapies for patients who have suffered a myocardial infarct. The serum cholesterol level of such patients may well be measured at the time when a patient is admitted to the study, and at regular intervals of time thereafter. This variable is then a time-dependent variable, and will be denoted  $X(t)$ . It is then plausible that the hazard of death for any particular patient, the  $i$ th, say, at time  $t$ ,  $h_i(t)$ , is more likely to be influenced by the value of the explanatory variable  $X(t)$  at time  $t$ , than its value at the time origin, where  $t = 0$ .

Now suppose that the  $i$ th individual dies at time  $t_i$  and that there are two other individuals, labelled  $r$  and  $s$ , in the risk set at time  $t_i$ . We further suppose that individual  $r$  dies at time  $t_r$ , where  $t_r > t_i$ , and that the survival time of individual  $s$ ,  $t_s$ , is censored at some time after  $t_r$ . The situation is illustrated graphically in Figure 8.1. In this figure, the vertical dotted lines refer to points in patient time when the value of  $X(t)$  is measured.

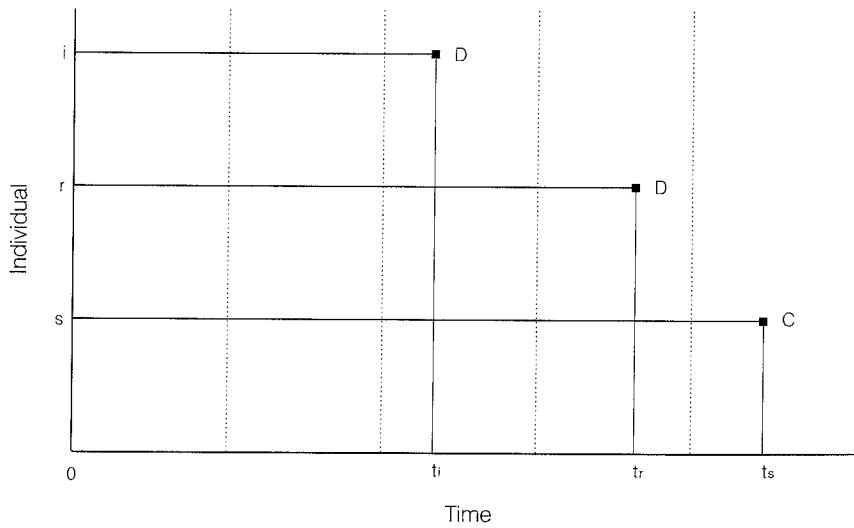


Figure 8.1 Survival times of three patients in patient time.

If individuals  $r$  and  $s$  are the only two in the risk set at time  $t_i$ , and  $X$  is the only explanatory variable that is measured, the contribution of the  $i$ th individual to the log-likelihood function in expression (8.2) will be

$$\beta x_i(t_i) - \log \sum_l \exp\{\beta x_l(t_i)\},$$

where  $x_l(t_i)$  is the value of  $X(t)$  for the  $l$ th individual at their death time,  $t_i$ , and  $l$  in the summation takes the values  $i$ ,  $r$ , and  $s$ . This expression is therefore equal to

$$\beta x_i(t_i) - \log \left\{ e^{\beta x_i(t_i)} + e^{\beta x_r(t_i)} + e^{\beta x_s(t_i)} \right\}.$$

This shows that the value of the time-dependent variable  $X(t)$  is needed at the death time of the  $i$ th individual, and at time  $t_i$  for individuals  $r$  and  $s$ . In addition, the value of the variable  $X(t)$  will be needed for individuals  $r$  and  $s$  at  $t_r$ , the death time of individual  $r$ .

For terms in a model that are explicit functions of time, such as interactions between time and a variable or factor measured at baseline, there is no difficulty in obtaining the values of the time-dependent variables at any time for any individual. Indeed, it is usually straightforward to incorporate such variables in the Cox model when using statistical software that has facilities for dealing with time-dependent variables. For other variables, such as serum cholesterol level, the values of the time-dependent variable at times other than that at which it was measured has to be approximated. There are then several possibilities.

One option is to use the last recorded value of the variable before the time at which the value of the variable is needed. When the variable has been recorded for an individual before and after the time when the value is required, the value closest to that time might be used. Another possibility is to use linear interpolation between consecutive values of the variable. Figure 8.2 illustrates these approximations.

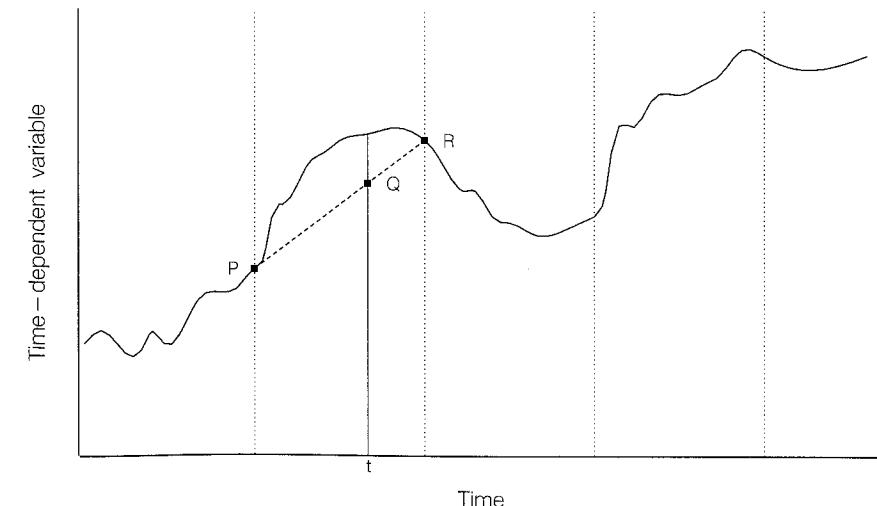


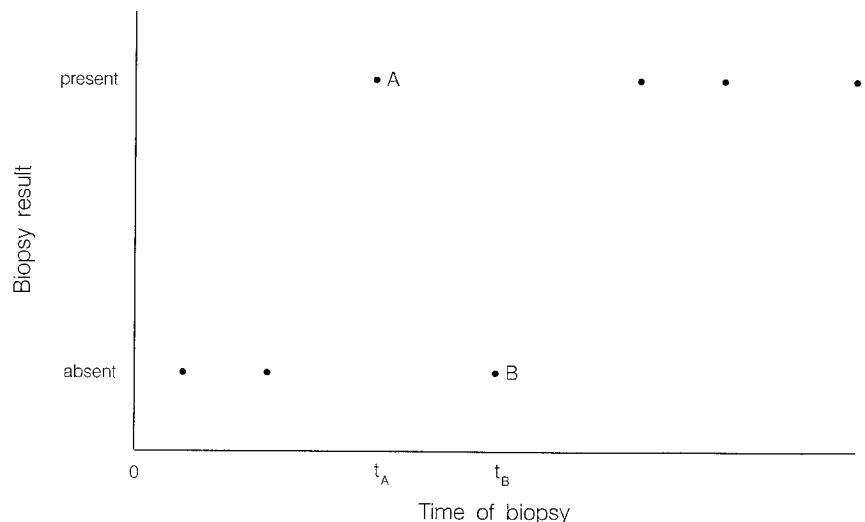
Figure 8.2 Computation of the value of a time-dependent variable at intermediate times.

In this figure, the continuous curve depicts the actual value of a time-

dependent variable at any time, and the dotted vertical lines signify times when the variable is actually measured. If the value of the variable is required at time  $t$  in this figure, we could use either the value at P, the last recorded value of the variable, the value at R, the value closest to  $t$ , or the value at Q, the linearly interpolated value between P and R.

Linear interpolation is clearly not an option when a time-dependent variable is a categorical variable. In addition, some categorical variables may be such that individuals can only progress through the levels of the variable in a particular direction. For example, the performance status of an individual may only be expected to deteriorate, so that the value of this categorical variable might only change from "good" to "fair" and from "fair" to "poor". As another example, following a biopsy, a variable associated with the occurrence of a tumour will take two values corresponding to absence and presence. It might then be very unlikely for the status to change from "present" to "absent" in consecutive biopsies.

Anomalous changes in the values of time-dependent variables can be detected by plotting the values of the variable against time for each patient. This may then lead on to a certain amount of data editing. For example, consider the plot in Figure 8.3, which shows the biopsy result, absent or present, for a particular patient at a number of time points. In this diagram, at least one of the observations made at times  $t_A$  and  $t_B$  must be incorrect. The observation at  $t_A$  might then be changed to "absent" or that at  $t_B$  to "present".



**Figure 8.3** Values of a time-dependent categorical variable.

If inferences of interest turn out to be sensitive to the method of interpolation used, extreme caution must be exercised when interpreting the results. Indeed, this feature could indicate that the value of the time-dependent

variable is subject to measurement error, substantial inherent variation, or perhaps the values have not been recorded sufficiently regularly.

### 8.2.2\* Estimation of baseline hazard and survivor functions

After a Cox regression model that includes time-dependent variables has been fitted, the baseline hazard function,  $h_0(t)$  and the corresponding baseline survivor function,  $S_0(t)$ , can be estimated. This involves an adaptation of the results given in Section 3.8 of Chapter 3 to cope with the additional complication of time-dependent variables, in which the values of the explanatory variables need to be updated with their time-specific values. In particular, the Nelson-Aalen estimate of the baseline cumulative hazard function, given in equation (3.25), becomes

$$\tilde{H}_0(t) = -\log \tilde{S}_0(t) = \sum_{j=1}^k \frac{d_j}{\sum_{l \in R(t_{(j)})} \exp\{\hat{\beta}' x_l(t)\}}, \quad (8.3)$$

for  $t_{(k)} \leq t < t_{(k+1)}$ ,  $k = 1, 2, \dots, r-1$ , where  $x_l(t)$  is the vector of values of the explanatory variables for the  $l$ th individual at time  $t$ , and  $d_j$  is the number of events at the  $j$ th ordered event time,  $t_{(j)}$ ,  $j = 1, 2, \dots, r$ . Similar modifications can be made to the other results in Section 3.8. With this modification, computation of the summation over the risk set is much more complicated, since for every event time,  $t_{(j)}$ ,  $j = 1, 2, \dots, r$ , the value of each time-dependent variable, for all individuals in the risk set, is needed at that event time.

Having obtained an estimate of the cumulative hazard function, the corresponding baseline hazard function can be estimated using equation (3.27), and an estimate of the baseline survivor function is  $\tilde{S}_0(t) = \exp\{-\tilde{H}_0(t)\}$ .

The survivor function for a particular individual is much more difficult to estimate. This is because the result that  $S_i(t)$  can be expressed as a power of the baseline survivor function,  $S_0(t)$ , given in equation (3.22) of Chapter 3, no longer holds. Instead, the survivor function for the  $i$ th individual is obtained from the integrated hazard function, which, from equation (1.6) in Chapter 1, is given by

$$S_i(t) = \exp \left\{ - \int_0^t \exp \left( \sum_{j=1}^p \beta_j x_{ji}(u) \right) h_0(u) du \right\}. \quad (8.4)$$

This survivor function therefore depends not only on the baseline hazard function  $h_0(t)$ , but also on the values of the time-dependent variables over the interval from 0 to  $t$ . The survivor function may therefore depend on future values of the time-dependent variables in the model, which will generally be unknown. However, approximate conditional probabilities of surviving a certain time interval can be found from the probability that an individual survives over an interval of time, from  $t$  to  $t+h$ , say, conditional on being alive at time  $t$ . This probability is  $P(T_i \geq t+h | T_i \geq t)$ , where  $T_i$  is the random variable associated with the survival time of the  $i$ th individual. Using the

standard result for conditional probability, given in Section 3.3.1 of Chapter 3, this probability becomes  $P(T_i \geq t+h)/P(T_i \geq t)$ , which is the ratio of the survivor functions at times  $t+h$  and  $t$ , that is,  $S_i(t+h)/S_i(t)$ . We now assume that any time-dependent variable remains constant through this interval, so that from equation (8.4), the approximate conditional probability is

$$\begin{aligned} P_i(t, t+h) &= \frac{\exp\left\{-\exp\left(\sum_{j=1}^p \beta_j x_{ji}(t)\right) \int_0^{t+h} h_0(u) du\right\}}{\exp\left\{-\exp\left(\sum_{j=1}^p \beta_j x_{ji}(t)\right) \int_0^t h_0(u) du\right\}}, \\ &= \exp\left[-\{H_0(t+h) - H_0(t)\} \exp\left(\sum_{j=1}^p \beta_j x_{ji}(t)\right)\right], \end{aligned}$$

where  $H_0(t)$  is the baseline cumulative hazard function. An estimate of this approximate conditional probability of surviving through the interval  $(t, t+h)$  is then

$$\tilde{P}_i(t, t+h) = \exp\left[-\left\{\tilde{H}_0(t+h) - \tilde{H}_0(t)\right\} \exp\left(\sum_{j=1}^p \hat{\beta}_j x_{ji}(t)\right)\right], \quad (8.5)$$

where  $\tilde{H}_0(t)$  is the estimated baseline cumulative hazard function obtained on fitting the Cox regression model with  $p$  possibly time-dependent variables with values  $x_{ji}(t)$ ,  $j = 1, 2, \dots, p$ , for the  $i$ th individual,  $i = 1, 2, \dots, n$ , and  $\hat{\beta}_j$  is the estimated coefficient of the  $j$ th time-dependent variable. This result was given by Altman and De Stavola (1994).

Corresponding estimates of the conditional probability of an event in the interval  $(t, t+h)$  are  $1 - \tilde{P}_i(t, t+h)$ , and these quantities can be used to obtain an estimate of the expected number of events in each of a number of successive intervals of width  $h$ . Comparing these values with the observed number of events in these intervals leads to an informal assessment of model adequacy.

### 8.3 Model comparison and validation

Models for survival data that include time-dependent variables can be compared in the same manner as Cox proportional hazards models, using the procedure described in Section 3.5 of Chapter 3. In particular, the model-fitting process leads to a maximised partial likelihood function, from which the value of the statistic  $-2 \log \hat{L}$  can be obtained. Changes in the value of this statistic between alternative nested models may then be compared to percentage points of the chi-squared distribution, with degrees of freedom equal to the difference in the number of  $\beta$ -parameters being fitted. For this reason, the model-building strategies discussed in Chapter 3 apply equally in situations where there are time-dependent variables.

#### 8.3.1 Comparison of treatments

In order to examine the magnitude of a treatment effect after taking account of variables that change over time, the value of  $-2 \log \hat{L}$  for a model that contains

the time-dependent variables and any other prognostic factors is compared with that for the model that contains the treatment term, in addition to these other variables. But in this analysis, if no treatment effect is revealed, one explanation could be that the time-dependent variable has *masked* the treatment difference.

To fix ideas, consider the example of a study to compare two cytotoxic drugs in the treatment of patients with leukaemia. Here, a patient's survival time may well depend on subsequent values of that patient's white blood cell count. If the effect of the treatment is to increase white blood cell count, no treatment difference will be identified after including white blood cell count as a time-dependent variable in the model. On the other hand, the treatment effect may appear in the absence of this variable. An interpretation of this is that the time-dependent variable has accounted for the treatment difference, and so provides an explanation as to how the treatment has been effective.

In any event, much useful information will be gained from a comparison of the results of an analysis that incorporates time-dependent variables with an analysis that uses baseline values alone.

#### 8.3.2 Assessing model adequacy

After fitting a model that includes time-dependent variables, a number of the techniques for evaluating the adequacy of the model, described in Chapter 4, can be implemented. In particular, an overall martingale residual can be computed for each subject, from an adaptation of the result in equation (4.6). The martingale residual for the  $i$ th subject is now

$$r_{Mi} = \delta_i - \exp\{\hat{\beta}' \mathbf{x}_i(t_i)\} \tilde{H}_0(t_i),$$

where  $\mathbf{x}_i(t_i)$  is the vector of values of explanatory variables for the  $i$ th individual, which may be time-dependent, evaluated at  $t_i$ , the event time of that individual. Also,  $\hat{\beta}$  is the vector of coefficients,  $\delta_i$  is the event indicator that takes the value unity if  $t_i$  is an event and zero otherwise, and  $\tilde{H}_0(t_i)$  is the estimated baseline cumulative hazard function at  $t_i$ , obtained from equation (8.3). The deviance residuals may also be computed from the martingale residuals, using equation (4.7) of Chapter 4.

The plots described in Section 4.2.2 of Chapter 4 will often be helpful. In particular, an index plot of the martingale residuals will enable outlying observations to be identified. However, diagnostic plots for assessing the functional form of covariates, described in Section 4.2.3, turn out to be not so useful when a time-dependent variable is being studied. This is because there will then be a number of values of the time-dependent covariate for any one individual, and it is not clear what the martingale residuals for the null model should be plotted against.

For detecting influential values, the delta-betas, introduced in Section 4.3.1 of Chapter 4, provide a helpful means of investigating the effect of each observation on individual parameter estimates. Changes in the value of the  $-2 \log \hat{L}$

statistic, on omitting each observation in turn, can give valuable information about the effect of each observation on the set of parameter estimates.

#### 8.4 Some applications of time-dependent variables

One application of time-dependent variables is in connection with evaluating the assumption of proportional hazards. This was discussed in detail in Section 4.4.3 of Chapter 4. In this application, a variable formed from the product of an explanatory variable,  $X$ , and time,  $t$ , is added to the linear part of the Cox model, and the null hypothesis that the coefficient of  $Xt$  is zero is tested. If this estimated coefficient is found to be significantly different from zero, there is evidence that the assumption of proportional hazards is not valid.

In many circumstances, the waiting time from the occurrence of some catastrophic event until a patient receives treatment may be strongly associated with the patient's survival. For example, in a study of factors affecting the survival of patients who have had a myocardial infarct, the time from the infarct to when the patient arrives in an Intensive Care Unit (ICU) may be crucial. Some patients may die before receiving treatment in the ICU, while those who arrive at the ICU soon after their infarct will tend to have a more favourable prognosis than those for whom treatment is delayed. It will be important to take account of this aspect of the data when assessing the effects of other explanatory variables on the survival times of these patients.

In a similar example, Crowley and Hu (1977) show how a time-dependent variable can be used in organ transplantation studies. Here, one feature of interest is the effect of a transplant on the patient's survival time. Suppose that in a study on the effectiveness of a particular type of organ transplant, a patient is judged to be suitable for a transplant at some time  $t_0$ . They then wait some period of time until a suitable donor organ is found, during which time the patients are unlikely to receive any beneficial therapy. Suppose that if the patient survives this period, they receive a transplant at time  $t_1$ .

In studies of this sort, the survival times of patients who have received a transplant cannot be compared with those who have not had a transplant in the usual way. The reason for this is that in order to receive a transplant, a patient must survive the waiting time to transplant. Consequently, the group who survive to transplant is not directly comparable with the group who receive no such transplant. Similarly, it is not possible to compare the times that the patients who receive a transplant survive after the transplant with the survival times of the group not receiving a transplant. Here, the time origin would be different for the two groups, and so the groups are not comparable at the time origin. This means that it is not possible to identify a time origin from which standard methods for survival analysis can be used.

The solution to this problem is to introduce a time-dependent variable  $X_1(t)$ , which takes the value zero if a patient has not received a transplant at time  $t$ , and unity otherwise. Adopting a Cox regression model, the hazard of

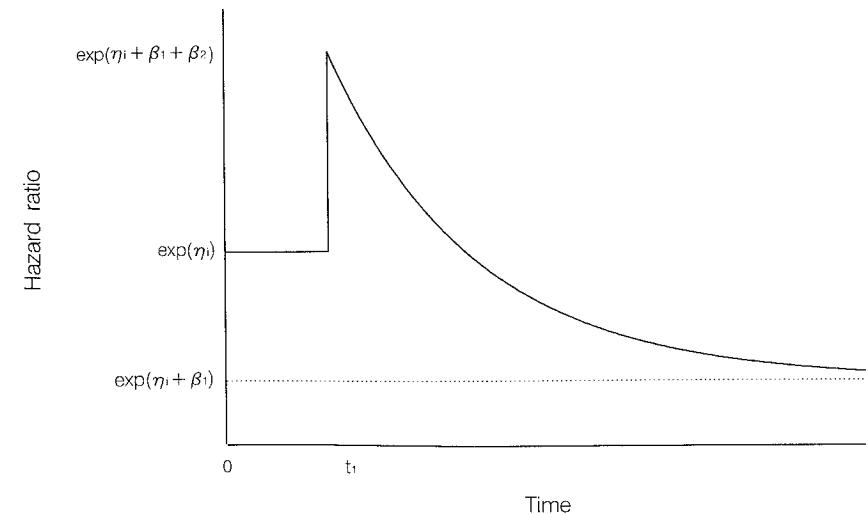
death for the  $i$ th individual at time  $t$  is then

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

where  $\eta_i$  is a linear combination of the explanatory variables that are not time-dependent, whose values have been recorded at the time origin for the  $i$ th individual, and  $x_{1i}(t)$  is the value of  $X_1$  for that individual at time  $t$ .

Under this model, the hazard function is  $\exp(\eta_i)h_0(t)$  for patients who do not receive a transplant before time  $t$ , and  $\exp\{\eta_i + \beta_1\}h_0(t)$  thereafter. The effect of transplant on the patient's survival experience is then reflected in  $\beta_1$ . In particular, for two patients who have the same values of other explanatory variables in a model,  $e^{\beta_1}$  is the hazard of death at time  $t$  for the patient who receives a transplant before time  $t$ , relative to the hazard at that time for the patient who does not. Values of  $-2 \log \hat{L}$  can be compared after fitting the models with and without the time-dependent variable  $X_1$ ; a significant difference in these values means that the transplant has an effect on survival.

In a refinement to this model, Cox and Oakes (1984) suggested that the term  $\beta_1 x_{1i}(t)$  be replaced by  $\beta_1 + \beta_2 \exp\{-\beta_3(t - t_1)\}$  for patients receiving a transplant at time  $t_1$ . In this model, the effect of the transplant is to increase the hazard to some value  $\exp(\eta_i + \beta_1 + \beta_2)h_0(t)$  immediately after the transplant, when  $t = t_1$ , and to then decrease exponentially to  $\exp(\eta_i + \beta_1)h_0(t)$ , which is less than the initial hazard  $\exp(\eta_i)h_0(t)$  if  $\beta_1 < 0$ . See Figure 8.4, which shows graphically the behaviour of the hazard ratio,  $h_i(t)/h_0(t)$ , for a transplant patient for whom  $\eta_i$  is the linear component of the model. Although this is an attractive model, it does have the disadvantage that specialist software is required to fit it.



**Figure 8.4** The hazard ratio  $\exp\{\eta_i + \beta_1 + \beta_2 e^{-\beta_3(t-t_1)}\}$ ,  $t > t_1$ , for individual  $i$  who receives a transplant at  $t_1$ .

In situations where a particular explanatory variable is changing rapidly, new variables that reflect such changes may be defined. The dependence of the hazard on the values of such variables can then be explored. For example, in an oncological study, the percentage increase in the size of a tumour over a period of time might be a more suitable prognostic variable than either the size of the tumour at the time origin, or the time-dependent version of that variable. If this route is followed, the computational burden of fitting time-dependent variables can be avoided.

### 8.5 Three examples

In this section, three examples of survival analyses that involve time-dependent variables are given. In the first, data from a study concerning the use of bone marrow transplantation in the treatment of leukaemia is used to illustrate how a variable that is associated with the state of a patient, and whose value changes over the follow-up period, can be included in a model. In the second example, data from Example 5.10 of Chapter 5 on the comparison of two chemotherapy treatments for ovarian cancer are analysed to explore whether there is an interaction between age and survival time. The third example is designed to illustrate how information on a time-varying explanatory variate recorded during the follow-up period can be incorporated in a model for survival times. Studies in which the values of certain explanatory variables are recorded regularly throughout the follow-up period of each patient generate large sets of data. For this reason, artificial data from a small number of individuals will be used in Example 8.3 to illustrate the methodology.

#### *Example 8.1 Bone marrow transplantation in the treatment of leukaemia*

Patients suffering from acute forms of leukaemia often receive a bone marrow transplant. This provides the recipient with a new set of parent blood-forming cells, known as stem cells, which in turn produce a supply of healthy red and white blood cells. Klein and Moeschberger (1997) describe a multicentre study of factors that affect the prognosis for leukaemia patients treated in this manner. This study involved patients suffering from acute lymphoblastic leukaemia (ALL) and acute myelocytic leukaemia (AML), with those suffering from AML being further divided into low-risk and high-risk, according to their status at the time of transplantation. The survival time from the date of the transfusion is available for each patient, together with the values of a number of explanatory variables concerning the characteristics of the donor and recipient, and adverse events that occurred during the recovery process. Before the bone marrow transplant, patients were treated with a combination of cyclophosphamide and busulfan, in order to destroy all abnormal blood cells. The time taken for the blood platelets to return to a normal level is then an important variable in terms of the prognosis for a patient, and so the values of this variable were also recorded.

This example is based on the data from just one hospital, St. Vincent in Sydney, Australia. The observed survival time of each patient was recorded in

days, together with the values of an event indicator which is unity if a patient died, and zero if the patient was still alive at the end of the study period. The prognostic variables to be used in this example concern the disease group, the ages of the patient and the bone marrow donor, an indicator variable that denotes whether the platelet count returned to a normal level of  $40 \times 10^9$  per litre, and the time taken to return to this value. Two patients, numbers 7 and 21, died before the platelet count had returned to normal, and so for these patients, no value is given for the time to return to a normal platelet count. The variables recorded are therefore as follows:

- Time*: Survival time in days,
- Status*: Event indicator ( $0 =$  censored,  $1 =$  event),
- Group*: Disease group ( $1 =$  ALL,  $2 =$  low-risk AML,  
 $3 =$  high-risk AML),
- Page*: Age of patient,
- Dage*: Age of donor,
- P*: Platelet recovery indicator ( $0 =$  no,  $1 =$  yes),
- Ptime*: Time in days to return of platelets to normal level (if  $P = 1$ ).

The data base used in this example is given in Table 8.1.

The aim of the analysis of these data is to examine whether there are any differences between the survival times of patients in the three disease groups, after adjusting for prognostic variables. In order to investigate the effect of the time taken for platelets to return to their normal level on patient survival, a time-dependent variable,  $\text{Plate}(t)$ , is defined. This variable takes the value zero at times  $t$  when the platelets have not returned to normal levels, and then switches to unity once a normal level has been achieved. Formally,

$$\text{Plate}(t) = \begin{cases} 0 & \text{if } t < \text{time at which platelets returned to normal}, \\ 1 & \text{if } t \geq \text{time at which platelets returned to normal}, \end{cases}$$

so that  $\text{Plate}(t) = 0$  for all  $t$  when a patient dies before platelet recovery.

We first fit a Cox proportional hazards model that contains the variables associated with the age of the patient and donor, *Page* and *Dage*. When either of these variables is added on their own or in the presence of the other, there is no significant reduction in the value of the  $-2 \log \hat{L}$  statistic.

The time-dependent variable  $\text{Plate}(t)$  is now added to the null model. The value of  $-2 \log \hat{L}$  is reduced from 67.13 to 62.21, a reduction of 4.92 on 1 d.f., which is significant at the 5% level ( $P = 0.026$ ). This suggests that time to platelet recovery does affect survival. After allowing for the effects of this variable, there is still no evidence that the hazard of death is dependent on the age of the patient or donor.

The estimated coefficient of  $\text{Plate}(t)$  in the fitted model is  $-2.696$ , and the fact that this is negative indicates that there is a greater hazard of death at any given time for a patient whose platelets are not at a normal level. The hazard ratio at any given time is  $\exp(-2.696) = 0.067$ , and so a patient whose platelets have recovered to normal at a given time has about one fifteenth

**Table 8.1** Survival times of patients following bone marrow transplantation.

Patient	Time	Status	Group	Page	Dage	P	Ptime
1	1199	0	1	24	40	1	29
2	1111	0	1	19	28	1	22
3	530	0	1	17	28	1	34
4	1279	1	1	17	20	1	22
5	110	1	1	28	25	1	49
6	243	1	1	37	38	1	23
7	86	1	1	17	26	0	
8	466	1	1	15	18	1	100
9	262	1	1	29	32	1	59
10	1850	0	2	37	36	1	9
11	1843	0	2	34	32	1	19
12	1535	0	2	35	32	1	21
13	1447	0	2	33	28	1	24
14	1384	0	2	21	18	1	19
15	222	1	2	28	30	1	52
16	1356	0	2	33	22	1	14
17	1136	0	3	47	27	1	15
18	845	0	3	40	39	1	20
19	392	1	3	43	50	1	24
20	63	1	3	44	37	1	16
21	97	1	3	48	56	0	
22	153	1	3	31	25	1	59
23	363	1	3	52	48	1	19

the risk of death at that time. However, a 95% confidence interval for the corresponding true hazard ratio is (0.006, 0.751), which shows that the point estimate of the relative risk is really quite imprecise.

To quantify the effect of disease group on survival, the change in  $-2 \log \hat{L}$  when the two terms corresponding to the factor *Group* are added to the model that contains the time-dependent variable *Plate(t)* is 6.49 on 2 d.f., which is significant at the 5% level ( $P = 0.039$ ). The parameter estimates associated with disease group show that the hazard of death is much greater for those suffering from ALL and those in the high-risk group of AML sufferers. The hazard ratios for an ALL patient relative to a low-risk AML patient is 7.97 and that for a high-risk AML patient relative to a low-risk one is 11.77.

For the model that contains the factor *Group* and the time-dependent variable *Plate(t)*, the estimated baseline cumulative hazard and survivor functions are given in Table 8.2. These have been obtained using the estimate of the baseline cumulative hazard function given in equation (8.3).

In this table,  $\tilde{H}_0(t)$  and  $\tilde{S}_0(t)$  are the estimated cumulative hazard and survivor functions for an individual with AML and for whom the platelet recovery indicator, *Plate(t)*, remains at zero throughout the study. Also given in this

**Table 8.2** Estimated baseline cumulative hazard,  $\tilde{H}_0(t)$ , baseline survivor function,  $\tilde{S}_0(t)$ , and survivor function for an ALL patient with  $Plate(t) = 1$ ,  $\tilde{S}_1(t)$ .

Time, $t$	$\tilde{H}_0(t)$	$\tilde{S}_0(t)$	$\tilde{S}_1(t)$
0	0.0000	1.0000	1.0000
63	0.1953	0.8226	0.9810
86	0.3962	0.6728	0.9618
97	0.6477	0.5232	0.9383
110	1.2733	0.2799	0.8823
153	1.9399	0.1437	0.8264
222	2.6779	0.0687	0.7685
243	3.4226	0.0326	0.7143
262	4.2262	0.0146	0.6600
363	5.0987	0.0061	0.6057
392	6.0978	0.0022	0.5491
466	7.2663	0.0007	0.4895
1279	13.0700	0.0000	0.2766

table are the values of the estimated survivor function for an individual with ALL, but for whom *Plate(t) = 1* for all values of  $t$ , denoted  $\tilde{S}_1(t)$ . Since the value of *Plate(t)* is zero for each patient at the start of the study, and for most patients this changes to unity at some later point in time, these two estimated survivor functions illustrate the effect of platelet recovery at any specific time. For example, the probability of an ALL patient surviving beyond 97 days is only 0.52 if their platelets have not recovered to a normal level by this time. On the other hand, if such a patient has experienced platelet recovery by this time, they would have an estimated survival probability of 0.94. The estimated survivor function for an ALL patient whose platelet recovery status changes at some time  $t_0$  from 0 to 1 can also be obtained from Table 8.2, since this will be  $\tilde{S}_0(t)$  for  $t \leq t_0$  and  $\tilde{S}_1(t)$  for  $t > t_0$ . Estimates of the survivor function may also be obtained for individuals in the other disease groups.

In this illustration, the data from two patients who died before their platelet count had reached a normal level have a substantial impact on inferences about the effect of platelet recovery. If patients 7 and 21 are omitted from the data base, the time-dependent variable is no longer significant when added to the null model ( $P = 0.755$ ). The conclusion about the effect of platelet recovery time on survival is therefore dramatically influenced by the data for these two patients.

#### Example 8.2 Chemotherapy in ovarian cancer patients

When a Cox proportional hazards model that contains the variables *Age*, the age of a patient at the time origin, and *Treat*, the treatment group, is fitted to the data on the survival times of patients with ovarian cancer, the estimated

hazard function for the  $i$ th of 26 patients in the study is

$$\hat{h}_i(t) = \exp\{0.147 \text{Age}_i - 0.796 \text{Treat}_i\} h_0(t).$$

The value of the statistic  $-2 \log \hat{L}$  for this model is 54.148.

We now fit a model that contains *Age* and *Treat*, and a term corresponding to an interaction between age and survival time. This interaction will be modelled by including the time-dependent variable *Tage*, whose values are formed from the product of *Age* and the survival time  $t$ , that is,  $\text{Tage} = \text{Age} \times t$ . Since the values of *Tage* are dependent upon  $t$ , this time-dependent variable cannot be fitted in the same manner as *Age* and *Treat*. When *Tage* is added to the model, the fitted hazard function becomes

$$\hat{h}_i(t) = \exp\{0.216 \text{Age}_i - 0.664 \text{Treat}_i - 0.0002 \text{Age}_i t\} h_0(t).$$

Under this model, the hazard of death at  $t$  for a patient of a given age on the combined treatment (*Treat* = 2), relative to one of the same age on the single treatment (*Treat* = 1), is  $\exp(-0.664) = 0.52$ , which is not very different from the value of 0.45 found using the model that does not contain the variable *Tage*. However, the log-hazard ratio for a patient aged  $a_2$  years, relative to one aged  $a_1$  years, is

$$0.216(a_2 - a_1) - 0.0002(a_2 - a_1)t$$

at time  $t$ . This model therefore allows the log-hazard ratio for *Age* to be linearly dependent on survival time.

The value of  $-2 \log \hat{L}$  for the model that contains *Age*, *Treat* and *Tage* is 53.613. The change in  $-2 \log \hat{L}$  on adding the variable *Tage* to a model that contains *Age* and *Treat* is therefore 0.53, which is not significant ( $P = 0.465$ ). We therefore conclude that the time-dependent variable *Tage* is not in fact needed in the model.

#### *Example 8.3 Data from a hypothetical cirrhosis study*

Although the data to be used in this example are artificial, it is useful to provide a background against which these data can be considered. Suppose therefore that 12 patients have been recruited to a study on the treatment of cirrhosis of the liver. The patients are randomised to receive either a placebo or a new treatment that will be referred to as Liverol. Six patients are allocated to Liverol and six to the placebo. At the time when the patient is entered into the study, the age and baseline value of the patient's bilirubin level are recorded. The natural logarithm of the bilirubin value (in  $\mu\text{mol/l}$ ) will be used in this analysis. The variables measured are summarised below:

- Time*: Survival time of the patient in days,
- Status*: Event indicator (0 = censored, 1 = uncensored),
- Treat*: Treatment group (0 = placebo, 1 = Liverol),
- Age*: Age of the patient in years,
- Lbr*: Logarithm of bilirubin level.

The values of these variables are given in Table 8.3.

**Table 8.3** Survival times of 12 patients in a study on cirrhosis of the liver.

Patient	Time	Status	Treat	Age	Lbr
1	281	1	0	46	3.2
2	604	0	0	57	3.1
3	457	1	0	56	2.2
4	384	1	0	65	3.9
5	341	0	0	73	2.8
6	842	1	0	64	2.4
7	1514	1	1	69	2.4
8	182	0	1	62	2.4
9	1121	1	1	71	2.5
10	1411	0	1	69	2.3
11	814	1	1	77	3.8
12	1071	1	1	58	3.1

Patients are supposed to return to the clinic three, six and twelve months after the commencement of treatment, and yearly thereafter. On these occasions, the bilirubin level is again measured and recorded. Data are therefore available on how the bilirubin level changes in each patient throughout the duration of the study. Table 8.4 gives the values of the logarithm of the bilirubin value at each time in the follow-up period for each patient.

In taking  $\log(\text{bilirubin})$  to be a time-dependent variable, the value of the variate at any time  $t$  is that recorded at the last follow-up visit before  $t$ , for each patient. In this calculation, the change to a new value will be assumed to take place immediately after the time that the reading was taken, so that patient 1, for example, is assumed to have a  $\log(\text{bilirubin})$  value of 3.2 from the time origin until day 47, after which it increases to 3.8, and so on. This is equivalent to the assumption that the values of *Lbr* for a given individual follow a step-function in which the values are assumed constant between any two adjacent time points.

The data are first analysed using the baseline  $\log(\text{bilirubin})$  value alone. A Cox proportional hazards model is used, and the values of  $-2 \log \hat{L}$  on fitting particular models are as shown in Table 8.5.

Both *Age* and *Lbr* appear to be needed in the model, although the evidence for including *Age* as well as *Lbr* is not very strong. When *Treat* is added to the model that contains *Age* and *Lbr*, the reduction in the value of  $-2 \log \hat{L}$  is 5.182 on 1 d.f. This is significant at the 5% level ( $P = 0.023$ ). The coefficient of *Treat* is  $-3.052$ , indicating that the drug Liverol is effective in reducing the hazard of death. Indeed, other things being equal, Liverol reduces the hazard of death by a factor of 0.047.

We now analyse these data, taking the  $\log(\text{bilirubin})$  values to be time-dependent. Let *Lbrt* be the time-dependent variate formed from the values of

**Table 8.4** Follow-up times and log(bilirubin) values for the 12 patients in the cirrhosis study.

Patient	Follow-up time	Log(bilirubin)
1	47	3.8
	184	4.9
	251	5.0
2	94	2.9
	187	3.1
	321	3.2
3	61	2.8
	97	2.9
	142	3.2
	359	3.4
	440	3.8
4	92	4.7
	194	4.9
	372	5.4
5	87	2.6
	192	2.9
	341	3.4
6	94	2.3
	197	2.8
	384	3.5
	795	3.9
	74	2.9
7	202	3.0
	346	3.0
	917	3.9
	1411	5.1
	90	2.5
8	182	2.9
	101	2.5
	410	2.7
9	774	2.8
	1043	3.4
	182	2.2
	847	2.8
	1051	3.3
10	1347	4.9
	167	3.9
	498	4.3
11	108	2.8
	187	3.4
	362	3.9
	694	3.8

**Table 8.5** Values of  $-2 \log \hat{L}$  for models without a time-dependent variable.

Terms in model	$-2 \log \hat{L}$
null model	25.121
<i>Age</i>	22.135
<i>Lbr</i>	21.662
<i>Age, Lbr</i>	18.475

log(bilirubin). The values of  $-2 \log \hat{L}$  on fitting Cox regression models to the data are then given in Table 8.6.

**Table 8.6** Values of  $-2 \log \hat{L}$  for models with a time-dependent variable.

Terms in model	$-2 \log \hat{L}$
null model	25.121
<i>Age</i>	22.135
<i>Lbrt</i>	12.050
<i>Age, Lbrt</i>	11.145

It is clear from this table that the hazard function depends on the time-dependent variable *Lbrt*, and that after allowing for this, the effect of *Age* is slight. We therefore add the treatment effect *Treat* to the model that contains *Lbrt* alone. The effect of this is that  $-2 \log \hat{L}$  is reduced from 12.050 to 10.676, a reduction of 1.374 on 1 d.f. This reduction is not significant ( $P = 0.241$ ) leading to the conclusion that after taking account of the dependence of the hazard of death on the evolution of the log(bilirubin) values, no treatment effect is discernible.

The estimated hazard function for the  $i$ th individual is given by

$$\hat{h}_i(t) = \exp\{3.605 Lbr_i(t) - 1.479 Treat_i\} h_0(t),$$

where  $Lbr_i(t)$  is the value of log(bilirubin) for the  $i$ th patient at time  $t$ . The estimated ratio of the hazard of death at time  $t$  for two individuals on the same treatment who have values of *Lbr* that differ by one unit at  $t$  is  $e^{3.605} = 36.77$ .

One possible explanation for the difference between the results of these two analyses is that the effect of the treatment is to change the values of the bilirubin level, so that after changes in these values over time have been allowed for, no treatment effect is visible.

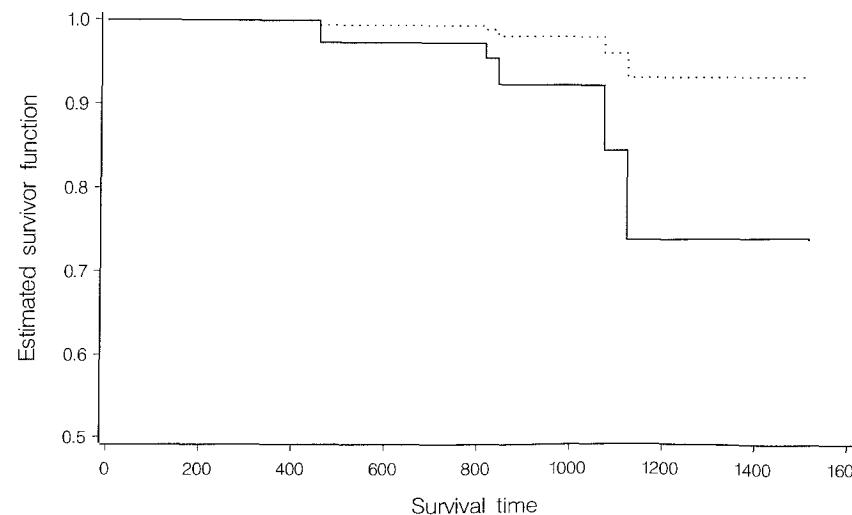
The baseline cumulative hazard function may now be estimated for the

model that contains the time-dependent variable  $Lbrt$  and  $Treat$ . The estimated values of this function are tabulated in Table 8.7.

**Table 8.7** Estimated baseline cumulative hazard function,  $\tilde{H}_0(t)$ , for the cirrhosis study.

Follow-up time ( $t$ )	$\tilde{H}_0(t)$
0	0.000
281	$0.009 \times 10^{-6}$
384	$0.012 \times 10^{-6}$
457	$0.541 \times 10^{-6}$
814	$0.908 \times 10^{-6}$
842	$1.577 \times 10^{-6}$
1071	$3.318 \times 10^{-6}$
1121	$6.007 \times 10^{-6}$
1514	$6.053 \times 10^{-6}$

This table shows that the cumulative hazard function is increasing in a non-linear fashion, which indicates that the baseline hazard is not constant, but increases with time. The corresponding baseline survivor function could be obtained from this estimate. However, a model with  $Lbrt = 0$  for all  $t$  is not at all easy to interpret, and so the estimated survivor function is obtained for an individual for whom  $Lbrt = 3$ . This function is shown in Figure 8.5, for a patient in either treatment group.



**Figure 8.5** Estimated survivor function for a patient with  $Lbr = 3$ , for all  $t$ , who is on placebo (—) or Liverol (···).

This figure clearly shows that patients on placebo have a much poorer prognosis than those on Liverol.

Finally, we illustrate how estimates of the conditional probabilities of survival from time  $t$  to time  $t + 360$  days can be obtained, using the result given in equation (8.5). Using the values of the time-dependent variable given in Tables 8.3 and 8.4, the values of  $\sum_{j=1}^p \hat{\beta}_j x_{ji}(t)$ , the prognostic index at time  $t$  for the  $i$ th patient,  $i = 1, 2, \dots, 12$ , can be calculated for  $t = 0, 360, 720, 1080$ , and 1440. For this calculation, the  $\log(\text{bilirubin})$  value at one of these times is taken to be the value recorded at the immediately preceding follow-up time. Table 8.7 is then used to obtain the values of  $\tilde{H}_0(t + 360) - \tilde{H}_0(t)$ , and then  $\tilde{P}_i(t, t + 360)$  is obtained from equation (8.5). The full set of results will not be given here, but as an example, the estimated approximate conditional probabilities of surviving through consecutive intervals of 360 days, for patients 1 and 7, are shown in Table 8.8.

**Table 8.8** Approximate conditional survival probabilities for patients 1 and 7.

Time interval	$\tilde{P}_1(t, t + h)$	$\tilde{P}_7(t, t + h)$
0–	0.999	1.000
360–	0.000	0.994
720–	0.000	0.969
1080–	0.000	0.457
1440–	0.045	0.364

Note that because these estimates are survival probabilities conditional on being alive at the start of an interval, they are not necessarily monotonic. These estimates again show that patients on Liverol have a greater probability of surviving for a period of one year, if they are alive at the start of that year, than patients on the placebo. Finally, the values of  $1 - \tilde{P}_i(t, t + h)$  are approximate estimates of the probability of death within each interval, conditional on a patient being alive at the start. Summing these estimates over all 12 patients leads to the values 0.02, 2.46, 5.64, 6.53, 3.16, for each of the five intervals, respectively. These can be compared to the observed numbers of deaths in each interval, which are 1, 2, 3, 1 and 1, respectively. There is therefore a tendency for the model to overestimate the numbers of deaths, but because of the small size of the data set, this does not provide a very reliable assessment of the predictive power of the model.

## 8.6 Further reading

The possibility of incorporating time-dependent variables in a proportional hazards model was raised by Cox (1972). The appropriate partial likelihood function was given in his paper, and discussed in greater detail in Cox (1975). Kalbfleisch and Prentice (2002) include a detailed account of the construc-

tion of the partial likelihood function. The classification of time-dependent variables outlined in Section 8.1 is due to Prentice and Kalbfleisch (1979), who amplify on this in Kalbfleisch and Prentice (2002). Andersen (1992) reviews the uses of time-dependent variables in survival analysis and includes an example on which the hypothetical study of Example 8.3 is loosely based.

A number of practical problems encountered in the analysis of survival data with time-dependent variables are discussed by Altman and De Stavola (1994), who include a review of software available at that time. A comprehensive analysis of data on primary biliary cirrhosis, which includes an assessment of conditional survival probabilities is also provided. See Christensen (1986) for a further illustration. Klein and Moeschberger (1997) give the full data set on survival following bone marrow transplantation, part of which was used in Example 8.1, and use this in a number of detailed illustrative examples.

The model described in Section 8.4 in connection with organ transplantation was presented by Crowley and Hu (1977) and Cox and Oakes (1984) in an analysis of the ‘Stanford heart transplant data’. This famous data set is given in Crowley and Hu (1977) and an update is provided by Cox and Oakes (1984). See also Aitkin, Laird and Francis (1983) and the ensuing discussion.

Relatively little work has been done on incorporating time-dependent variables in a fully parametric model for survival data. However, Petersen (1986) shows how a parametric model with time-dependent variables can be fitted.

## CHAPTER 9

# Interval-censored survival data

In many studies where the response variable is a survival time, the exact time of the event of interest will not be known. Instead, the event will be known to have occurred during a particular interval of time. Data in this form are known as *grouped* or *interval-censored* survival data.

Interval-censored data commonly arise in studies where there is a non-lethal end-point, such as the recurrence of a disease or condition. However, most survival analyses are based on interval-censored data, in the sense that the survival times are often taken as the nearest day, week or month. In this chapter, some methods for analysing interval-censored data will be described and illustrated. Models in which specific assumptions are made about the form of the underlying hazard function are considered in Sections 9.1 to 9.4, and fully parametric models are discussed in Section 9.5.

### 9.1 Modelling interval-censored survival data

In this chapter, a number of methods for the analysis of interval-censored survival data will be discussed in the context of a study on disease recurrence. In the management of patients who have been cured of ulcers, carcinomas or other recurrent conditions, the patients are usually provided with medication to maintain their recovery. These patients are subsequently examined at regular intervals in order to detect whether a recurrence has occurred. Naturally, some patients may experience symptoms of a recurrence and be subsequently diagnosed as having had a recurrence at a time other than one of the scheduled screening times.

Now suppose that the study is designed to compare two maintenance therapies, a new and a standard treatment, say, and that a number of explanatory variables are recorded for each individual when they are recruited to the study. The vector  $\mathbf{x}_i$  will be used to denote the set of values of  $p$  explanatory variables,  $X_1, X_2, \dots, X_p$ , for the  $i$ th individual in the study. The first of these variables,  $X_1$ , will be taken to be an indicator variable corresponding to the treatment group, where  $X_1 = 0$  if an individual is on the standard treatment and  $X_1 = 1$  if on the new treatment.

Clearly, one way of analysing such data is to ignore the interval censoring. A survival analysis is then carried out on the times of a detected recurrence. However, the data set used in this analysis will be based on a mixture of recurrences detected at scheduled screening times, known as *screen-detected recurrences* and recurrences diagnosed following the occurrence of symptoms

or *interval-detected recurrences*. This leads to a difficulty in interpreting the results of the analysis.

To illustrate the problem, consider a study to compare two treatments for suppressing the recurrence of an ulcer, a new and a standard treatment, say. Also suppose that both treatments have exactly the same effect on the recurrence time, but that the new treatment suppresses symptoms. The recurrence of an ulcer in a patient on the new treatment will then tend to be detected later than that in a patient on the standard treatment. Therefore, interval-detected recurrences will be identified sooner in a patient on the standard treatment. The interval-detected recurrence times will then be shorter for this group of patients, indicating an apparent advantage of the new treatment over the standard.

If the time interval between successive screenings is short, relative to the average time to recurrence, there will be few interval-detected recurrences. The standard application of methods for survival analysis will not then be inappropriate.

#### *Example 9.1 Recurrence of an ulcer*

In a double blind clinical trial to compare treatments for the inhibition of relapse after primary therapy has healed an ulcer, patients are randomised to receive one or other of two treatments, labelled *A* and *B*. Regular visits to a clinic were arranged for the patients, and endoscopies were performed 6 months and 12 months after randomisation. A positive endoscopy result indicates that an ulcer has recurred in the time since the last negative result. Information is therefore obtained on whether or not an ulcer has recurred in the interval from 0 to 6 months or in the interval from 6 to 12 months. Additionally, some patients presented at the clinic between scheduled visits, suffering from symptoms of recurrence. These patients had endoscopies at these times in order to detect if a recurrence had in fact occurred.

At entry to the trial, the age of each person, in years, and the duration of verified disease (1 = less than five years, 2 = greater than or equal to five years) was recorded, in addition to the treatment group (*A* or *B*). There are two variables associated with ulcer detection in the data set, namely the time of the last visit, in months, and the result of the endoscopy (1 = no ulcer detected, 2 = ulcer detected). Those with times other than 6 or 12 months had presented with symptoms between scheduled visits.

The study itself was multinational and the full set of data is given in Whitehead (1989). In this example, only the data from Belgium will be used, and the relevant data are given in Table 9.1.

Once an ulcer is detected by endoscopy, a patient is treated for this and is then no longer in the study. There were some patients who either did not have an endoscopy six months after trial entry, or who dropped out after a negative unscheduled endoscopy in the first six months. These patients have been omitted from the data set on the grounds that there is no information about whether an ulcer has recurred in the first six months of the study. This

**Table 9.1** Data on the recurrence of an ulcer following treatment for the primary disease.

Patient	Age	Duration	Treatment	Time of last visit	Result
1	48	2	<i>B</i>	7	2
2	73	1	<i>B</i>	12	1
3	54	1	<i>B</i>	12	1
4	58	2	<i>B</i>	12	1
5	56	1	<i>A</i>	12	1
6	49	2	<i>A</i>	12	1
7	71	1	<i>B</i>	12	1
8	41	1	<i>A</i>	12	1
9	23	1	<i>B</i>	12	1
10	37	1	<i>B</i>	5	2
11	38	1	<i>B</i>	12	1
12	76	2	<i>B</i>	12	1
13	38	2	<i>A</i>	12	1
14	27	1	<i>A</i>	6	2
15	47	1	<i>B</i>	6	2
16	54	1	<i>A</i>	6	1
17	38	1	<i>B</i>	10	2
18	27	2	<i>B</i>	7	2
19	58	2	<i>A</i>	12	1
20	75	1	<i>B</i>	12	1
21	25	1	<i>A</i>	12	1
22	58	1	<i>A</i>	12	1
23	63	1	<i>B</i>	12	1
24	41	1	<i>A</i>	12	1
25	47	1	<i>B</i>	12	1
26	58	1	<i>A</i>	3	2
27	74	2	<i>A</i>	2	2
28	75	2	<i>A</i>	6	1
29	72	1	<i>A</i>	12	1
30	59	1	<i>B</i>	12	2
31	52	1	<i>B</i>	12	1
32	75	1	<i>B</i>	12	2
33	76	1	<i>A</i>	12	1
34	34	2	<i>A</i>	6	1
35	36	1	<i>B</i>	12	1
36	59	1	<i>B</i>	12	1
37	44	1	<i>A</i>	12	2
38	28	2	<i>B</i>	12	1
39	62	1	<i>B</i>	12	1
40	23	1	<i>A</i>	12	1
41	49	1	<i>B</i>	12	1
42	61	1	<i>A</i>	12	1
43	33	2	<i>B</i>	12	1

means that those patients in Table 9.1 whose last visit was greater than 6 months after randomisation would have had a negative endoscopy at 6 months.

In modelling the data from this study, duration of disease is denoted by an indicator variable *Dur*, which is zero when the duration is less than 5 years and unity otherwise. The treatment effect is denoted by a variable *Treat*, which takes the value zero if an individual is on treatment *A* and unity if on treatment *B*. The patients age is reflected in the continuous variate *Age*.

We first analyse the recurrence times in Table 9.1 ignoring the interval censoring. The recurrence times of those patients who have not had a detected recurrence by the time of their last visit are taken to be censored. The data are now analysed using the Cox proportional hazards model, described in Chapter 3.

**Table 9.2** Values of  $-2 \log \hat{L}$  on fitting a Cox regression model to data on the time to recurrence of an ulcer.

Variables in model	$-2 \log \hat{L}$
None	79.189
<i>Dur</i>	79.157
<i>Age</i>	78.885
<i>Age + Dur</i>	78.872
<i>Age + Dur + Treat</i>	78.747
<i>Treat</i>	79.097

From the values of the  $-2 \log \hat{L}$  statistic for different models, given in Table 9.2, it is clear that neither age nor duration of disease are important prognostic factors. Moreover, the reduction in  $-2 \log \hat{L}$  on adding the treatment effect to the model, adjusted or unadjusted for the prognostic factors, is nowhere near significant.

The estimated coefficient of *Treat* in the model that contains *Treat* alone is 0.189, and the standard error of this estimate is 0.627. The estimated hazard of a recurrence under treatment *B* (*Treat* = 1), relative to that under treatment *A* (*Treat* = 0), is therefore  $\exp(0.189) = 1.21$ . The standard error of the estimated hazard ratio is found using equation (3.12) in Chapter 3, and is 0.758. The fact that the estimated hazard ratio is greater than unity gives a slight indication that treatment *A* is superior to treatment *B*, but not significantly so.

## 9.2 Modelling the recurrence probability in the follow-up period

Suppose that patients are followed up to time  $t_s$ , at which time the last scheduled screening test is carried out. Information on whether or not a recurrence was detected at any time up to and including the last screen is then recorded.

Let  $p_i(t_s)$  be the probability of a recurrence up to time  $t_s$  for the  $i$ th patient,  $i = 1, 2, \dots, n$ , with explanatory variables  $\mathbf{x}_i$ . We now adopt a Cox proportional hazards model, according to which the hazard of a recurrence at  $t_s$ , for the  $i$ th patient, is given by

$$h_i(t_s) = \exp(\eta_i) h_0(t_s),$$

where  $\eta_i = \beta' \mathbf{x}_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}$ , and  $h_0(t_s)$  is the baseline hazard function at  $t_s$ .

The probability that the  $i$ th individual experiences a recurrence after time  $t_s$  is the survivor function  $S_i(t_s)$ , so that  $S_i(t_s) = 1 - p_i(t_s)$ . Now, from equation (3.22) in Section 3.8 of Chapter 3,

$$S_i(t_s) = \{S_0(t_s)\}^{\exp(\eta_i)}, \quad (9.1)$$

where  $S_0(t_s)$  is the value of the survivor function at  $t_s$  for an individual on the standard treatment for whom all the other explanatory variables are zero. The probability of a recurrence up to time  $t_s$  under this model is therefore

$$p_i(t_s) = 1 - \{S_0(t_s)\}^{\exp(\eta_i)},$$

and so

$$\log[-\log\{1 - p_i(t_s)\}] = \eta_i + \log\{-\log S_0(t_s)\}.$$

Writing  $\beta_0 = \log\{-\log S_0(t_s)\}$ , the model can be expressed as

$$\log[-\log\{1 - p_i(t_s)\}] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}. \quad (9.2)$$

This is a linear model for the complementary log-log transformation of the probability of a recurrence up to time  $t_s$ . The model can be fitted to data on a binary response variable that takes the value zero for those individuals in the study who have not experienced a recurrence before  $t_s$ , the time of the last screen, and unity otherwise.

As in modelling survival data, models fitted to data on a binary response variable can be compared on the basis of the statistic  $-2 \log \hat{L}$ . Here,  $\hat{L}$  is the maximised likelihood of the binary data, and  $-2 \log \hat{L}$  is generally known as the *deviance*. Differences in the deviance for two nested models have an asymptotic chi-squared distribution, and so models fitted to binary data can be compared in the same manner as the models used in survival analysis.

When the model in equation (9.2) is fitted to the observed data, the estimate of the constant,  $\hat{\beta}_0$ , is an estimate of  $\log\{-\log S_0(t_s)\}$ , from which an estimate of the baseline survivor function at  $t_s$  can be obtained. Also, the ratio of the hazard of a recurrence for an individual on the new treatment, relative to one on the standard, is  $\exp(\hat{\beta}_1)$ . This can be estimated by  $\exp(\hat{\beta}_1)$ , where  $\hat{\beta}_1$  is the parameter estimate corresponding to  $X_1$ , the indicator variable that corresponds to the treatment group. Values of the hazard ratio less than unity suggest that the risk of a recurrence at any time is smaller under the new treatment than under the standard. A confidence interval for the hazard ratio may be obtained from the standard error of  $\hat{\beta}_1$  in the usual manner.

This method of estimating the hazard ratio from interval-censored survival

data is not particularly efficient, since data on the times that a recurrence is detected are not utilised. However, the method is appropriate when interest simply centres on the risk of a recurrence in a specific time period. It is also the method that would be adopted in modelling quantities such as the probability of a relapse in the first year of treatment, or the probability of no recurrence in a five-year period after trial entry.

#### *Example 9.2 Recurrence of an ulcer*

We now model the probability of an ulcer recurring in the 12 months following recruitment to the study described in Example 9.1. Of the 43 patients in the data set, 11 of them had experienced a recurrence in this 12-month period, namely patients 1, 10, 14, 15, 17, 18, 26, 27, 30, 32 and 37. A binary response variable is now defined, which takes the value unity if a patient has experienced a recurrence and zero otherwise. A model in which the complementary log-log transformation of the recurrence probability is related to age, duration of disease and treatment group is then fitted to the binary observations.

Table 9.3 gives the deviances on fitting complementary log-log models with different terms to the binary response variable. All the models fitted include a constant term.

**Table 9.3** Deviances on fitting complementary log-log models to data on the recurrence of an ulcer in 12 months.

Variables in model	Deviance	d.f.
Constant	48.902	42
<i>Dur</i>	48.899	41
<i>Age</i>	48.573	41
<i>Treat</i>	48.531	41
<i>Dur + Age</i>	48.565	40
<i>Dur + Treat</i>	48.531	40
<i>Age + Treat</i>	48.175	40
<i>Dur + Age + Treat</i>	48.172	39
<i>Dur + Age + Treat + Treat × Age</i>	47.944	38
<i>Dur + Age + Treat + Treat × Dur</i>	48.062	38

In this example, the effects of age, duration of disease and treatment group have been modelled using the variates *Age*, *Dur*, and *Treat*, defined in Example 9.1. However, factors corresponding to duration and treatment could have been used in conjunction with packages that allow factors to be included directly. This would not make any difference to the deviances in Table 9.3, but it may have an effect on the interpretation of the parameter estimates. See Sections 3.2 and 3.7 for fuller details.

It is clear from Table 9.3 that no variable reduces the deviance by a significant amount. For example, the change in the deviance on adding *Treat* to the model that only contains a constant is 0.371, which is certainly not significant when compared to percentage points of the chi-squared distribution on 1 d.f.

Approximately the same change in deviance results when *Treat* is added to the model that contains *Age* and *Dur*, showing that the treatment effect is of a similar magnitude after allowing for these two variables. Moreover, there is no evidence whatsoever of an interaction between treatment and the variables *Age* and *Dur*.

On fitting a model that contains *Treat* alone, the estimated coefficient of *Treat* is 0.378, with a standard error of 0.629. Thus the ratio of the hazard of a recurrence before 12 months in a patient on treatment *B* (*Treat* = 1), relative to that for a patient on treatment *A* (*Treat* = 0), is  $\exp(0.378) = 1.46$ . The risk of a recurrence in the year following randomisation is thus greater under treatment *B* than it is under treatment *A*, but not significantly so. This hazard ratio is not too different from the value of 1.21 obtained in Example 9.1. The standard error of the estimated hazard ratio, again found using equation (3.12) in Chapter 3, is 0.918, which is also very similar to that found in Example 9.1.

A 95% confidence interval for the log-hazard ratio has limits of  $0.378 \pm 1.96 \times 0.629$ , and so the corresponding interval estimate for the hazard ratio itself is (0.43, 5.01). Notice that this interval includes unity, a result which was foreshadowed by the non-significant treatment effect.

The estimated constant term in this fitted model is  $-1.442$ . This is an estimate of  $\log\{-\log S_0(12)\}$ , the survivor function at 12 months for a patient on treatment *A*. The estimated probability of a recurrence after 12 months for a patient on treatment *A* is therefore  $\exp(-e^{-1.442}) = 0.79$ . The corresponding value for a patient on treatment *B* is  $0.79^{\exp(0.378)} = 0.71$ . The probabilities of a recurrence in the first 12 months are therefore 0.21 for a patient on treatment *A*, and 0.29 for a patient on treatment *B*. This again shows that patients on treatment *B* have a slightly higher probability of the recurrence of an ulcer in the year following randomisation.

#### 9.3\* Modelling the recurrence probability at different times

In this procedure for analysing interval-censored survival data, information about whether or not a recurrence is detected at different examination times is taken into account.

Suppose that patients enter a study at time 0 and are followed up to time  $t_k$ . During the course of this follow-up period, the individuals are screened on a regular basis in order to detect a recurrence of the disease or condition under study. Denote the examination times by  $t_1, t_2, \dots, t_k$ , which are such that  $t_1 < t_2 < \dots < t_k$ . Further, let  $t_0$  denote the time origin, so that  $t_0 = 0$  and let  $t_{k+1} = \infty$ .

For each individual, information will be recorded on whether or not a recurrence has occurred at times  $t_1, t_2, \dots, t_k$ . It can then be determined whether a given individual has experienced a recurrence in the  $j$ th time interval from  $t_{j-1}$  to  $t_j$ . Thus a patient who has a recurrence detected at time  $t_j$  has an actual recurrence time of  $t$ , where  $t_{j-1} \leq t < t_j$ ,  $j = 1, 2, \dots, k$ . Note that

the study will not provide any information about whether a recurrence occurs after the final screening time,  $t_k$ .

Now let  $p_{ij}$  be the probability of a recurrence being detected in the  $i$ th patient,  $i = 1, 2, \dots, n$ , at time  $t_j$ , so that  $p_{ij}$  is the probability that patient  $i$  experiences a recurrence in the  $j$ th time interval,  $j = 1, 2, \dots, k$ . Also let  $\pi_{ij}$  be the probability that the  $i$ th of  $n$  patients is found to be free of the disease at time  $t_{j-1}$  and has a recurrence in the  $j$ th time interval,  $j = 1, 2, \dots, k$ . This is therefore the conditional probability of a recurrence in the  $j$ th interval, given that the recurrence occurs after  $t_{j-1}$ . Using  $T_i$  to denote the random variable associated with the recurrence time of the  $i$ th individual, we therefore have

$$p_{ij} = P(t_{j-1} \leq T_i < t_j),$$

and

$$\pi_{ij} = P(t_{j-1} \leq T_i < t_j \mid T_i \geq t_{j-1}),$$

for  $j = 1, 2, \dots, k$ .

We now consider individuals who have not had a detected recurrence by the last examination time,  $t_k$ . For these individuals, we define  $T_i$  to be the random variable associated with the time to either a recurrence or death, and the corresponding probability of a recurrence or death in the interval from time  $t_k$  is given by

$$p_{i,k+1} = P(T_i \geq t_k) = 1 - \sum_{j=1}^k p_{ij}.$$

Also, the corresponding conditional probability of a recurrence or death in the interval  $(t_k, \infty)$  is

$$\pi_{i,k+1} = P(T_i \geq t_k \mid T_i \geq t_k) = 1.$$

It then follows that

$$p_{ij} = (1 - \pi_{i1})(1 - \pi_{i2}) \cdots (1 - \pi_{i,j-1})\pi_{ij}, \quad (9.3)$$

for  $j = 2, 3, \dots, k+1$ , with  $p_{i1} = \pi_{i1}$ .

Now let  $r_{ij}$  be unity if the  $i$ th patient has a recurrence detected in the interval from  $t_{j-1}$  to  $t_j$ ,  $j = 1, 2, \dots, k+1$ , and zero if no recurrence is detected in that interval, with  $r_{i,k+1} = 1$ . Also let  $s_{ij}$  be unity if a patient has a detected recurrence after  $t_j$ , and zero otherwise. Then,

$$s_{ij} = r_{i,j+1} + r_{i,j+2} + \cdots + r_{i,k+1},$$

for  $j = 1, 2, \dots, k$ .

The sample likelihood of the  $n(k+1)$  values  $r_{ij}$  is

$$\prod_{i=1}^n \prod_{j=1}^{k+1} p_{ij}^{r_{ij}},$$

and on substituting for  $p_{ij}$  from equation (9.3), the likelihood function be-

comes

$$\prod_{i=1}^n \prod_{j=1}^{k+1} \{(1 - \pi_{i1}) \cdots (1 - \pi_{i,j-1})\pi_{ij}\}^{r_{ij}}.$$

This function can be written as

$$\prod_{i=1}^n \pi_{i1}^{r_{i1}} \{(1 - \pi_{i1})\pi_{i2}\}^{r_{i2}} \cdots \{(1 - \pi_{i1}) \cdots (1 - \pi_{ik})\pi_{i,k+1}\}^{r_{i,k+1}},$$

which reduces to

$$\prod_{i=1}^n \pi_{i,k+1}^{r_{i,k+1}} \prod_{j=1}^k \pi_{ij}^{r_{ij}} (1 - \pi_{ij})^{s_{ij}}. \quad (9.4)$$

However,  $\pi_{i,k+1} = 1$ , and so the likelihood function in equation (9.4) becomes

$$\prod_{i=1}^n \prod_{j=1}^k \pi_{ij}^{r_{ij}} (1 - \pi_{ij})^{s_{ij}}. \quad (9.5)$$

This is the likelihood function for  $nk$  observations  $r_{ij}$  from a binomial distribution with response probability  $\pi_{ij}$ , and where the binomial denominator is  $r_{ij} + s_{ij}$ . This denominator is equal to unity when a patient is at risk of having a detected recurrence after time  $t_j$ , and zero otherwise. In fact, the denominator is zero when both  $r_{ij}$  and  $s_{ij}$  are equal to zero, and the likelihood function in expression (9.5) is unaffected by observations for which  $r_{ij} + s_{ij} = 0$ . Data records for which the binomial denominator is zero are therefore uninformative, and so they can be omitted from the data set. If there are  $m$  observations remaining after these deletions, so that  $m \leq nk$ , the likelihood function in expression (9.5) is that of  $m$  observations from binomial distributions with parameters 1 and  $\pi_{ij}$ , in other words,  $m$  observations from a *Bernoulli distribution*.

The next step is to note that for the  $i$ th patient,

$$1 - \pi_{ij} = P(T_i \geq t_j \mid T_i \geq t_{j-1}),$$

so that

$$1 - \pi_{ij} = \frac{S_i(t_j)}{S_i(t_{j-1})}.$$

Adopting a proportional hazards model for the recurrence times, the hazard of a recurrence being detected at time  $t_j$  in the  $i$ th individual can be expressed as

$$h_i(t_j) = \exp(\eta_i)h_0(t_j),$$

where  $h_0(t_j)$  is the baseline hazard at  $t_j$ , and  $\eta_i$  is the risk score for the  $i$ th individual. Notice that this assumption means that the hazards need only be proportional at the scheduled screening times  $t_j$ , and not at intermediate times. This is less restrictive than the usual proportional hazards assumption, which requires that hazards be proportional at every time.

Using the result in equation (9.1),

$$1 - \pi_{ij} = \left\{ \frac{S_0(t_j)}{S_0(t_{j-1})} \right\}^{\exp(\eta_i)},$$

and on taking logarithms we find that

$$\log(1 - \pi_{ij}) = \exp(\eta_i) \log \{S_0(t_j)/S_0(t_{j-1})\}.$$

Consequently,

$$\begin{aligned} \log\{-\log(1 - \pi_{ij})\} &= \eta_i + \log [-\log \{S_0(t_j)/S_0(t_{j-1})\}] \\ &= \eta_i + \gamma_j, \end{aligned}$$

say. This is a linear model for the complementary log-log transformation of  $\pi_{ij}$ , in which the parameters  $\gamma_j$ ,  $j = 1, 2, \dots, k$ , are associated with the  $k$  time intervals. The model can be fitted using standard methods for modelling binary data.

In modelling the probability of a recurrence in the  $j$ th time interval for the  $i$ th patient,  $\pi_{ij}$ , the data are the values  $r_{ij}$ . Data records for which both  $r_{ij}$  and  $s_{ij}$  are equal to zero are omitted, and so the binomial denominator is unity for each remaining observation. The parameters  $\gamma_j$  are incorporated in the model by fitting terms corresponding to a  $k$ -level factor associated with the period of observation, or by including suitable indicator variables as described in Section 3.2. Note that a constant term is not included in the model. The estimates of the  $\beta$ -coefficients in  $\eta_i$ , obtained on fitting this model, can again be interpreted as log-hazard ratios. Also, estimates of the  $\gamma_j$  can be used to obtain estimates of the  $\pi_{ij}$ . This process is illustrated in Example 9.3 below.

#### *Example 9.3 Recurrence of an ulcer*

The data on the time to detection of an ulcer recurrence, given in Example 9.1, are now analysed using the method described above. To prepare the data set for analysis using this approach, the two additional variables, *Period* and *R*, are introduced. The first of these, *Period*, is used to signify the period, and the variable is given the value unity for each observation. The second variable, *R*, contains the values  $r_{i1}$ ,  $i = 1, 2, \dots, 43$ , and so *R* is equal to unity if an ulcer is detected in the first period and zero otherwise. For these data, patients 10, 14, 15, 26 and 27 experienced a recurrence in the interval from 0 to 6 months, and so the value of *R* is unity for these five individuals and zero for the remaining 38.

We then add a second block of data to this set. This block is a duplication of the records for the patients who have not had a recurrence at the six-month mark and for whom the last visit is made after 6 months. There are 38 patients who have not had a recurrence at six months, but three of these, patients 16, 28 and 34, took no further part in the study. The second block of data therefore contains 35 records. The variable *Period* now takes the value 2 for these 35 observations, since they correspond to the second time period. The variable *R* contains the values  $r_{i2}$  for this second block of data. Therefore, *R* takes the value unity for patients 1, 17, 18, 30, 32 and 37 and zero otherwise,

since these are the only six patients who have a detectable recurrence at 12 months.

The combined set of data has  $43 + 35 = 78$  rows, and includes the variable *Period*, which defines the period in which an endoscopy is performed ( $1 = 0\text{--}6$  months,  $2 = 6\text{--}12$  months), and the variable *R*, which defines the endoscopy result ( $0 = \text{negative}$ ,  $1 = \text{positive}$ ). The value of  $s_{ij}$  is unity for all records except those for which  $r_{ij} = 1$ , when it is zero. The binomial denominators  $r_{ij} + s_{ij}$  are therefore equal to unity for each patient, since every patient in the extended data set is at risk of a detectable recurrence. Instead of giving a full listing of the modified data set, the records for the combinations of patient and period, for the first 18 patients, are shown in Table 9.4.

The dependence of the complementary log-log transformation of the probabilities  $\pi_{ij}$  on certain explanatory variables can now be investigated by fitting models to the binary response variable *R*. Each model includes a two-level factor, *Period*, associated with the period, but no constant term. The term in the model that corresponds to the period effect is  $\gamma_j$ ,  $j = 1, 2$ . The deviances for the models fitted are summarised in Table 9.5.

From this table we see that the effect of adding either *Age* or *Dur* to the model that contains *Period* alone is to reduce the deviance by less than 0.3. There is therefore no evidence that the age of a person or the duration of disease are associated with a recurrence. Adding *Treat* to the model that contains *Period* alone, the reduction in deviance is 0.10 on 1 d.f. This leads us to conclude that there is no significant difference between the two treatments. The treatment effect, after adjusting for the variables *Age* and *Dur*, is of a similar magnitude.

To check whether there are interactions between treatment and the two prognostic factors, we look at the effect of adding the terms *Treat*  $\times$  *Age* and *Treat*  $\times$  *Dur* to that model that contains *Period*, *Age* and *Dur*. From Table 9.5, the resulting change in deviance is very small, and so there is no evidence of any such interactions.

In summary, the modelling process shows that  $\pi_{ij}$ , the probability that the  $i$ th patient has a recurrence in the  $j$ th period, does not depend on the patient's age or the duration of the disease, and, more importantly, does not depend on the treatment group.

To further quantify the treatment effect, consider the model that includes both *Treat* and *Period*. The equation of the fitted model can be written as

$$\log\{-\log(1 - \hat{\pi}_{ij})\} = \hat{\gamma}_j + \hat{\beta} \text{ } \text{Treat}_i, \quad (9.6)$$

where  $\gamma_j$  is the effect of the  $j$ th period,  $j = 1, 2$ , and  $\text{Treat}_i$  is the value of the indicator variable *Treat*, for the  $i$ th individual. This variable is zero if that patient is on treatment *A* and unity otherwise.

The estimated coefficient of *Treat* in this model is 0.195 and the standard error of this estimate is 0.626. The hazard of a recurrence on treatment *B* at any given time, relative to that on treatment *A*, is  $\exp(0.195) = 1.21$ . Since this exceeds unity, there is the suggestion that the risk of recurrence is less on treatment *A* than on treatment *B*, but the evidence for this is not statis-

**Table 9.4** Modified data on the recurrence of an ulcer in two periods, for the first 18 patients.

Patient	Age	Duration	Treatment	Time of last visit	Result	Period	R
1	48	2	B	7	2	1	0
1	48	2	B	7	2	2	1
2	73	1	B	12	1	1	0
2	73	1	B	12	1	2	0
3	54	1	B	12	1	1	0
3	54	1	B	12	1	2	0
4	58	2	B	12	1	1	0
4	58	2	B	12	1	2	0
5	56	1	A	12	1	1	0
5	56	1	A	12	1	2	0
6	49	2	A	12	1	1	0
6	49	2	A	12	1	2	0
7	71	1	B	12	1	1	0
7	71	1	B	12	1	2	0
8	41	1	A	12	1	1	0
8	41	1	A	12	1	2	0
9	23	1	B	12	1	1	0
9	23	1	B	12	1	2	0
10	37	1	B	5	2	1	1
11	38	1	B	12	1	1	0
11	38	1	B	12	1	2	0
12	76	2	B	12	1	1	0
12	76	2	B	12	1	2	0
13	38	2	A	12	1	1	0
13	38	2	A	12	1	2	0
14	27	1	A	6	2	1	1
15	47	1	B	6	2	1	1
16	54	1	A	6	1	1	0
17	38	1	B	10	2	1	0
17	38	1	B	10	2	2	1
18	27	2	B	7	2	1	0
18	27	2	B	7	2	2	1

tically significant. The standard error of the estimated hazard ratio is 0.757. For comparison, from Example 9.2, the estimated hazard ratio at 12 months was found to be 1.46, with a standard error of 0.918. These values are very similar to those obtained in this example. Moreover, the results of analyses that accommodate interval censoring are comparable to those found in Example 9.1, in which the Cox proportional hazards model was used without taking account of the fact that the data are interval-censored.

The model in equation (9.6) can be used to provide estimates of the  $\pi_{ij}$ . The estimates of the period effects in this model are  $\hat{\gamma}_1 = -2.206$ ,  $\hat{\gamma}_2 = -1.794$ ,

**Table 9.5** Deviances on fitting complementary log-log models that do not include a constant to the variable R.

Terms fitted in model	Deviance	d.f.
Period	62.982	76
Period + Age	62.685	75
Period + Dur	62.979	75
Period + Age + Dur	62.685	74
Period + Age + Dur + Treat	62.566	73
Period + Age + Dur + Treat + Treat × Age	62.278	72
Period + Age + Dur + Treat + Treat × Dur	62.552	72
Period + Treat	62.884	75

and so the estimated probability of a recurrence in the first period, for a patient on treatment A, denoted  $\hat{\pi}_{A1}$ , is given by

$$\log\{-\log(1 - \hat{\pi}_{A1})\} = \hat{\gamma}_1 + \hat{\beta} \times 0.$$

Therefore,

$$\log\{-\log(1 - \hat{\pi}_{A1})\} = -2.206,$$

and from this,  $\hat{\pi}_{A1} = 0.104$ . Other fitted probabilities can be calculated in a similar manner, and the results of these calculations are shown in Table 9.6. The corresponding observed proportions of individuals with a recurrence for each combination of treatment and period are also displayed. The agreement between the observed and fitted probabilities is good, which indicates that the model is a good fit.

**Table 9.6** Fitted and observed probabilities of an ulcer recurring in the two time periods.

Period	Treatment A		Treatment B	
	Fitted	Observed	Fitted	Observed
(0, 6)	0.104	0.158	0.125	0.083
(6, 12)	0.153	0.077	0.183	0.227

If desired, probabilities of a recurrence in either period 1 or period 2 could also be estimated. The probability that a patient on treatment A has a recurrence in either period 1 or period 2 is

$$P\{\text{recurrence in } (0, 6)\} + P\{\text{recurrence in } (6, 12) \text{ and no recurrence in } (0, 6)\}.$$

The joint probability of a recurrence in (6, 12) and no recurrence in (0, 6) can be expressed as

$$P\{\text{recurrence in } (6, 12) \mid \text{no recurrence in } (0, 6)\} \times P\{\text{no recurrence in } (0, 6)\},$$

and so the required probability is estimated by

$$\hat{\pi}_{A1} + \hat{\pi}_{A2}(1 - \hat{\pi}_{A1}) = 0.104 + 0.153 \times 0.896 = 0.241.$$

Similarly, that for treatment  $B$  is

$$\hat{\pi}_{B1} + \hat{\pi}_{B2}(1 - \hat{\pi}_{B1}) = 0.125 + 0.183 \times 0.875 = 0.285.$$

This again indicates the superiority of treatment  $A$ , but there is insufficient data for this effect to be declared significant.

#### 9.4\* Arbitrarily interval-censored survival data

The methods of analysis described in the previous sections may be adopted when different individuals have the same observation times. In this section, we consider a more general form of interval censoring, where the observation times differ between individuals. Then, each individual may have a different time interval in which the event of interest has occurred, and data in this form are referred to as *arbitrarily interval-censored data*. A method for analysing such data, assuming proportional hazards, that is based on a non-linear model for binary data, proposed by Farrington (1996), is now developed.

##### 9.4.1 Modelling arbitrarily interval-censored data

Suppose that the event time for the  $i$ th of  $n$  individuals is observed to occur in the interval  $(a_i, b_i]$ , where the use of different types of bracket indicates that the actual event time is greater than  $a_i$ , but less than or equal to  $b_i$ . In other words, the event has not occurred by time  $a_i$ , but has occurred by time  $b_i$ , where the values of  $a_i$  and  $b_i$  may well be different for each individual in the study. We will further suppose that the values of a number of explanatory variables have also been recorded for each individual in the study.

When the values of both  $a_i$  and  $b_i$  are observed for an individual, the interval-censored observation is said to be *confined*. If the event time for an individual is left-censored at time  $b_i$ , so that the event is only known to have occurred some time before  $b_i$ , then  $a_i = 0$ . Similarly, if the event time right-censored at time  $a_i$ , so that the event is only known to have occurred after time  $a_i$ , the upper limit of the interval,  $b_i$ , is then effectively infinite.

The survivor function for the  $i$ th individual will be denoted by  $S_i(t)$ , so that the probability of an event occurring in the interval  $(a_i, b_i]$  is  $S_i(a_i) - S_i(b_i)$ . The likelihood function for the  $n$  observations is then

$$\prod_{i=1}^n \{S_i(a_i) - S_i(b_i)\}. \quad (9.7)$$

Now suppose that the  $n$  observations consist of  $l$  left-censored observations,  $r$  right-censored observations, and  $c$  observations that are confined, with  $n = l + r + c$ . For the purpose of this exposition, we will assume that the data have been arranged in such a way that the first  $l$  observations are left-censored ( $a_i = 0$ ), the next  $r$  are right-censored ( $b_i = \infty$ ), and the remaining  $c$  observations are

confined ( $0 < a_i < b_i < \infty$ ). Since  $S_i(0) = 1$  and  $S_i(\infty) = 0$ , the contributions of a left- and right-censored observation to the likelihood function will be  $1 - S_i(b_i)$  and  $S_i(a_i)$ , respectively. Consequently, from equation (9.7), the overall likelihood function can be written as

$$\prod_{i=1}^l \{1 - S_i(b_i)\} \prod_{i=l+1}^{l+r} S_i(a_i) \prod_{i=l+r+1}^n \{S_i(a_i) - S_i(b_i)\},$$

and a re-expression of the final product in this function gives

$$\prod_{i=1}^l \{1 - S_i(b_i)\} \prod_{i=l+1}^{l+r} S_i(a_i) \prod_{i=l+r+1}^n S_i(a_i) \{1 - S_i(b_i)/S_i(a_i)\}. \quad (9.8)$$

We now show that this likelihood is equivalent to that for a corresponding set of  $n + c$  independent binary observations,  $y_1, y_2, \dots, y_{n+c}$ , where the  $i$ th is assumed to be an observation from a Bernoulli distribution with response probability  $p_i$ ,  $i = 1, 2, \dots, n+c$ . The likelihood function for this set of binary data is then

$$\prod_{i=1}^{n+c} p_i^{y_i} (1 - p_i)^{1-y_i}, \quad (9.9)$$

where  $y_i$  takes the value 0 or 1, for  $i = 1, 2, \dots, n+c$ .

To see the relationship between the response probabilities,  $p_i$ , in expression (9.9), and the values of the survivor function in expression (9.8), consider first the left-censored observations. Suppose that each of these  $l$  observations contributes a binary observation with  $y_i = 1$  and  $p_i = 1 - S_i(b_i)$ ,  $i = 1, 2, \dots, l$ . The contribution of these  $l$  observations to expression (9.9) is then

$$\prod_{i=1}^l p_i = \prod_{i=1}^l \{1 - S_i(b_i)\},$$

which is the first term in expression (9.8). For a right-censored observation, we take  $y_i = 0$  and  $p_i = 1 - S_i(a_i)$  in expression (9.9), and the contribution to the likelihood function in expression (9.9) from  $r$  such observations is

$$\prod_{i=l+1}^{l+r} (1 - p_i) = \prod_{i=l+1}^{l+r} S_i(a_i),$$

and this is the second term in expression (9.8). The situation is a little more complicated for an observation that is confined to the interval  $(a_i, b_i]$ , since two binary observations are needed to give the required component of expression (9.8). One of these is taken to have  $y_i = 0$ ,  $p_i = 1 - S_i(a_i)$ , while the other is such that  $y_{c+i} = 1$ ,  $p_{c+i} = 1 - \{S_i(b_i)/S_i(a_i)\}$ , for  $i = l+r+1, l+r+2, \dots, n$ . Combining these two terms leads to a component of the likelihood in expression (9.9) of the form

$$\prod_{i=l+r+1}^n (1 - p_i)p_{c+i},$$

which corresponds to

$$\prod_{i=l+r+1}^n S_i(a_i) \{1 - S_i(b_i)/S_i(a_i)\}$$

in expression (9.8).

This shows that by suitably defining a set of  $n + c$  binary observations, with response probabilities expressed in terms of the survivor functions for the three possible forms of interval-censored observation, the likelihood function in expression (9.9) is equivalent to that in expression (9.8). Accordingly, maximisation of the log-likelihood function for  $n + c$  binary observations is equivalent to maximising the log-likelihood for the interval-censored data.

#### 9.4.2 Proportional hazards model for the survivor function

The next step in the development of a procedure for modelling arbitrarily interval-censored survival data is to construct expressions for the survivor functions that make up the likelihood function in expression (9.8). A proportional hazards model will be assumed, so that from equation (9.1),

$$S_i(t) = S_0(t)^{\exp(\beta' \mathbf{x}_i)}, \quad (9.10)$$

where  $S_0(t)$  is the baseline survivor function and  $\mathbf{x}_i$  is the vector of values of  $p$  explanatory variables for the  $i$ th individual,  $i = 1, 2, \dots, n$ , with coefficients that make up the vector  $\beta$ .

The baseline survivor function will be modelled as a step function, where the steps occur at the  $k$  ordered censoring times,  $t_{(1)}, t_{(2)}, \dots, t_{(k)}$ , where  $0 < t_{(1)} < t_{(2)} < \dots < t_{(k)}$ , which are a subset of the times at which observations are interval-censored. This means that the  $t_{(j)}$ ,  $j = 1, 2, \dots, k$ , are a subset of the values of  $a_i$  and  $b_i$ ,  $i = 1, 2, \dots, n$ . Exactly how these times are chosen will be described later in Section 9.4.3.

We now define

$$\theta_j = \log \frac{S_0(t_{(j-1)})}{S_0(t_{(j)})},$$

where  $t_{(0)} = 0$ , so that  $\theta_j \geq 0$ , and at times  $t_{(j)}$ , we have

$$S_0(t_{(j)}) = e^{-\theta_j} S_0(t_{(j-1)}), \quad (9.11)$$

for  $j = 1, 2, \dots, k$ .

Since the first step in the baseline survivor function occurs at  $t_{(1)}$ ,  $S_0(t) = 1$  for  $0 \leq t < t_{(1)}$ . From time  $t_{(1)}$ , the baseline survivor function, using equation (9.11), has the value  $S_0(t_{(1)}) = \exp(-\theta_1) S_0(t_{(0)})$ , which, since  $t_{(0)} = 0$ , means that  $S_0(t) = \exp(-\theta_1)$ , for  $t_{(1)} \leq t < t_{(2)}$ . Similarly, from time  $t_{(2)}$ , the survivor function is  $\exp(-\theta_2) S_0(t_{(1)})$ , that is  $S_0(t) = \exp\{-(\theta_1 + \theta_2)\}$ ,  $t_{(2)} \leq t < t_{(3)}$ , and so on, until  $S_0(t) = \exp\{-(\theta_1 + \theta_2 + \dots + \theta_k)\}$ ,  $t \geq t_{(k)}$ . Consequently,

$$S_0(t) = \exp\left(-\sum_{r=1}^j \theta_r\right), \quad (9.12)$$

for  $t_{(j)} \leq t < t_{(j+1)}$ , and so the baseline survivor function, at any time  $t_i$ , is given by

$$S_0(t_i) = \exp\left(-\sum_{j=1}^k \theta_j d_{ij}\right), \quad (9.13)$$

where

$$d_{ij} = \begin{cases} 1 & \text{if } t_{(j)} \leq t_i, \\ 0 & \text{if } t_{(j)} > t_i, \end{cases}$$

for  $j = 1, 2, \dots, k$ . The quantities  $d_{ij}$  will be taken to be the values of  $k$  indicator variables,  $D_1, D_2, \dots, D_k$ , for the  $i$ th observation in the augmented data set. Note that the values of the  $D_j$ ,  $j = 1, 2, \dots, k$ , will differ at each observation time,  $t_i$ .

Combining the results in equations (9.10) and (9.13), the survivor function for the  $i$ th individual, at times  $a_i, b_i$ , can now be obtained. In particular,

$$S_i(a_i) = S_0(a_i)^{\exp(\beta' \mathbf{x}_i)} = \left\{ \exp\left(-\sum_{j=1}^k \theta_j d_{ij}\right) \right\}^{\exp(\beta' \mathbf{x}_i)},$$

which can be expressed in the form

$$S_i(a_i) = \exp\left\{-\exp(\beta' \mathbf{x}_i) \sum_{j=1}^k \theta_j d_{ij}\right\},$$

where  $d_{ij} = 1$  if  $t_{(j)} \leq a_i$ , and  $d_{ij} = 0$ , otherwise. An expression for  $S_i(b_i)$  can be obtained in a similar way, leading to

$$S_i(b_i) = \exp\left\{-\exp(\beta' \mathbf{x}_i) \sum_{j=1}^k \theta_j d_{ij}\right\},$$

where  $d_{ij} = 1$  if  $t_{(j)} \leq b_i$ , and  $d_{ij} = 0$ , otherwise.

From these expressions for  $S_i(a_i)$  and  $S_i(b_i)$ , the response probabilities,  $p_i$ , used in expression (9.9), can be expressed in terms of the unknown parameters  $\theta_1, \theta_2, \dots, \theta_k$ , and the unknown coefficients of the  $p$  explanatory variables in the model,  $\beta_1, \beta_2, \dots, \beta_p$ . Specifically, as in Section 9.4.1, for a left-censored observation,  $p_i = 1 - S_i(b_i)$ , and for a right-censored observation,  $p_i = 1 - S_i(a_i)$ . In the case of a confined observation,  $p_i = 1 - S_i(a_i)$  for one of the two binary observations. For the other,

$$p_{c+i} = 1 - S_i(b_i)/S_i(a_i), \\ = 1 - \frac{\exp\left\{-\exp(\beta' \mathbf{x}_i) \sum_{j=1}^k \theta_j d_{1ij}\right\}}{\exp\left\{-\exp(\beta' \mathbf{x}_i) \sum_{j=1}^k \theta_j d_{2ij}\right\}},$$

where the values  $d_{1ij}$  in the numerator are equal to unity if  $t_{(j)} \leq b_i$ , and zero otherwise, and the values  $d_{2ij}$  in the denominator are equal to unity if  $t_{(j)} \leq a_i$ , and zero otherwise. Consequently, the  $\theta$ -terms in the numerator for which  $t_{(j)} \leq a_i$  cancel with those in the denominator, and this leaves

$$p_{c+i} = 1 - \exp\left\{-\exp(\beta' \mathbf{x}_i) \sum_{j=1}^k \theta_j d_{ij}\right\},$$

where here  $d_{ij} = 1$  if  $a_i < t_{(j)} \leq b_i$ .

It then follows that in each case, the response probability can be expressed in the form

$$p_i = 1 - \exp \left\{ -\exp(\beta' x_i) \sum_{j=1}^k \theta_j d_{ij} \right\}, \quad (9.14)$$

where

$$d_{ij} = \begin{cases} 1 & \text{if } t_{(j)} \text{ is in the interval } A_i, \\ 0 & \text{otherwise,} \end{cases}$$

for  $j = 1, 2, \dots, k$ , and the intervals  $A_i$  are as shown in Table 9.7.

**Table 9.7** Definition of intervals,  $A_i$ , used for constructing indicator variables.

Type of observation	Value of $y_i$	Interval, $A_i$
Left-censored	1	$(0, b_i]$ , $i = 1, 2, \dots, l$
Right-censored	0	$(0, a_i]$ , $i = l+1, l+2, \dots, l+r$
Confined	0	$(0, a_i]$ , $i = l+r+1, l+r+2, \dots, n$
	1	$(a_{i-c}, b_{i-c}]$ , $i = n+1, n+2, \dots, n+c$

This leads to a non-linear model for a set of binary response variables, with values  $y_i$ , and corresponding response probabilities  $p_i$ , found from equation (9.14), for  $i = 1, 2, \dots, n+c$ . The model contains  $k+p$  unknown parameters, namely  $\theta_1, \theta_2, \dots, \theta_k$  and  $\beta_1, \beta_2, \dots, \beta_p$ . This model is known as a *generalised non-linear model*, since it is not possible to express a simple function of  $p_i$  as a linear combination of the unknown parameters, except in the case where there are no explanatory variables in the model. The model can be fitted using computer software for generalised non-linear modelling, such as SAS proc nlmixed. Note that in the fitting process, the  $\theta$ -parameters should be constrained to be non-negative.

After fitting a model, the value of the statistic  $-2 \log \hat{L}$  can be found, and this may be used to compare alternative models in the usual manner. The general procedure is to fit the  $k$  terms involving the  $\theta$ -parameters, and to then examine the effect of adding and subtracting the explanatory variables in the data set.

Once an appropriate model has been found, the baseline survivor function in equation (9.12) is estimated using

$$\hat{S}_0(t) = \exp \left( -\sum_{r=1}^j \hat{\theta}_r \right), \quad (9.15)$$

for  $t_{(j)} \leq t < t_{(j+1)}$ ,  $j = 1, 2, \dots, k$ , where  $t_{(k+1)} = \infty$ , and  $\hat{\theta}_j$  is the estimated value of  $\theta_j$ . The estimated survivor function for the  $i$ th individual follows from

$$\begin{aligned} \hat{S}_i(t) &= \hat{S}_0(t)^{\exp(\beta' x_i)} \\ &= \exp \left\{ -\exp(\hat{\beta}' x_i) \sum_{r=1}^j \hat{\theta}_r \right\}, \end{aligned} \quad (9.16)$$

for  $i = 1, 2, \dots, n$ , where  $\hat{\beta}$  is the vector of estimated coefficients of the explanatory variables. Furthermore, the estimates of the  $\beta$ -parameters are interpretable as log-hazard ratios, in the usual manner, and their standard errors, produced during the fitting process, can be used to obtain confidence limits.

#### 9.4.3 Choice of the step times

We have seen that the baseline survivor function is assumed to have steps at times  $t_{(j)}$ ,  $j = 1, 2, \dots, k$ , which are a subset of the observed censoring times,  $a_i$  and  $b_i$ ,  $i = 1, 2, \dots, n$ . It might be considered to be desirable for the  $t_{(j)}$  to be formed from all distinct censoring times, that is all the unique values of  $a_i$  and  $b_i$ . However, this will generally lead to the introduction of far too many  $\theta$ -parameters in the non-linear model. Instead, a subset of the available times is chosen.

Each interval used in the binary data model, and denoted by  $A_i$  in the preceding section, must include at least one of the times  $t_{(j)}$ . If this is the case, at least one of the values of  $d_{ij}$  in equation (9.14) will be equal to unity and hence the term  $\sum_{j=1}^k \theta_j d_{ij}$  will be greater than zero. Suppose that the interval  $A_i$  is  $(u_i, v_i]$ . This requirement is then achieved by taking  $t_{(1)}$  to be the smallest of the values of  $v_i$ ,  $t_{(2)}$  to be the smallest  $v_i$  such that  $u_i \geq t_{(1)}$ ,  $t_{(3)}$  to be the smallest  $v_i$  such that  $u_i \geq t_{(2)}$ , and so on, until  $t_{(k)}$  is the smallest value of  $v_i$  such that  $u_i \geq t_{(k-1)}$ .

Once this subset of  $k$  times has been identified, the model can be fitted. Models containing explanatory variables are fitted, and for each model the estimates of the  $k$   $\theta$ -parameters and the relevant  $\beta$ -parameters are found. The fitting process will lead to a value of the  $-2 \log \hat{L}$  statistic, and these values can be compared for models with the same number of  $\theta$ -parameters, but different explanatory variables, in the usual way.

Sometimes, it may be desirable to increase the number of steps in the estimated baseline hazard function, by the addition of some of the remaining censoring times. This entails adding a new  $\theta$ -parameter for each additional time point. One way of doing this is to fit the minimal set of censoring times, and to then add each additional time point in turn, for the full set of explanatory variables. The time that leads to the biggest reduction in the value of  $-2 \log \hat{L}$  is then added to the minimal set. All remaining times are then added one by one, and again that which reduces  $-2 \log \hat{L}$  the most is added to the set. This process may be continued until the reduction on adding an additional  $\theta$ -parameter ceases to be significant at some chosen level, and so long as all the estimated  $\theta$ -parameters remain positive.

It is important to note that this modelling procedure is only valid if the set of possible censoring times is finite, that is, it does not increase with the number of observations. Otherwise, the number of  $\theta$ 's increases indefinitely and asymptotic results used in comparing models will no longer be valid.

This procedure for modelling arbitrarily interval-censored data is now illustrated.

*Example 9.4 Occurrence of breast retraction*

In the treatment of early breast cancer, a tumourectomy, followed by radiation therapy, is often used as an alternative to mastectomy. Chemotherapy may also be used in conjunction with the radiotherapy in order to enhance its effect, but there is evidence that this adjuvant chemotherapy increases the effect of the radiation on normal tissue. This in turn leads to breast retraction, which has a negative impact on the appearance of the breast. In a retrospective study to assess the effect of this type of treatment on breast cosmesis, 46 women who had been treated with radiotherapy alone were compared with 48 who had received a combination of radiotherapy and chemotherapy. Patients were observed every 4 to 6 months, but less frequently as their recovery progressed. On these occasions, the cosmetic effect of the treatment was monitored, with the extent of breast retraction being measured on a four-point scale: none, minimal, moderate, severe. The event of interest was the time in months to the first appearance of moderate or severe retraction. The exact time of occurrence of breast retraction will be unknown, and the only information available will concern whether or not retraction is identified when a patient visits the clinic. Moreover, since the visit times were not the same for each patient, and a number of patients failed to keep appointments, the data are regarded as arbitrarily interval-censored. The data obtained in this study were included in Finkelstein and Wolfe (1985), and are given in Table 9.8.

**Table 9.8** Data on the time in months to breast retraction in patients with breast cancer.

Radiotherapy		Radiotherapy and Chemotherapy		
(45, *)	(25, 37]	(37, *)	(8, 12]	(0, 5]
(6, 10]	(46, *)	(0, 5]	(0, 22]	(5, 8]
(0, 7]	(26, 40]	(18, *)	(24, 31]	(12, 20]
(46, *)	(46, *)	(24, *)	(17, 27]	(11, *)
(46, *)	(27, 34]	(36, *)	(17, 23]	(33, 40]
(7, 16]	(36, 44]	(5, 11]	(24, 30]	(31, *)
(17, *)	(46, *)	(19, 35]	(16, 24]	(13, 39]
(7, 14]	(36, 48]	(17, 25]	(13, *)	(19, 32]
(37, 44]	(37, *)	(24, *)	(11, 13]	(34, *)
(0, 8]	(40, *)	(32, *)	(16, 20]	(13, *)
(4, 11]	(17, 25]	(33, *)	(18, 25]	(16, 24]
(15, *)	(46, *)	(19, 26]	(17, 26]	(35, *)
(11, 15]	(11, 18]	(37, *)	(32, *)	(15, 22]
(22, *)	(38, *)	(34, *)	(23, *)	(11, 17]
(46, *)	(5, 12]	(36, *)	(44, 48]	(22, 32]
(46, *)			(14, 17]	(10, 35]
				(48, *)

In this data set, there are five patients for whom breast retraction had occurred before the first visit. For each of these patients, the start of the interval is set to zero, that is  $a_i = 0$ , and the observed times are left-censored,

so that  $l = 5$ . There are 38 patients who had not experienced retraction by their final visit. The upper limit of the time interval that includes the event time is then shown as an asterisk (\*) in Table 9.8. The observations for these patients are therefore right-censored and so  $r = 38$ . The remaining  $c = 51$  patients experience breast retraction within confined time intervals, and the total number of observations is  $n = l + c + r = 94$ .

The first step in fitting a model to these arbitrarily interval-censored data is to expand the data set by adding a further 51 lines of data, repeating that for the patients whose intervals are confined, so that the revised data base has  $n + c = 145$  observations. The values,  $y_i$ , of the binary response variable,  $Y$ , are then added. These are such that  $Y = 1$  for a left-censored observation, and  $Y = 0$  for a right-censored observation. For confined observations, where the data are duplicated, one of the pairs of observations has  $Y = 0$  and the other  $Y = 1$ . The treatment effect will be represented by the value of a variable labelled *Treat*, which will be zero for a patient on radiotherapy and unity for a patient on radiotherapy and chemotherapy. For illustration, the values of the binary response variable,  $Y$ , are shown for the first three patient treated with radiotherapy alone, in Table 9.9.

**Table 9.9** Augmented data set for the first three patients on radiotherapy.

Patient	A	B	U	V	Treat	Y
1	45	*	0	45	0	0
2	6	10	0	6	0	0
2	6	10	6	10	0	1
3	0	7	0	7	0	1

In this table, the variables *A* and *B* refer to the times of the start and end of each interval, and so their values are  $a_i, b_i$ , and the variables *U* and *V* contain the values of  $u_i, v_i$  that form the limits of the intervals,  $A_i$ , in the binary data model.

We now determine the time points,  $t_{(j)}$ , that are to be used in calculating the baseline survivor function, using the procedure described in Section 9.4.3. The first of these is the smallest of the values  $v_i$ , which form the variable *V* in the data set. This is found for the observation  $(0, 4]$ , for which  $V = 4$ , so that  $t_{(1)} = 4$ . The smallest value of *V* for which  $U \geq 4$  occurs for the observation  $(4, 8]$ , and so  $t_{(2)} = 8$ . Next, the smallest value of *V* with  $U \geq 8$  occurs for the observation  $(8, 12]$ , giving  $t_{(3)} = 12$ . There are six other times found in this way, namely 17, 23, 30, 34, 39, and 48, and so the minimal subset of times has  $k = 9$  in this example.

Variables with values  $d_{ij}$ , that correspond to each of the  $k$  censoring times,  $t_{(j)}$ , are now added to the data base that has 145 records. Since  $k = 9$  in this case, nine variables,  $D_1, D_2, \dots, D_9$ , are introduced, where the values of  $D_j$ ,  $j = 1, 2, \dots, 9$ , are  $d_{ij}$ , for  $i = 1, 2, \dots, 145$ . These values are such that  $d_{ij} = 1$

if  $u_i < t_{(j)} \leq v_i$ , and zero otherwise, and so they can straightforwardly be obtained from the variables  $U$  and  $V$  in Table 9.9. The values of  $D_1, D_2, \dots, D_9$ , for the three patients included in Table 9.9, are shown in Table 9.10.

**Table 9.10** Data base for binary data analysis, for the first three patients on radiotherapy.

Patient	Treat	$Y$	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$	$D_6$	$D_7$	$D_8$	$D_9$
1	0	0	1	1	1	1	1	1	1	1	0
2	0	0	1	0	0	0	0	0	0	0	0
2	0	1	0	1	0	0	0	0	0	0	0
3	0	1	1	0	0	0	0	0	0	0	0

We now have the data base to which a non-linear model for the binary response data in  $Y$  is fitted. The model is such that we take  $Y$  to have a Bernoulli distribution with response probability as given in equation (9.14), that is a binomial distribution with parameters 1,  $p_i$ . The SAS procedure proc nlmixed has been used to fit the non-linear model to the binary data.

On fitting the null model, that is the model that contains all 9  $D$ -variables, but not the treatment effect, the value of the statistic  $-2 \log \hat{L}$  is 285.417. On adding *Treat* to the model, the value of  $-2 \log \hat{L}$  is reduced to 276.983. This reduction of 8.43 on 1 d.f. is significant at the 1% level ( $P = 0.0037$ ), and so we conclude that the interval-censored data do provide strong evidence for a treatment effect. The estimated coefficient of *Treat* is 0.8212, with a standard error of 0.2881. The corresponding hazard ratio for a patient on the combination of radiotherapy and chemotherapy, relative to a patient on radiotherapy alone, is  $\exp(0.8212) = 2.273$ . The interpretation of this is that patients on the combined treatment have just over twice the risk of breast retraction, compared to patients on radiotherapy. A 95% confidence interval for the corresponding true hazard ratio has limits  $\exp(0.8212 \pm 1.96 \times 0.2881)$ , which leads to the interval (1.29, 4.00).

The minimal subset of times at which the estimated baseline survivor function is estimated can be enlarged by adding additional censoring times from the data set. However, there are no additional times that lead to a significant reduction in the value of the  $-2 \log \hat{L}$  statistic, with the estimated  $\theta$ -parameters remaining positive.

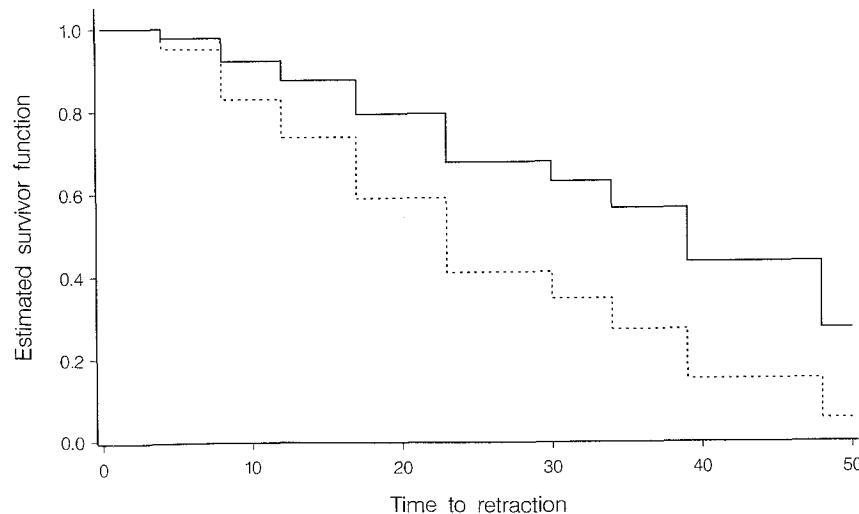
The estimated values of the coefficients of the  $D$ -variables are the values  $\hat{\theta}_j$ , for  $j = 1, 2, \dots, 9$ , and these can be used to provide an estimate of the survivor function for the two treatment groups. Equation (9.15), gives the form of the estimated baseline survivor function, which is the estimated survivor function for the patients on radiotherapy alone. The corresponding estimate for the patients who receive adjuvant chemotherapy is obtained using equation (9.16), and is just  $\{\hat{S}_0(t_{(j)})\}^{\exp(\hat{\beta})}$ , with  $\hat{\beta} = 0.8212$ . On fitting the model that contains the treatment effect and the 9  $D$ -variables, the estimated values of  $\theta_j$ , for  $j = 1, 2, \dots, 9$ , are 0.0223, 0.0603, 0.0524, 0.0989, 0.1620, 0.0743,

0.1098, 0.2633 and 0.4713, respectively. From these values, the baseline survivor function, at the times  $t_{(j)}$ ,  $j = 1, 2, \dots, 9$ , can be estimated, and this estimate is given as  $\hat{S}_0(t_{(j)})$  in Table 9.11. Also given is the estimated survivor function for patients on the combined treatment, denoted  $\hat{S}_1(t_{(j)})$ .

**Table 9.11** Estimated survivor functions for a patient on radiotherapy alone,  $\hat{S}_0(t)$ , and adjuvant chemotherapy,  $\hat{S}_1(t)$ .

Time interval	$\hat{S}_0(t)$	$\hat{S}_1(t)$
0–	1.000	1.000
4–	0.978	0.951
8–	0.921	0.829
12–	0.874	0.736
17–	0.791	0.588
23–	0.673	0.407
30–	0.625	0.343
34–	0.560	0.268
39–	0.430	0.147
48–	0.269	0.050

These functions are also plotted in Figure 9.1.



**Figure 9.1** Estimated survivor functions for a patient on radiotherapy (—) and the combination of radiotherapy and chemotherapy (···).

From the estimated survivor functions, the median time to breast retraction for patients on radiotherapy is estimated to be 39 months, while that

for patients who received adjuvant chemotherapy is 23 months. More precise estimates of these median times could be obtained if a greater number of censoring times were used in the analysis.

### 9.5 Parametric models for interval-censored data

The methods described for modelling interval-censored data that have been described in previous sections are based on the Cox proportional hazards model. In fact, it is much more straightforward to model such data assuming a parametric model for the survival times. In Section 9.4, the likelihood function for  $n$  arbitrarily interval-censored observations consisting of  $l$  that are left-censored at time  $a_i$ ,  $r$  that are right-censored at  $b_i$ , and  $c$  that are confined to the interval  $(a_i, b_i)$ , was given as

$$\prod_{i=1}^l \{1 - S_i(b_i)\} \prod_{i=l+1}^{l+r} S_i(a_i) \prod_{i=l+r+1}^n \{S_i(a_i) - S_i(b_i)\}. \quad (9.17)$$

If a parametric model for the survival times is assumed, then  $S_i(t)$  has a fully parametric form. For example, if the survival times have a Weibull distribution with scale parameter  $\lambda$  and shape parameter  $\gamma$ , from equation (5.37) in Section 5.5 of Chapter 5, the survivor function for the  $i$ th individual is

$$S_i(t) = \exp\{-\exp(\beta' \mathbf{x}_i)\lambda t^\gamma\},$$

where  $\mathbf{x}_i$  is the vector of values of explanatory variables for that individual, with coefficients  $\beta$ . Alternatively, the accelerated failure time form of the survivor function, given in equation (6.17) of Section 6.5 of Chapter 6, may be used. The Weibull survivor function leads to expressions for the survivor functions in expression (9.17), for any values of  $a_i$  and  $b_i$ . The corresponding log-likelihood function can then be maximised with respect to the parameters  $\lambda$ ,  $\gamma$  and the  $\beta$ 's. No new principles are involved. The same procedure can be adopted for any other parametric model described in Chapter 6.

In some situations, the data base may be a combination of censored and uncensored observations. The likelihood function in expression (9.17) may then be further extended to allow for the uncensored observations. This is achieved by including an additional factor of the form  $\prod_i f(t_i)$ , which is the product of the density functions at the event times, over the uncensored observations. A number of software packages include facilities for the parametric modelling of interval-censored data, including the SAS procedure proc lifereg.

#### *Example 9.5 Occurrence of breast retraction*

The interval-censored data on the times to breast retraction in women being treated for breast cancer, given in Example 9.4, are now used to illustrate parametric modelling for interval-censored data.

If a common Weibull distribution is assumed for the data in both treatment groups, the value of  $-2 \log \hat{L}$  is 297.585. If the treatment effect is added to this model, and the assumption of proportional hazards is made, so that the shape parameter is constant, the value of  $-2 \log \hat{L}$  reduces to 286.642. This

reduction of 10.943 on 1 d.f. is highly significant ( $P = 0.001$ ). Comparing this with the results of the analysis in Example 9.4, we find that the significance of the treatment effect is a little greater under the assumed Weibull model. Furthermore, the estimated hazard ratio for a patient on radiotherapy with adjuvant chemotherapy, relative to a patient on radiotherapy alone, is 2.501, and a corresponding 95% confidence interval is (1.44, 4.35). These values do not differ much from those found in Example 9.4.

### 9.6 Discussion

In many situations, the method for analysing interval-censored survival data that has been presented in Section 9.4, or the fully parametric approach of Section 9.5, will be the most appropriate. Even when studies are designed to be such that the examination times are the same for each patient in the study, missed, postponed or cancelled appointments may lead to observation times that do differ across the patients. In this case, and for studies where this is a natural feature, methods for handling arbitrarily interval-censored data will be required.

When the observation times are the same for each patient, the method for analysing interval-censored data that has been presented in Section 9.3 will generally be the most suitable. However, this approach is not optimal, since recurrences detected between scheduled examinations, that is, interval-detected recurrences, are only counted at the next examination time. If the intervals between successive examination times are not too large, the difference between the results of an analysis based on the model in Section 9.3, and one that uses the actual times of interval-detected recurrences, will be negligible. In fact, if the number of intervals is not too small, and the time between successive examinations not too large, the results will not be too different from an analysis that assumes the recurrence times to be continuous, outlined in Section 9.1.

As mentioned earlier, the model described in Section 9.3 only requires hazards to be proportional at scheduled screening times. This means that the model is useful when the hazards are not necessarily proportional between screening times. Furthermore, the model could be relevant in situations where although actual survival times are available, the hazards can only be taken to be proportional at specific times. On the other hand, the method for analysing arbitrarily interval-censored data in Section 9.4, requires hazards to be proportional at each of the times used in constructing the baseline survivor function, which is more restrictive. Further comments on methods for analysing survival data where hazards are non-proportional are included in Chapter 10.

### 9.7 Further reading

Much of Sections 9.1 to 9.3 of this chapter are based on the summary of methods for processing interval-censored survival data given by Whitehead (1989). The approach described in Section 9.3 is based on Prentice and Gloeck-

ler (1978). A method for fitting the proportional hazards model to interval-censored data was proposed by Finkelstein (1986), but the method for modelling arbitrarily interval-censored data in Section 9.4 is due to Farrington (1996), who also develops additive and multiplicative models for such data.

A number of approaches to the analysis of interval-censored data involve the analysis of binary observations. Collett (2003) describes a model-based approach to the analysis of binary data. This book includes a description of the facilities available for modelling binary data in the major software packages. Other books that include material on the analysis of binary data include Hosmer and Lemeshow (2000), Dobson (2001) and Morgan (1992).

The use of the complementary log-log transformation in the analysis of interval-censored data was described by Thompson (1981). Becker and Melbye (1991) show how a log-linear model can be used to obtain an estimate of the survivor function from interval-censored data, assuming a constant hazard in each interval. Whitehead (1997) shows how the log-rank test for comparing two treatment groups can be adapted for use in processing interval-censored survival data. However, since this approach does not allow other explanatory variables to be taken account of, details are not given here.

There are a number of other approaches to the analysis of interval-censored data, some of which are described in the tutorial provided by Lindsey and Ryan (1998). Lindsey (1998) reviews the use of parametric models for the analysis of interval-censored data. Pan (2000) suggests using multiple imputation, based on the approximate Bayesian bootstrap, to impute values for censored observations. Farrington (2000) provides a comprehensive account of diagnostic methods for use with proportional hazards models for interval-censored data.

## CHAPTER 10

# Sample size requirements for a survival study

There are many aspects of the design of a clinical trial that must be considered when the response variable is a survival time. These include all the usual matters, such as patient eligibility, definition of the treatments, the method of randomisation to be employed in allocating patients to treatment group, and the use of blinding. In addition, care must be taken to define both the time origin and the end-point of the study in a clear and unambiguous manner. Consideration might also be given to whether the study should be based on a fixed number of patients, or whether a sequential design should be adopted, in which the study continues until there is a sufficient number of events to be able to distinguish between two treatments. The desirability of performing interim analyses might also need to be addressed.

In a book of this nature, there is insufficient space to be able to go into any of these topics in detail. Fortunately there are a number of excellent books that include material on the design of clinical trials, some of which include material on the design of trials for survival analysis, such as Pocock (1983).

However, there is one matter in the design of fixed sample size studies that will be discussed here. This is the crucial issue of the number of patients that are required in such studies. If too few patients are recruited, there may be insufficient information available in the data to enable a treatment difference to be pronounced significant. On the other hand, it is unethical to waste resources in studies that are unnecessarily large.

### 10.1 Distinguishing between two treatment groups

Many survival studies are concerned with distinguishing between two alternative treatments. For this reason, a study to compare the survival times of patients who receive a new treatment with those who receive a standard will be used as the focus for this chapter.

Suppose that in this study, there are two groups of patients, and that the standard treatment is allocated to the patients in Group I, while the new treatment is allocated to those in Group II. Assuming a proportional hazards model for the survival times, the hazard of death at time  $t$  for a patient on the new treatment,  $h_N(t)$ , can be written as

$$h_N(t) = \psi h_S(t),$$

where  $h_S(t)$  is the hazard function at  $t$  for a patient on the standard treatment

and  $\psi$  is the unknown hazard ratio. We will also define  $\theta = \log \psi$  to be the log-hazard ratio. If  $\theta$  is zero, there is no treatment difference. On the other hand, negative values of  $\theta$  indicate that survival is longer under the new treatment, while positive values of  $\theta$  indicate that patients survive longer on the standard treatment.

In order to test the null hypothesis that  $\theta = 0$ , the log-rank test described in Section 2.6 can be used. As was shown in Section 3.9, this is equivalent to using the score test of the null hypothesis of equal hazards in the Cox regression model. In this chapter, sample size requirements will be based on the log-rank test statistic, but the results will naturally apply when an analysis based on the Cox regression model is envisaged.

In a survival study, the occurrence of censoring means that it is not usually possible to measure the actual survival times of all patients in the study. However, it is the number of actual deaths that is important in the analysis, rather than the total number of subjects. Accordingly, the first step in determining the number of patients in a study is to calculate the number of deaths that must be observed. We then go on to determine the required number of subjects.

## 10.2 Calculating the required number of deaths

To determine the sample size requirement for a study, we calculate the number of subjects needed for there to be a certain chance of declaring  $\theta$  to be significantly different from zero when the true, but unknown, log-hazard ratio is  $\theta_R$ . Here,  $\theta_R$  is the *reference value* of  $\theta$ . It will be a reflection of the magnitude of the treatment difference that it is important to detect, using the test of significance. In a study to compare a new treatment with a standard, there is likely to be a minimum worthwhile improvement and a maximum envisaged improvement. The actual choice of  $\theta_R$  will then lie between these two values. In practice,  $\theta_R$  might be chosen on the basis of the increase in the median survival time that is to be detected, or in terms of the probability of survival beyond some time. This is discussed further and illustrated later in Example 10.1.

More formally, the required number of deaths is taken to be such that there is a probability of  $1 - \beta$  of declaring the observed log-hazard ratio to be significantly different from zero, using a hypothesis test with a specified significance level of  $\alpha$ , when in fact  $\theta = \theta_R$ . The quantity  $1 - \beta$  is the probability of rejecting the null hypothesis when it is in fact false, and is known as the *power* of the test. The term  $\beta$  is the probability of not rejecting the null hypothesis when it is false and is sometimes known as the *type II error*. Both  $\alpha$  and  $\beta$  are taken to be small. Typical values will be  $\alpha = 0.05$  and  $\beta = 0.1$ , and with these values there would be a 90% chance of declaring the observed difference between two treatments to be significant at the 5% level. The exact specification of  $\alpha$  and  $\beta$  will to some extent depend on the circumstances. If it is important to detect a difference as being significant at a lower level of significance, or if

there needs to be a higher chance of declaring a result to be significant,  $\alpha$  and  $\beta$  will need to be modified accordingly.

The required number of deaths in a survival study,  $d$ , can be obtained from the equation

$$d = \frac{4(z_{\alpha/2} + z_{\beta})^2}{\theta_R^2}, \quad (10.1)$$

where  $z_{\alpha/2}$  and  $z_{\beta}$  are the upper  $\alpha/2$ - and upper  $\beta$ -points, respectively, of the standard normal distribution. It is convenient to write  $c(\alpha, \beta) = (z_{\alpha/2} + z_{\beta})^2$  in equation (10.1), giving

$$d = 4c(\alpha, \beta)/\theta_R^2. \quad (10.2)$$

The values of  $c(\alpha, \beta)$  for commonly chosen values of the significance level  $\alpha$  and power  $1 - \beta$  are given in Table 10.1.

**Table 10.1** Values of the function  $c(\alpha, \beta)$ .

Value of $\alpha$	Value of $1 - \beta$			
	0.80	0.90	0.95	0.99
0.10	6.18	8.56	10.82	15.77
0.05	7.85	10.51	13.00	18.37
0.01	11.68	14.88	17.81	24.03
0.001	17.08	20.90	24.36	31.55

Calculation of the required number of deaths then requires that a value for  $\theta_R$  be identified, and appropriate values of  $\alpha$  and  $\beta$  chosen. Table 10.1 is then used in conjunction with equation (10.2) to give the number of deaths required in a study.

The derivation of the result in equation (10.2) assumes that the same number of individuals is to be assigned to each treatment group. If this is not the case, a modification has to be made. In particular, if the proportion of individuals to be allocated to Group I is  $\pi$ , so that a proportion  $1 - \pi$  will be allocated to Group II, the required total number of deaths becomes

$$d = \frac{c(\alpha, \beta)}{\pi(1 - \pi)\theta_R^2}.$$

Notice that an imbalance in the number of individuals in the two treatment groups leads to an increase in the total number of deaths required. The derivation also includes an approximation, which means that the calculated number of deaths could be an underestimate. Some judicious rounding up of the calculated value is therefore suggested to compensate for this.

The actual derivation of the formula for the required number of deaths is important and so details are given below in Section 10.2.1. This section can be omitted without loss of continuity. It is followed by an example that illustrates the calculations.

### 10.2.1\* Derivation of the required number of deaths

An expression for the required number of deaths is now derived on the basis of a log-rank test to compare two treatment groups. As in Section 2.6, suppose that there are  $r$  distinct death times,  $t_{(1)} < t_{(2)} < \dots < t_{(r)}$ , among the individuals in the study, and that in the  $i$ th group there are  $d_{ij}$  deaths at the  $j$ th ordered death time  $t_{(j)}$ , for  $i = 1, 2$  and  $j = 1, 2, \dots, r$ . Also suppose that the number at risk at  $t_{(j)}$  in the  $i$ th group is  $n_{ij}$ , and write  $n_j = n_{1j} + n_{2j}$  for the total number at risk at  $t_{(j)}$  and  $d_j = d_{1j} + d_{2j}$  for the number who die at  $t_{(j)}$ . The log-rank statistic is then

$$U = \sum_{j=1}^r (d_{1j} - e_{1j}),$$

where  $e_{1j}$  is the expected number of deaths in Group I at  $t_{(j)}$ , given by  $e_{1j} = n_{1j}d_j/n_j$ , and the variance of the log-rank statistic is

$$V = \sum_{j=1}^r \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)}. \quad (10.3)$$

When using the log-rank test, the null hypothesis that  $\theta = 0$  is rejected if the absolute value of  $U$  is sufficiently large, that is, if  $|U| > k$ , say, where  $k > 0$  is a constant. We therefore require that

$$P(|U| > k; \theta = 0) = \alpha, \quad (10.4)$$

and

$$P(|U| > k; \theta = \theta_R) = 1 - \beta. \quad (10.5)$$

for a two-sided  $100\alpha\%$  significance test to have a power  $1 - \beta$ .

We now quote without proof a result given in Sellke and Siegmund (1983), according to which the log-rank statistic,  $U$ , has an approximate normal distribution with mean  $\theta V$  and variance  $V$ , for small values of  $\theta$ . Indeed, the result that  $U \sim N(0, V)$  under the null hypothesis  $\theta = 0$ , is used as a basis for the log-rank test. Then, since

$$P(|U| > k; \theta = 0) = P(U > k; \theta = 0) + P(U < -k; \theta = 0),$$

and  $U$  has an  $N(0, V)$  distribution when  $\theta = 0$ , a distribution that is symmetric about zero,

$$P(U > k; \theta = 0) = P(U < -k; \theta = 0).$$

It then follows from equation (10.4) that

$$P(U > k; \theta = 0) = \frac{\alpha}{2}. \quad (10.6)$$

Next, we note that

$$P(|U| > k; \theta = \theta_R) = P(U > k; \theta = \theta_R) + P(U < -k; \theta = \theta_R).$$

For the sort of values of  $k$  that are likely to be used in the hypothesis test, either  $P(U < -k; \theta = \theta_R)$  or  $P(U > k; \theta = \theta_R)$  will be negligible. For example,

if the new treatment is expected to increase survival so that  $\theta_R$  is taken to be less than zero, the probability of  $U$  having a value in excess of  $k$ ,  $k > 0$ , will be small. So without loss of generality we will take

$$P(|U| > k; \theta = \theta_R) \approx P(U < -k; \theta = \theta_R).$$

We now denote the upper  $100p\%$  point of the standard normal distribution by  $z_p$ . Then  $\Phi(z_p) = 1 - p$ , where  $\Phi(\cdot)$  stands for the standard normal distribution function. The quantity  $\Phi(z_p)$  therefore represents the area under a standard normal density function to the left of the value  $z_p$ . Now, since  $U \sim N(0, V)$  when  $\theta = 0$ ,

$$P(U > k; \theta = 0) = 1 - P(U \leq k; \theta = 0) = 1 - \Phi\left(\frac{k}{\sqrt(V)}\right),$$

and using equation (10.6) we have that

$$\Phi\left(\frac{k}{\sqrt(V)}\right) = 1 - (\alpha/2).$$

Therefore,

$$\frac{k}{\sqrt(V)} = z_{\alpha/2},$$

where  $z_{\alpha/2}$  is the upper  $\alpha/2$ -point of the standard normal distribution, and so  $k$  can be expressed as

$$k = z_{\alpha/2}\sqrt(V). \quad (10.7)$$

In a similar manner, since  $U \sim N(\theta_R V, V)$  when  $\theta = \theta_R$ ,

$$P(U < -k; \theta = \theta_R) = \Phi\left(\frac{-k - \theta_R V}{\sqrt(V)}\right) \approx 1 - \beta,$$

and so we take

$$\frac{-k - \theta_R V}{\sqrt(V)} = z_\beta,$$

where  $z_\beta$  is the upper  $\beta$ -point of the standard normal distribution. If we now substitute for  $k$  from equation (10.7), we get

$$-z_{\alpha/2}\sqrt(V) - \theta_R V = z_\beta\sqrt(V),$$

and so  $V$  needs to be such that

$$V = (z_{\alpha/2} + z_\beta)^2 / \theta_R^2, \quad (10.8)$$

to meet the specified requirements.

When the number of deaths is few relative to the number at risk, the expression for  $V$  in equation (10.3) is approximately

$$\sum_{j=1}^r \frac{n_{1j}n_{2j}d_j}{n_j^2}. \quad (10.9)$$

Moreover, if  $\theta$  is small, and recruitment to each treatment group proceeds at

a similar rate, then  $n_{1j} \approx n_{2j}$ , for  $j = 1, 2, \dots, r$ , and so

$$\frac{n_{1j}n_{2j}}{n_j^2} = \frac{n_{1j}n_{2j}}{(n_{1j} + n_{2j})^2} \approx \frac{n_{1j}^2}{(2n_{1j})^2} = \frac{1}{4}.$$

Then,  $V$  is given by

$$V \approx \sum_{j=1}^r d_j/4 = d/4,$$

say, where  $d = \sum_{j=1}^r d_j$  is the total number of deaths among the individuals in the study.

Finally, using equation (10.8), we now require  $d$  to be such that

$$\frac{d}{4} = \frac{(z_{\alpha/2} + z_{\beta})^2}{\theta_R^2},$$

which leads to the required number of deaths being that given in equation (10.1).

At later death times, that is, when the values of  $j$  in expression (10.9) are close to  $r$ , the numbers of subjects at risk in the two groups will be small. This is likely to mean that  $n_{1j}$  and  $n_{2j}$  will be quite different at the later death times, and so  $n_{1j}n_{2j}/n_j^2$  will be less than 0.25. This in turn means that  $V < d/4$  and so the required number of deaths will tend to be underestimated.

#### *Example 10.1 Survival from chronic active hepatitis*

Patients suffering from chronic active hepatitis rapidly progress to an early death from liver failure. A new treatment has become available and so a clinical trial is planned to evaluate the effect of this new treatment on the survival times of patients suffering from the disease. As a first step, information is obtained on the survival times in years of patients in a similar age range who have received the standard therapy. The Kaplan-Meier estimate of the survivor function derived from such data is shown in Figure 10.1.

From this estimate of the survivor function, the median survival time is 3.3 years, and the survival rates at two, four and six years can be taken to be given by  $S(2) = 0.70$ ,  $S(4) = 0.45$ , and  $S(6) = 0.25$ .

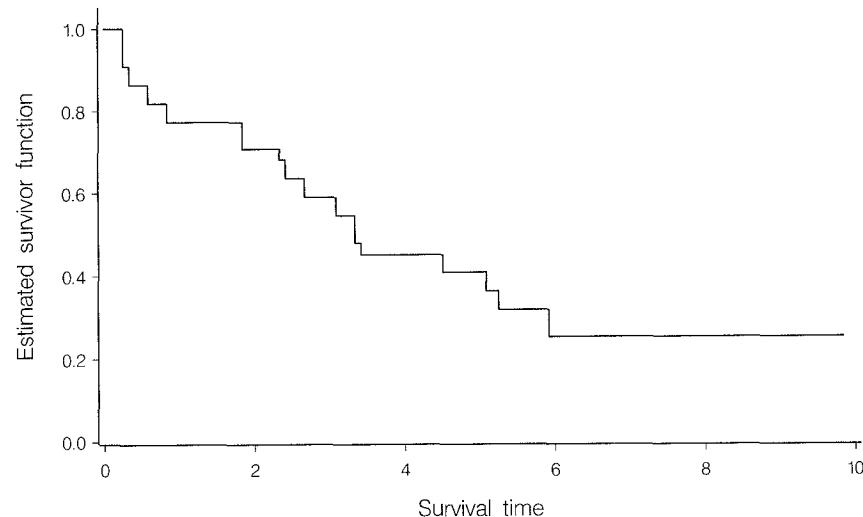
The new treatment is expected to increase the survival rate at five years from 0.41, the value under the standard treatment, to 0.60. This information can be used to calculate a value for  $\theta_R$ . To do this, we use the result that if the hazard functions are assumed to be proportional, the survivor function for an individual on the new treatment at time  $t$  is

$$S_N(t) = [S_S(t)]^\psi, \quad (10.10)$$

where  $S_S(t)$  is the survivor function for an individual on the standard treatment at  $t$  and  $\psi$  is the hazard ratio. Therefore,

$$\psi = \frac{\log S_N(t)}{\log S_S(t)},$$

and so the value of  $\psi$  corresponding to an increase in  $S(t)$  from 0.41 to 0.60



**Figure 10.1** Estimated survivor function for patients receiving a standard treatment for hepatitis.

is

$$\psi_R = \frac{\log(0.60)}{\log(0.41)} = 0.57.$$

With this information, the survivor function for a patient on the new treatment can be estimated by  $[S_S(t)]^{\psi_R}$ . In particular,  $S_N(2) = 0.82$ ,  $S_N(4) = 0.63$ , and  $S_N(6) = 0.45$ . A plot of the two survivor functions is shown in Figure 10.2.

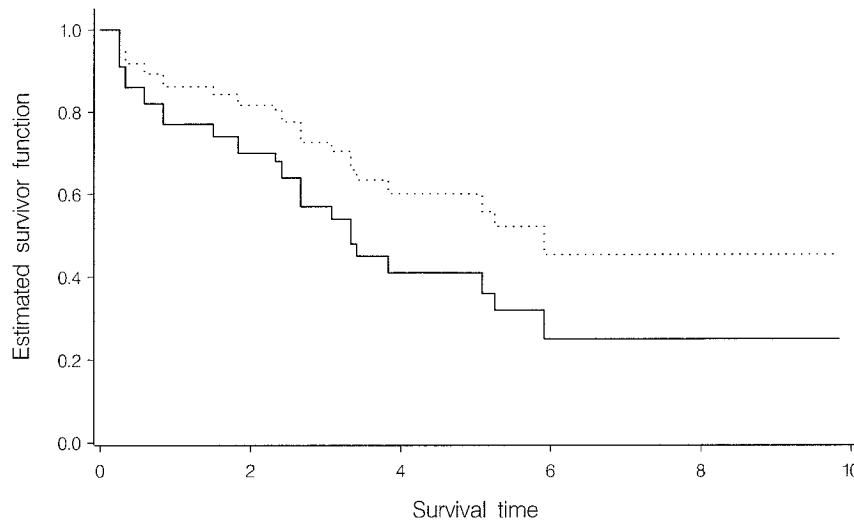
The median survival time under the new treatment can be estimated from this estimate of the survivor function. Using Figure 10.2, the median survival time under the new treatment is estimated to be about six years.

To calculate the number of deaths that would be required in a study to compare the two treatments, we will take  $\alpha = 0.05$  and  $1 - \beta = 0.90$ . With these values of  $\alpha$  and  $\beta$ , the value of the function  $c(\alpha, \beta)$  from Table 10.1 is 10.51. Substituting for  $c(0.05, 0.1)$  in equation (10.2) and taking  $\theta_R = \log \psi_R = \log(0.57) = -0.562$ , the number of deaths required to have a 90% chance of detecting a hazard ratio of 0.57 to be significant at the 5% level is then given by

$$d = \frac{4 \times 10.51}{0.562^2} = 133.$$

Allowing for possible underestimation, this can be rounded up to 140 deaths in total. This means that approximately 70 deaths would need to be observed in each treatment group.

The calculations described above are only going to be of direct use when a study is to be continued until a given number of those entering the study have



**Figure 10.2** Estimated survivor functions for individuals on the standard treatment (—) and the new treatment (···).

died. Most trials will be designed on the basis of the number of patients to be recruited and so we must now examine how this number can be calculated.

### 10.3 Calculating the required number of patients

In order to calculate the actual number of patients that are required in a survival study, we need to consider the probability of death over the duration of a study. Typically, patients are recruited over an *accrual period* of length  $a$ . After recruitment is complete, there is an additional *follow-up period* of length  $f$ . The total duration of a study will therefore be of length  $a + f$ . Notice that if  $f$  is small, or even zero, there will need to be correspondingly more patients recruited in order to achieve a specific number of deaths.

Once the probability of a patient dying in the study has been evaluated, the required number of patients will be found from

$$n = \frac{d}{P(\text{death})}, \quad (10.11)$$

where  $d$  is the required number of deaths found from equation (10.2). According to a result derived in the next section, the probability of death can be taken as

$$P(\text{death}) = 1 - \frac{1}{6}\{\bar{S}(f) + 4\bar{S}(0.5a + f) + \bar{S}(a + f)\}. \quad (10.12)$$

where

$$\bar{S}(t) = \frac{S_S(t) + S_N(t)}{2},$$

and  $S_S(t)$  and  $S_N(t)$  are the estimated values of the survivor functions for individuals on the standard and new treatments, respectively, at time  $t$ .

The above result shows how the required number of patients can be calculated for a trial with an accrual period of  $a$  and a follow-up period of  $f$ . Of course, the duration of the accrual period and follow-up period will depend on the recruitment rate. So suppose that the recruitment rate is expected to be  $m$  patients per month and that  $d$  deaths are required. If  $n$  patients are to be entered into the study over a period of  $a$  months, this means that  $n/a$  need to be recruited in each month. In practice, information is likely to be available on the accrual rate,  $m$ , that can be expected. The number recruited in an accrual period of length  $a$  is then  $ma$  and so the expected number of deaths in the study is

$$ma \times P(\text{death}).$$

Values of  $a$  and  $f$  which make this value close to the number of deaths required can then be found numerically, for example, by trying out different values of  $a$  and  $f$ . This algorithm could be computerised and an optimisation method used to find the value of  $a$  that makes

$$d - \{ma \times P(\text{death})\} \quad (10.13)$$

close to zero for a range of values of  $f$ . Alternatively, the value of  $f$  that yields the result in equation (10.13) for a range of values of  $a$  can be found. A two-way table giving the required number of patients for different combinations of values of  $a$  and  $f$  will be particularly useful in planning a study.

The following section gives details underlying the derivation of the result in equation (10.12), and can again be omitted without loss of continuity.

#### 10.3.1\* Derivation of the required number of patients

We begin with the general result from distribution theory that the marginal probability of a patient dying during the course of a study can be obtained from the joint probability of death and entry to the study at time  $t$  using

$$P(\text{death}) = \int_0^a P(\text{death and entry at time } t) dt. \quad (10.14)$$

The joint probability can in turn be found from the result

$$P(\text{death and entry at time } t) = P(\text{death} | \text{entry at } t) \times P(\text{entry at } t), \quad (10.15)$$

which is simply a version of the result that  $P(A|B) = P(AB)/P(B)$ .

We now assume a uniform recruitment rate over the accrual period. The distribution of entry times to the study can then be taken to be uniform over the time interval  $(0, a)$ . Therefore, the probability of an individual being recruited to the study at time  $t$  is  $a^{-1}$ , for any value of  $t$  in the interval  $(0, a)$ . From equations (10.14) and (10.15), we have

$$P(\text{death}) = \int_0^a P(\text{death} | \text{entry at } t) a^{-1} dt,$$

so that

$$P(\text{death}) = 1 - \frac{1}{a} \int_0^a P(\text{survival} \mid \text{entry at } t) dt.$$

A patient entering the study at time  $t$  who survives for the duration of the study, that is, to time  $a+f$ , must have been alive for a period of length  $a+f-t$  after entry. The conditional probability  $P(\text{survival} \mid \text{entry at } t)$  is therefore the probability of survival beyond  $a+f-t$ . This probability is the value of the survivor function for that individual at  $a+f-t$ , that is,  $S(a+f-t)$ . Consequently,

$$P(\text{death}) = 1 - \frac{1}{a} \int_0^a S(a+f-t) dt,$$

and on writing  $u = a+f-t$ , this result becomes

$$P(\text{death}) = 1 - \frac{1}{a} \int_f^{a+f} S(u) du. \quad (10.16)$$

The integral of the survivor function is now approximated using numerical integration. According to Simpson's rule,

$$\int_u^v f(x) dx \approx \frac{v-u}{6} \left\{ f(u) + 4f\left(\frac{u+v}{2}\right) + f(v) \right\},$$

so that

$$\int_f^{a+f} S(u) du \approx \frac{a}{6} \{S(f) + 4S(0.5a+f) + S(a+f)\},$$

and hence, using equation (10.16), the probability of death during the study is given by

$$P(\text{death}) = 1 - \frac{1}{6} \{S(f) + 4S(0.5a+f) + S(a+f)\}.$$

From this result, the approximate probability of death for an individual in Group I, for whom the survivor function is  $S_S(t)$ , is

$$P(\text{death; Group I}) = 1 - \frac{1}{6} \{S_S(f) + 4S_S(0.5a+f) + S_S(a+f)\},$$

and similarly that for an individual in Group II is

$$P(\text{death; Group II}) = 1 - \frac{1}{6} \{S_N(f) + 4S_N(0.5a+f) + S_N(a+f)\}.$$

On the assumption that there is an equal probability of an individual being assigned to either of the two treatment groups, the overall probability of death is the average of these two probabilities, so that

$$P(\text{death}) = \frac{P(\text{death; Group I}) + P(\text{death; Group II})}{2}.$$

On substituting for the probabilities of death in the two treatment groups, and writing  $\bar{S}(t) = \{S_S(t) + S_N(t)\}/2$ , we get

$$P(\text{death}) = 1 - \frac{1}{6} \{\bar{S}(f) + 4\bar{S}(0.5a+f) + \bar{S}(a+f)\},$$

as in equation (10.12).

If the proportion of individuals to be allocated to Group I is  $\pi$ , the overall probability of death becomes

$$\pi P(\text{death; Group I}) + (1-\pi) P(\text{death; Group II}),$$

and the result for the overall probability of death is modified accordingly.

#### *Example 10.2 Survival from chronic active hepatitis*

In Example 10.1, it was shown that 140 deaths needed to be observed for the study on chronic hepatitis to have sufficient power to detect a hazard ratio of 0.57 as significant. Suppose that patients are to be recruited to the study over an 18-month accrual period and that there is to be a subsequent follow-up period of 24 months. From equation (10.12), the probability of death in the 42 months of the study will then be given by

$$P(\text{death}) = 1 - \frac{1}{6} \{\bar{S}(24) + 4\bar{S}(33) + \bar{S}(42)\}.$$

Now, using the estimated survivor functions shown in Figure 10.2,

$$\bar{S}(24) = \frac{S_S(24) + S_N(24)}{2} = \frac{0.70 + 0.82}{2} = 0.76,$$

$$\bar{S}(33) = \frac{S_S(33) + S_N(33)}{2} = \frac{0.57 + 0.73}{2} = 0.65,$$

$$\bar{S}(42) = \frac{S_S(42) + S_N(42)}{2} = \frac{0.45 + 0.63}{2} = 0.54,$$

and so the probability of death is

$$1 - \frac{1}{6} \{0.76 + (4 \times 0.65) + 0.54\} = 0.350.$$

From equation (10.11), the required number of patients is

$$n = \frac{140}{0.350} = 400,$$

and so 400 patients will need to be recruited to the study over the accrual period of 18 months. This demands a recruitment rate of about 22 patients per month.

If it is only expected that 18 patients can be found each month, the accrual period will need to be extended to ensure that there is a sufficient number of individuals to give the required number of deaths. The number of individuals that could be recruited in a period of  $a$  months would be  $18a$ . Various values of  $a$  can then be tried in order to make this approximately equal to the value obtained from equation (10.11). For example, if we take  $a = 24$  and continue with  $f = 24$ , the probability of death over the four years of the study is

$$P(\text{death}) = 1 - \frac{1}{6} \{\bar{S}(24) + 4\bar{S}(36) + \bar{S}(48)\}.$$

From Figure 10.2, the survivor functions for patients on each treatment at 24, 36 and 48 months can be estimated, and we find that  $\bar{S}(24) = 0.76$ ,

$\bar{S}(36) = 0.65$ , and  $\bar{S}(48) = 0.50$ . The probability of death then turns out to be 0.357 and the required number of patients to give 140 deaths is now 393. This is broadly consistent with an estimated recruitment rate of 18 per month.

Now suppose that it is decided that the study will not have a follow-up period, so that the accrual period is equal to the duration of the study. If the accrual period is taken to be 20 months, so that  $a = 20$  and  $f = 0$ , the probability of death is given by

$$P(\text{death}) = 1 - \frac{1}{6}\{\bar{S}(0) + 4\bar{S}(10) + \bar{S}(20)\}.$$

Now,  $\bar{S}(0) = 1.00$ ,  $\bar{S}(10) = 0.82$ , and  $\bar{S}(20) = 0.79$ , and the probability of death is 0.155. The required number of patients is now  $140/0.155 = 903$ , and this would just about be met by a recruitment rate of 45 patients per month. This shows that the absence of a follow-up period leads to an increase in the number of patients that must be entered into the study.

### 10.3.2 An approximate procedure

A much simpler procedure for calculating the required number of patients in a survival study to compare two treatments is outlined in this section. The basis for this result is that  $\{S_S(\tau) + S_N(\tau)\}/2$  is the average probability that a patient in the study survives beyond time  $\tau$ , where  $S_S(\tau)$  and  $S_N(\tau)$  are the survivor functions at time  $\tau$ , for patients on the standard and new treatments, respectively. The probability of death, in the period from the time origin to  $\tau$ , can then be approximated by

$$1 - \frac{S_S(\tau) + S_N(\tau)}{2}.$$

Using equation (10.11), the required number of patients becomes

$$n = \frac{2d}{2 - S_S(\tau) - S_N(\tau)},$$

where  $d$  is the required number of deaths.

A natural choice for the value of  $\tau$  to use in this calculation is the average length of the follow-up period for patients in the study,  $f + (a/2)$ , where  $a$  is the accrual period and  $f$  the follow-up period. This procedure is very approximate because it does not take account of patient follow-up times that extend beyond  $\tau$ . As a consequence, this result will tend to overestimate the required sample size.

#### Example 10.3 Survival from chronic active hepatitis

The study on chronic active hepatitis is now used to illustrate the approximate procedure for determining the required number of patients. As in Example 10.2, suppose that patients are recruited over an 18 month period and that there is a further follow-up period of 24 months. The average length of the follow-up period for a patient will then be  $\tau = f + (a/2) = 33$  months. From Figure 10.2, the survivor functions for patients in the two groups are given by

$S_S(33) = 0.57$  and  $S_N(33) = 0.73$ , respectively. The approximate probability of death in the period from 0 to 33 months is therefore 0.35 and the number of patients required to give 140 deaths is  $140/0.35 = 400$ . In this illustration, the sample size produced by the approximate result is identical to that found using the more complicated procedure illustrated in Example 10.2, but this will not usually be the case.

### 10.4 Further reading

Full details on the issues to be considered when designing a clinical trial are given in Pocock (1983) and Altman (1991). Whitehead (1997) gives an authoritative account of sequential methodology for survival trials. The computer package PEST4 (Planning and Evaluation of Sequential Trials), documented in MPS Research Unit (2000), includes facilities for the design and analysis of sequential clinical trials in which the response variable is a survival time.

The formula for the required number of deaths in equation (10.1) appears in many papers, including Bernstein and Lagakos (1978), Schoenfeld (1981), Schoenfeld and Richter (1982) and Schoenfeld (1983), although the assumptions on which the result is based are different. Bernstein and Lagakos (1978) obtain equation (10.1) on the assumption that the survival times in each group have exponential distributions. Lachin (1981), Rubinstein, Gail and Santner (1981) and Lachin and Foulkes (1986) also discuss sample size requirements in trials where the survival times are assumed to be exponentially distributed. See also the earlier work of George and Desu (1974).

Schoenfeld (1981) obtains the same result as Bernstein and Lagakos (1978) and others when the log-rank test is used to compare treatments, without making the assumption of exponentiality. Schoenfeld (1983) shows that equation (10.1) holds when information on the values of explanatory variables is allowed for.

Schoenfeld and Richter (1982) give nomograms that enable the required numbers of patients to be determined on the assumption of exponential survival times. Dupont and Plummer (1990), in their review of sample size formulae, only give results that assume exponential survival times, although they do describe a public domain computer program that can be used to determine sample size requirements. This is easier to use than Schoenfeld and Richter's nomograms!

The formulae for the required number of patients in Section 10.3.1 are based on Schoenfeld (1983). When the assumption of exponential survival times is made, these formulae do simplify to the results of Schoenfeld and Richter (1982). Although the resulting formulae are easier to use, it is dangerous to conduct sample size calculations on the basis of restrictive assumptions about survival time distributions.

A variant on the formula for the required number of deaths is given by Freedman (1982). Freedman's result has  $\{(1 + \psi)/(1 - \psi)\}^2$  in place of  $4/(\log \psi)^2$

in equation (10.1). However, for small values of  $\log \psi$ ,

$$\{(1 + \psi)/(1 - \psi)\}^2 \approx 4/(\log \psi)^2,$$

and so the two expressions will tend to give similar results. The approximate formula for the required number of patients, given in Section 10.3.2, is also due to Freedman (1982).

Lakatos (1988) presented a method for estimating the required number of patients to compare two treatments which can accommodate matters such as staggered entry, non-compliance, loss to follow-up and non-proportional hazards. Lakatos and Lan (1992) show that the Lakatos method procedure performs well in a variety of circumstances. This approach is based on a Markov model, and requires a computer program for its implementation; a SAS macro has been given by Shih (1995).

Extensive tables of sample size requirements in studies involving different types of response variable, including survival times, are provided by Machin *et al.* (1997). This book includes computer software designed to simplify the use of the tables. A number of commercially available software packages for sample size calculations, including PASS and nQuery Advisor, also implement methods for calculating the required number of patients in a survival study.

## CHAPTER 11

# Some additional topics

In this chapter, a number of additional topics that arise in the practical application of the methods of survival analysis described in this book are discussed. Many analyses are based on a model that assumes proportional hazards, and so some methods for processing survival data when this assumption is not satisfied are described in Section 11.1. In Section 11.2, the implications of informative censoring are outlined, and methods for detecting this feature are suggested. Models for survival data that contain a random effect, known as frailty models, are introduced in Section 11.3. Multistate models, used in modelling the natural progression of a disease, are introduced in Section 11.4. A number of other issues in survival analysis, namely the effect of adjusting for explanatory variables, measures of explained variation, modelling a cure probability, cross-over and sequential designs, are considered in less detail in subsequent sections. For each of these topics, references to the literature are given to enable the interested reader to obtain further information on them.

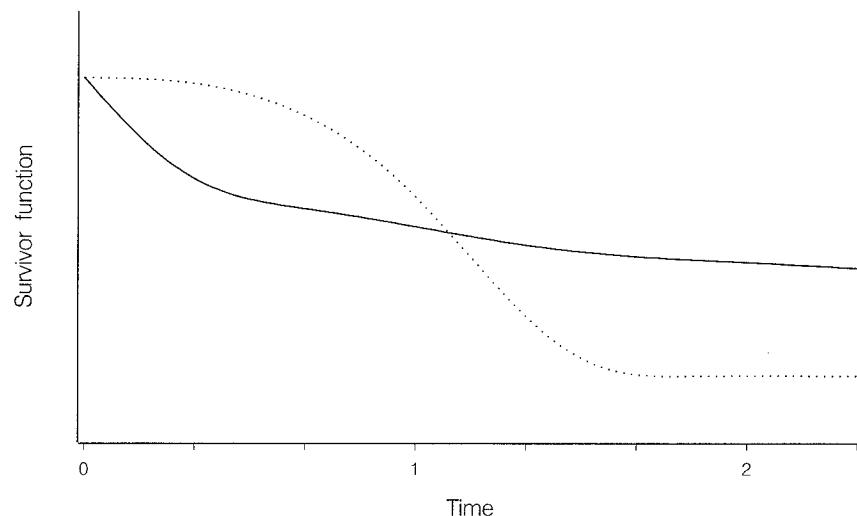
### 11.1 Non-proportional hazards

Some models that do not require the assumption of proportional hazards have been considered in previous chapters. These include the accelerated failure time model and the proportional odds model introduced in Chapter 6, and the Cox regression model that includes a time-dependent variable, described in Chapter 8. But often we are faced with a situation where the assumption of proportional hazards is untenable, and yet none of the above models is satisfactory.

As an illustration, consider a study to compare a surgical procedure with chemotherapy in the treatment of a particular form of cancer. Suppose that the survivor functions under the two treatments are as shown in Figure 11.1, where the time scale is in years. Clearly the hazards are non-proportional. Death at an early stage may be experienced by patients on the surgical treatment, as a result of patients not being able to withstand the surgery or complications arising from it. In the longer term, patients who have recovered from the surgery have a better prognosis.

A similar situation arises when an aggressive form of chemotherapy is compared to a standard. Here also, a long-term advantage to the aggressive treatment may be at the expense of short-term excess mortality.

One approach, which is useful in the analysis of data arising from situations such as these, is to define the end-point of the study to be survival beyond

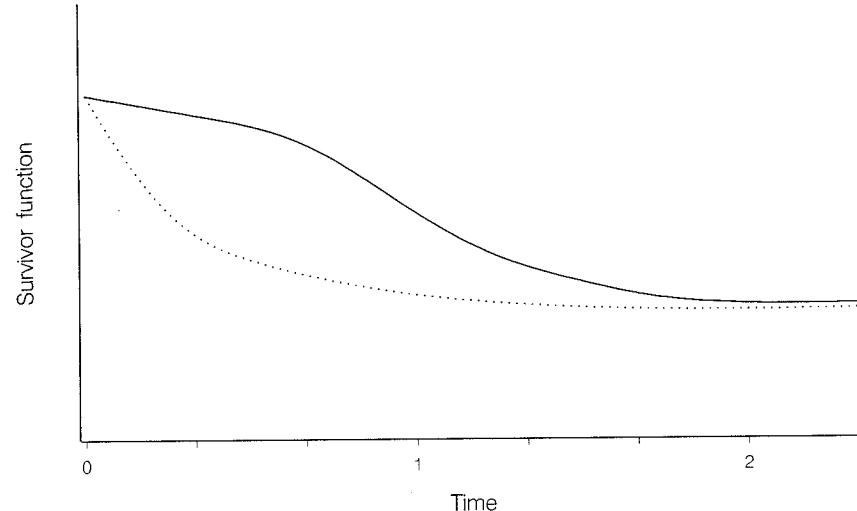


**Figure 11.1** Long-term advantage of surgery (—) over chemotherapy (···).

some particular time. For example, in the study leading to the survivor functions illustrated in Figure 11.1, the treatment difference is roughly constant after two years. The dependence of the probability of survival beyond two years on prognostic variables and treatment might therefore be modelled. This approach was discussed in connection with the analysis of interval-censored survival data in Section 9.2. As shown in that section, there are advantages in using a linear model for the complementary log-log transformation of the survival probability. In particular, the coefficients of the explanatory variables in the linear component of the model can be interpreted as logarithms of hazard ratios. The disadvantages of this approach are that all patients must be followed until the point in time when the survival rates are to be analysed, and that the death data cannot be used until this time. Moreover, faith in the long-term benefits of one or other of the two treatments will be needed to ensure that the trial is not stopped early because of excess mortality in one treatment group.

Strictly speaking, an analysis based on the survival probability at a particular time is only valid when that time is specified at the outset of the study. In other words, this end-point needs to be defined in the study protocol. This may be difficult to do. If the data are used to suggest end-points such as the probability of survival beyond two years, some caution will be needed in interpreting the results of a significance test.

In the study that leads to the survivor functions shown in Figure 11.1, it is clear that an analysis of the two-year survival rate will be appropriate. Now consider a study to compare the use of chemotherapy in addition to surgery with surgery alone, in which the survivor functions are as shown in Figure 11.2. Here, the short-term benefit of the chemotherapy may certainly be



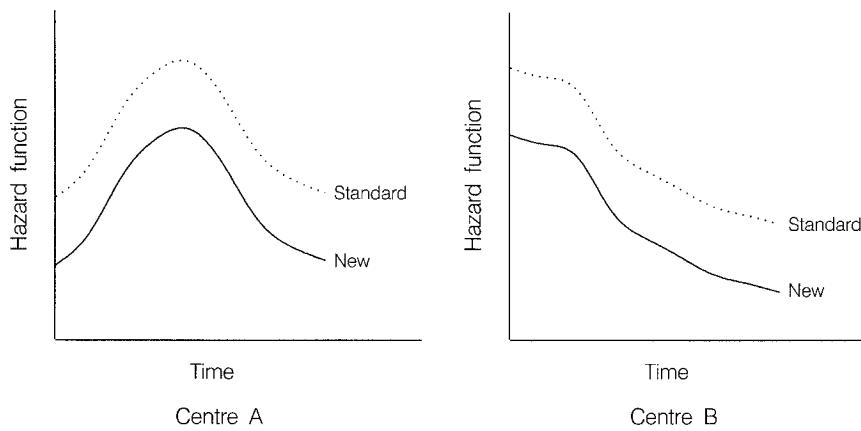
**Figure 11.2** Short-term advantage of chemotherapy and surgery (—) over surgery alone (···).

worthwhile, but an analysis of the two-year survival rates will fail to establish a treatment difference. The fact that the difference between the two survival rates is not constant makes it difficult to use an analysis based on survival rates at a given time. However, it might be reasonable to assume that the hazards are proportional over the first year of the study, and to carry out a survival analysis at that time.

### 11.1.1 Stratified proportional hazards models

A situation that sometimes occurs is that hazards are not proportional on an overall basis, but that they are proportional in different subgroups of the data. For example, consider a situation in which a new drug is being compared with a standard in the treatment of a particular disease. If the study involves two participating centres, it is possible that in each centre the new treatment halves the hazard of death, but that the hazard functions for the standard drug are different between the subjects recruited by each centre. Then, the hazards between centres for individuals on a given drug are not proportional. This situation is illustrated in Figure 11.3.

In problems of this kind, it may be assumed that patients in each of the subgroups, or *strata*, have a different baseline hazard function, but that all other explanatory variables satisfy the proportional hazards assumption within each stratum. Suppose that the patients in the  $j$ th stratum have a baseline hazard function  $h_{0j}(t)$ , for  $j = 1, 2, \dots, g$ , where  $g$  is the number of strata. The effect of explanatory variables on the hazard function can then be represented by a proportional hazards model for  $h_{ij}(t)$ , the hazard function for the  $i$ th



**Figure 11.3** Hazard functions for individuals on a new drug (—) and a standard drug (···) in two centres.

individual in the  $j$ th stratum, where  $i = 1, 2, \dots, n_j$ , say, and  $n_j$  is the number of individuals in the  $j$ th stratum. We then have the *stratified proportional hazards model*, according to which

$$h_{ij}(t) = \exp(\beta' \mathbf{x}_{ij}) h_{0j}(t),$$

where  $\mathbf{x}_{ij}$  is the vector of values of  $p$  explanatory variables,  $X_1, X_2, \dots, X_p$ , recorded on the  $i$ th individual in the  $j$ th stratum.

As an example of this model, consider the particular case where there are two treatments being compared in each of  $g$  strata, and no other explanatory variables. Let  $x_{ij}$  be the value of an indicator  $X$ , which is zero if the  $i$ th subject in the  $j$ th stratum is on the standard treatment and unity if on the new treatment. The hazard function for this individual is then

$$h_{ij}(t) = e^{\beta x_{ij}} h_{0j}(t).$$

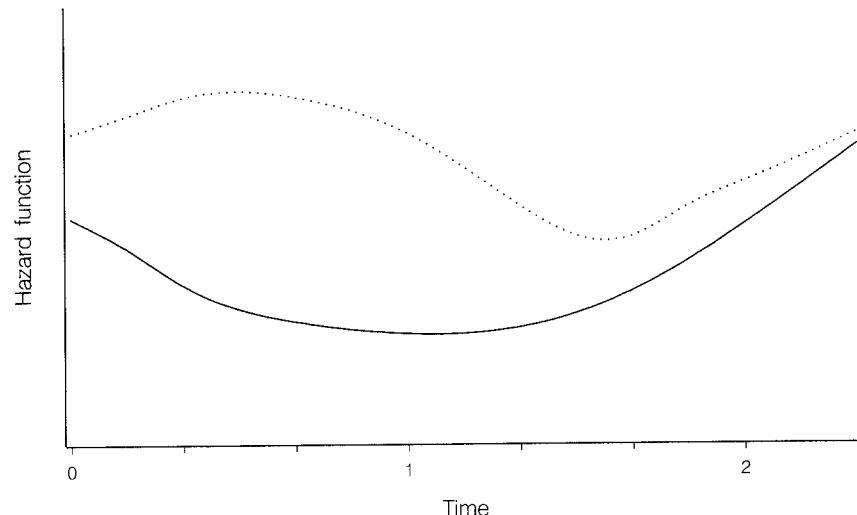
On fitting this model, the estimated value of  $\beta$  is the log-hazard ratio for an individual on the new treatment, relative to one on the standard, in each stratum.

This model for stratified proportional hazards is easily fitted using major software packages for survival analysis. Models can be compared using the  $-2 \log \hat{L}$  statistic and no new principles are involved. When two or more groups of survival data are being compared, the stratified proportional hazards model is in fact equivalent to the stratified log-rank test described in Section 2.8 of Chapter 2.

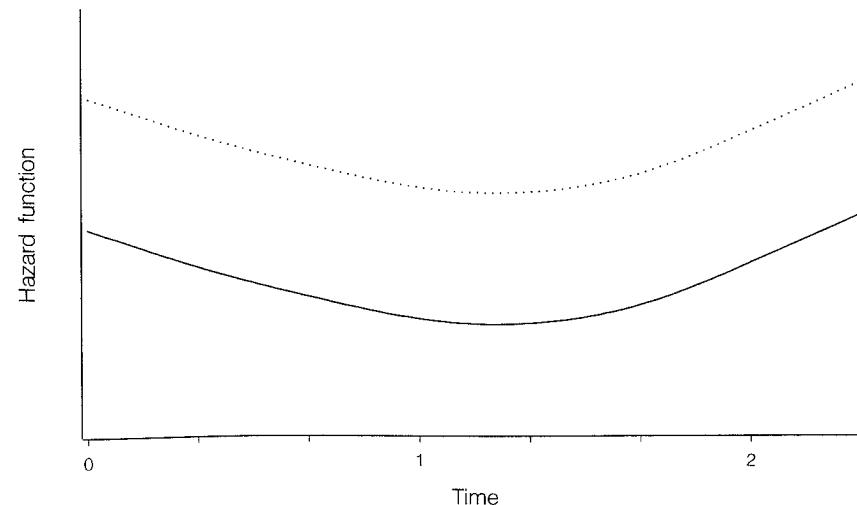
#### 11.1.2 Non-proportional hazards between treatments

If there are non-proportional hazards between two treatments, misleading inferences can result from ignoring this phenomenon. To illustrate this point,

suppose that the hazard function for two groups of individuals, on a new and standard treatment, are as shown in Figure 11.4. If a proportional hazards model were fitted, the resulting fitted hazard functions are likely to be as shown in Figure 11.5. Incorrect conclusions would then be drawn about the relative merit of the two treatments.



**Figure 11.4** Non-proportional hazards for individuals on a new treatment (—) and a standard treatment (···).



**Figure 11.5** The result of fitting a proportional hazards model to the hazard functions in Figure 11.4.

Non-proportional hazards between treatments can be modelled assuming proportional hazards in a series of consecutive time intervals. This is achieved using a *piecewise Cox model*, which is analogous to the piecewise exponential model introduced in Chapter 6. To illustrate the use of the model, suppose that the time period over which the hazard functions in Figure 11.4 are given is divided into three intervals, namely  $(0, t_1)$ ,  $(t_1, t_2)$  and  $(t_2, t_3)$ . Within each of these intervals, hazards might be assumed to be proportional.

Now let  $X$  be an indicator variable associated with the two treatments, where  $X = 0$  if an individual is on the standard treatment and  $X = 1$  if an individual is on the new treatment. The piecewise Cox regression model can then be fitted by defining two time-dependent variables,  $Z_2(t)$  and  $Z_3(t)$ , say, which are as follows:

$$Z_2(t) = \begin{cases} 1 & \text{if } t \in (t_1, t_2) \text{ and } X = 1, \\ 0 & \text{otherwise;} \end{cases}$$

$$Z_3(t) = \begin{cases} 1 & \text{if } t \in (t_2, t_3) \text{ and } X = 1, \\ 0 & \text{otherwise.} \end{cases}$$

In the absence of other explanatory variables, the model for the hazard function for the  $i$ th individual at  $t$  can be written as

$$h_i(t) = \exp\{\beta_1 x_i + \beta_2 z_{2i}(t) + \beta_3 z_{3i}(t)\} h_0(t)$$

where  $x_i$  is the value of  $X$  for the  $i$ th individual, and  $z_{2i}(t)$  and  $z_{3i}(t)$  are the values of the two time-dependent variables for the  $i$ th individual at  $t$ . Under this model, the log-hazard ratio for an individual on the new treatment, relative to one on the standard, is then  $\beta_1$  for  $t \in (0, t_1)$ ,  $\beta_1 + \beta_2$  for  $t \in (t_1, t_2)$  and  $\beta_1 + \beta_3$  for  $t \in (t_2, t_3)$ . This model can be fitted in the manner described in Chapter 8.

### 11.1.3 Further reading

Examples of survival analyses in situations where the proportional hazards model is not applicable have been given by Stablein *et al.* (1981) and Gore *et al.* (1984). Further details on the stratified proportional hazards model can be found in Kalbfleisch and Prentice (2002) and Lawless (2002), for example. A review of methods for dealing with non-proportional hazards in the Cox regression model is included in Schemper (1992). A discussion of strategies for dealing with non-proportional hazards is included in Chapter 6 of Therneau and Grambsch (2000).

## 11.2 Informative censoring

The methods described in this book for the analysis of censored survival data are only valid if the censoring is non-informative. Essentially, this means that the censoring is not related to any factors associated with the actual survival time, as pointed out in Section 1.1 of Chapter 1.

As an example of a situation where censoring is informative, suppose that

individuals are withdrawn from a survival study to compare two treatments because they experience life-threatening side effects from one particular treatment. The effect of this will be that the survival rates observed on this particular treatment will appear greater than they should be, leading to incorrect conclusions about the extent of the treatment difference.

The assumption of uninformative censoring can be examined in a number of ways. One possibility is to plot observed survival times against the values of explanatory variables, where the censored observations are distinguished from the uncensored. If a pattern is exhibited in the censoring, such as there being more censored observations at an earlier time on one treatment than on the other, or if there is a greater proportion of censored survival times in patients with a particular range of values of explanatory variables, informative censoring is suggested.

One approach to informative censoring is to examine the sensitivity of inferences of primary interest to the possibility that censored observations are informative. This is done by carrying out two further analyses. In the first, we suppose that individuals who contribute censored observations are actually those at high risk of an event. We therefore assume that individuals for whom the survival time is censored experience the event immediately after censoring, and replace all censored observations by events that occur at the time of censoring. In the second analysis, we suppose that the censored individuals are those at low risk of an event. We then assume that those individuals with censored observations experience the event only after the longest survival time among all individuals in the data set. The censored times are therefore replaced by the longest event time. The impact of these two assumptions on the results of the analysis can then be studied in detail. If essentially the same conclusions arise from the original analysis and the two supplementary analyses, it will be safe to assume that the results are not sensitive to the potential presence of informative censoring.

More formally, a model could be used to examine whether the probability of censoring was related to the explanatory variables in the model. In particular, a linear logistic model could be used in modelling a binary response variable that takes the value unity if an observed survival time is censored and zero otherwise. If particular explanatory variables in the data set lead to significant changes in the deviance when they are included in the model, the assumption of non-informative censoring may have been violated.

In situations where the reasons for censoring are available, this information may be used to throw light on differences between alternative treatments. For example, if there was a greater proportion of patients with censored survival times on one treatment, and the censoring was due to the occurrence of a particular side effect, this would give information about the merits of the two treatments. One method of analysis, useful when censoring does not occur too early in patient time, is to analyse the survival data at a time before any censoring has occurred. Alternatively, the probability of survival beyond such a time could be modelled. However, this approach is unlikely to be useful when informative censoring occurs early in the study.

In summary there is no satisfactory way to compare survival times of two or more groups of patients in the presence of informative censoring. The earlier warning that great care should be taken to ensure that informative censoring does not occur is therefore repeated.

### 11.2.1 Further reading

Models for data with informative censoring have been described by Wu and Carroll (1988) and Schluter (1992). More recently, Satten *et al.* (2001) and Scharfstein and Robins (2002) have shown how to estimate the survivor function in the presence of informative censoring.

## 11.3 Frailty models

In the analysis of survival data, situations where the survival times are not independent are frequently encountered. Such data tends to arise when different individuals have some feature in common. As an example, consider a multi-centre study, in which the survival experience of individuals from the same centre may be more similar than that for individuals from different centres. This could be because of different surgical teams in the different centres, or different nursing practices across the centres. Similarly, in an animal experiment, animals within a litter will be more alike than animals from different litters, because of genetic and environmental influences.

Some applications lead to repeated event times within an individual, and so here the event times within an individual will not be independent. Examples of this type of situation include studies on the times between successive adverse events, such as migraine or nausea, the times to the failure of tooth fillings in an individual, and times to the failure of transplanted kidneys in patients that receive more than one transplant.

When there is some characteristic that is shared by more than one observed event time, account can be taken of the effects of this characteristic by introducing a corresponding term into the model. Suppose that the characteristic of interest is represented by a factor  $Z$ , where the effect due to the  $j$ th level of  $Z$  is denoted by  $\zeta_j$ . Such effects may then be included in the linear component of a proportional hazards model, or an accelerated failure time model.

In this book, the effects corresponding to factors of interest have always been assumed to be fixed. Inference about the effect of the factor on the hazard of death is then based on estimates of the corresponding fixed effects. But in the applications envisaged here, the factor  $Z$  may have quite a large number of levels. For example, in a multicentre clinical trial involving 30 different centres, the centre effect would be represented by a factor with 29 levels. It is generally undesirable to include so many unknown parameters in a model, particularly when there is no structure among the factor levels. We might therefore represent the effect of the factor  $Z$  by a *random effect*. The effects  $\zeta_j$  are then assumed to be observations from a probability distribution

with zero mean and variance,  $\sigma_\zeta^2$ , and this variance is now the only unknown parameter that needs to be estimated.

In the context of survival analysis, a random effect is often referred to as a frailty. This is because individuals with a larger value of the effect will in consequence have a larger value of the hazard function, if the standard proportional hazards model is adopted. They are then more likely to die sooner, and are therefore deemed to be more frail. Models that include a random effect to represent a characteristic whose values are shared by groups of individuals, are referred to as *shared frailty models*.

There are a number of consequences that follow from the inclusion of a random effect to represent a shared characteristic in a survival model. Because the same realisation of the random component is common to all individuals in the group that have a given level of the characteristic, there will be some dependence between the survival times of individuals in the same group. Indeed, it is because of this that shared frailty models provide a method for modelling survival data when the survival times are not independent.

Inclusion of a random effect in a Cox regression model can lead to hazards not being proportional, while the introduction of a random effect into a Weibull proportional hazards model can lead to non-monotonic hazard function. This means that models that include random effects provide an alternative way of modelling data where the hazard function is not monotonic, or where the hazards are not proportional. As an illustration of this, consider a group of individuals who have experienced a non-fatal heart attack. For such individuals, the hazard of death is generally observed to decline with time, and there are two possible reasons for this. First, the individuals may simply adjust to any damage to the heart that has been caused by the heart attack. Alternatively, the hazard of death may be constant, but the observed decrease in the hazard may be due to frailty; higher risk individuals simply die earlier, so that at any time, the individuals who remain alive are those that are less frail. Note that the survival data cannot be used to distinguish between these two possible explanations for the apparent decline in the hazard of death.

### 11.3.1\* The shared frailty model

To formulate a shared frailty model for survival data, suppose that there are  $g$  groups of individuals with  $n_j$  individuals per group,  $j = 1, 2, \dots, g$ . For the proportional hazards model, the hazard of death at time  $t$  for the  $i$ th individual,  $i = 1, 2, \dots, n_j$ , in the  $j$ th group is then

$$h_{ij}(t) = \exp(\boldsymbol{\beta}' \mathbf{x}_{ij} + \zeta_j) h_0(t), \quad (11.1)$$

where  $\mathbf{x}_{ij}$  is a vector of values of  $p$  explanatory variables for the  $i$ th individual in the  $j$ th group,  $h_0(t)$  is the baseline hazard function, and  $\zeta_j$  is the random group effect. This is assumed to be a value from a probability distribution with density function  $f(\zeta_j)$ . The hazard can also be written in the form

$$h_{ij}(t) = \xi_j e^{\boldsymbol{\beta}' \mathbf{x}_{ij}} h_0(t),$$

where  $\xi_j = \exp(\zeta_j)$ . The distribution assumed for  $\xi_j$  is usually taken to have unit mean, and the gamma distribution is a popular choice. A lognormal distribution may also be adopted for  $\xi_j$ , in which case  $\zeta_j$  in equation (11.1) has a normal distribution.

The general accelerated failure time model that incorporates a shared frailty component is of the form

$$h_{ij}(t) = e^{-\eta_{ij}} h_0(t/e^{\eta_{ij}}),$$

where  $\eta_{ij} = \boldsymbol{\alpha}' \mathbf{x}_{ij} + \zeta_j$ . Equivalently, this model can be expressed in log-linear form as

$$\log T_{ij} = \mu + \boldsymbol{\alpha}' \mathbf{x}_{ij} + \zeta_j + \sigma \epsilon_{ij},$$

where  $T_{ij}$  is the random variable associated with the survival time of the  $i$ th individual in the  $j$ th group, and  $\epsilon_{ij}$  has some specified probability distribution.

Models in which  $h_0(t)$  is fully specified, such as the Weibull proportional hazards model or the accelerated failure time model, can be fitted using the method of maximum likelihood. Denote the observed data by the pairs  $(t_{ij}, \delta_{ij})$ , where  $\delta_{ij}$  is the event indicator, which takes the value zero for a censored observation and unity for an event. If the random effects  $\zeta_j$  had known values, the likelihood function would be

$$\prod_{j=1}^g \prod_{i=1}^{n_j} h_{ij}(t_{ij})^{\delta_{ij}} S_{ij}(t_{ij}),$$

in which  $S_{ij}(t_{ij})$  is the survivor function for the  $i$ th individual in group  $j$ . This function is given by

$$S_{ij}(t_{ij}) = \exp \{ -\exp(\boldsymbol{\beta}' \mathbf{x}_{ij} + \zeta_j) H_0(t_{ij}) \}$$

for the proportional hazards model, where  $H_0(t_{ij})$  is the cumulative hazard function evaluated at time  $t_{ij}$ . Expressions for the survivor function of an accelerated failure time model are given in Section 6.4 of Chapter 6.

However, the  $\zeta_j$  are not known, but are realisations of a random variable that has a probability distribution with density  $f(\zeta_j)$ . In this situation, we integrate the likelihood over possible values of the random effects, to give the likelihood function

$$\prod_{j=1}^g \int_0^\infty \prod_{i=1}^{n_j} h_{ij}(t_{ij})^{\delta_{ij}} S_{ij}(t_{ij}) f(\zeta_j) d\zeta_j.$$

Numerical methods are usually needed to maximise this function, or its logarithm. In principle, a similar procedure can be adopted to fit random effects in the Cox model, but because this is rather more complicated, details will be omitted.

After fitting a shared frailty model, nested models can be compared using the statistic  $-2 \log \hat{L}$ . The estimated coefficients of the explanatory variables in the model are interpreted in the usual manner, and estimates of the random effects can be found using what are known as *empirical Bayes methods*. Such estimates are particularly useful when the frailty term represents centres in

a multicentre study, since the rank order of the estimated centre effects provides information about the merits of the different centres in terms of patient survival. Estimates of the random effects also lead to estimates of the survivor function for individuals with given characteristics, and median survival times can easily be obtained.

### 11.3.2 Further reading

The excellent book of Hougaard (2000) includes an extensive discussion of frailty models, and a number of illustrative examples. General review papers include those of Aalen (1994), Aalen (1998), and Hougaard (1995). Klein (1992) and Klein and Moeschberger (1997) describe how the Cox regression model with a shared frailty component can be fitted using the EM algorithm. More general texts on mixed models include Brown and Prescott (2000), Bryk and Raudenbush (1992), Longford (1993) and Snijders and Bosker (1999).

## 11.4 Multistate models

The experience of a patient in a survival study can be thought of as a process that involves two *states*. At the point of entry to the study, the patient is in a state that corresponds to their being alive. Patients then transfer from this “live” state to the “dead” state at some *transition rate*  $h(t)$ , which is the hazard of death at a given time  $t$ . The situation is expressed diagrammatically in Figure 11.6. The dependence of the rate of transition from one state to the other on explanatory variables is then modelled.

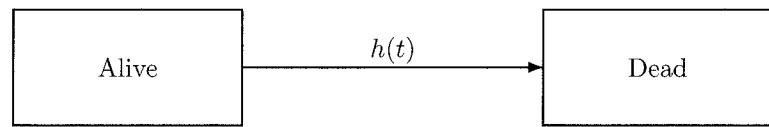
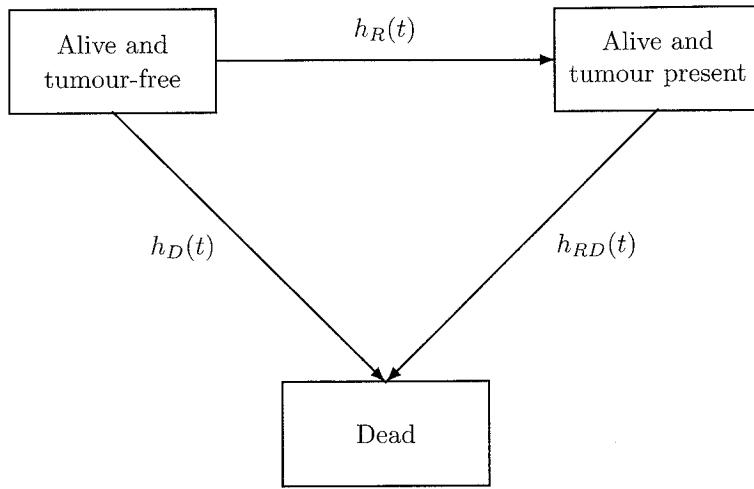


Figure 11.6 A two-state model for survival analysis.

In some studies, the state representing those alive can be partitioned into two or more states, each of which corresponds to a particular stage in the natural progression of the disease. To fix ideas, consider a study concerned with a type of cancer, in which the survival times of patients are recorded from the surgical removal of a primary tumour. Following removal of the tumour, a patient is at risk of death, but there is also the competing risk of a recurrence of the cancer. If the single event of death is used as the end-point of the study, no distinction is drawn between those patients who have experienced a recurrence and those who have not.

It is unlikely that the hazard of death in patients who have had a recurrence will be the same as that in patients who have not. Moreover, the prognostic factors associated with the hazard of death may be different in these two groups of patients. This suggests that the hazard of a recurrence should be

modelled as a function of explanatory variables, in much the same way as the time to death. However, this analysis cannot by itself shed light on factors affecting the time from surgery to death. We therefore adopt a three-state model for the study. The model is illustrated in Figure 11.7.



**Figure 11.7** A three-state model for survival analysis.

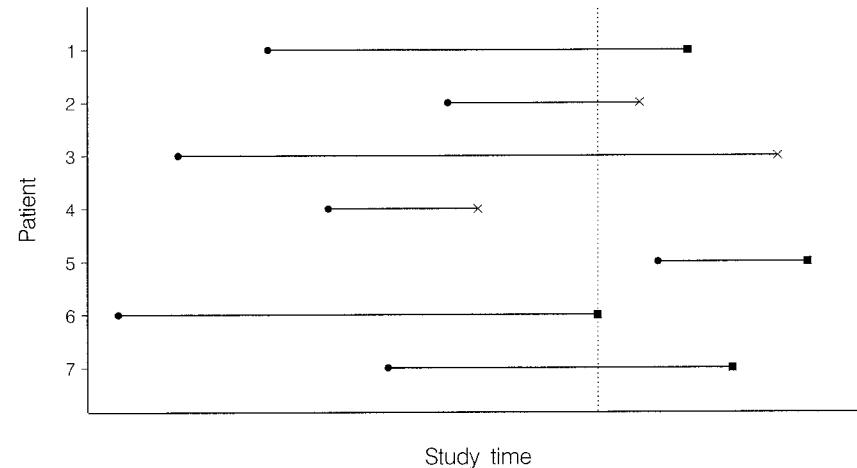
This model can be specified in terms of three hazard functions. The hazard of death at  $t$  without a recurrence of the disease is denoted  $h_D(t)$ , the hazard of a recurrence at  $t$  is denoted  $h_R(t)$ , and the hazard of death at  $t$  after a recurrence is  $h_{RD}(t)$ . Notice that although  $h_D(t)$  and  $h_{RD}(t)$  both denote the hazard of death at time  $t$ , the latter is conditional on the cancer having recurred.

Hsieh *et al.* (1983) have shown that each transition in the multistate model can be analysed separately, using a Cox proportional hazards model. It is straightforward to model the hazard of death without recurrence,  $h_D(t)$ , and the hazard of a recurrence,  $h_R(t)$ . In modelling  $h_D(t)$ , the survival times of those patients who have suffered a recurrence are taken to be censored at the recurrence time. Patients who are alive and without a recurrence also contribute censored survival times. When modelling  $h_R(t)$ , the end-point is the recurrence time. The survival times of those patients who have not suffered a recurrence are regarded as censored, irrespective of whether they are still alive, or have died without experiencing a recurrence.

It is not so straightforward to model the hazard of death in those who have experienced a recurrence,  $h_{RD}(t)$ . This is because the set of patients at risk of death at any time consists of those who have had a recurrence of the disease and are still alive at that time. Patients who have not yet had a recurrence cannot be in this risk set.

For example, consider the survival experience of seven patients who have

had a recurrence, shown schematically in Figure 11.8. In this figure, the recurrence times are denoted by a “●”, censored survival times are denoted by a “×” and death is denoted by a “■”. At the time of death of patient number 6, patients 1, 2, 3, 6 and 7 are in the risk set.



**Figure 11.8** Risk set for seven patients who have had a recurrence.

The  $\beta$ -parameters in the hazard function,  $h_{RD}(t)$ , can be estimated after restructuring the data set. In the new data set, the record for each individual who has died after a recurrence is followed by the records for all patients who are at risk at the death time of the individual who has died. When the data are expressed in this manner, the data can be envisaged as arising from a *matched case-control study*. The “cases” are those who die following a recurrence. For each case there is a set of “matched controls”, which are the patients in the risk set at the death time of the case.

In the example shown in Figure 11.8, there are four cases, namely patients 1, 5, 6 and 7. The sets of matched controls for patients 1, 6 and 7 are  $\{3, 5, 7\}$ ,  $\{1, 2, 3, 7\}$  and  $\{3, 5\}$ , respectively. There is no control corresponding to patient 5.

After this data set has been constructed, the data can be analysed using the technique of *conditional logistic regression*. This technique is described in Collett (2003) and provides a method for estimating the effects of different explanatory variables on the hazard function  $h_{RD}(t)$ . Fuller justification for this approach is given in the section below.

#### 11.4.1\* The likelihood function used in modelling $h_{RD}(t)$

Suppose that there are  $r$  individuals who die following a recurrence and that  $t_j$  is the death time of the  $j$ th such individual,  $j = 1, 2, \dots, r$ . Let  $\mathbf{x}_j$  be the vector of values of  $p$  explanatory variables,  $X_1, X_2, \dots, X_p$ , for the  $j$ th

individual, so that  $x_j$  is the set of values of these explanatory variables for the case in the restructured data set. Also, let  $x_l$  be the vector of values of the explanatory variables for the  $l$ th individual in the risk set at  $t_j$ . This is the set of individuals who have had a recurrence and are still alive at  $t_j$ , and is denoted  $R(t_j)$ . From equation (3.4), the partial likelihood function is

$$\prod_{j=1}^r \frac{\exp(\beta' x_j)}{\sum_{l \in R(t_j)} \exp(\beta' x_l)}.$$

This function can be written in the form

$$\prod_{j=1}^r \frac{\exp(\beta' x_j)}{\exp(\beta' x_j) + \sum_{l^* \in R^*(t_j)} \exp(\beta' x_{l^*})}, \quad (11.2)$$

where  $R^*(t_j)$  is the risk set at  $t_j$ , excluding the  $j$ th individual. This risk set is therefore formed from the matched controls corresponding to the case in the revised data set. If we now write expression (11.2) as

$$\prod_{j=1}^r \left[ 1 + \sum_{l^* \in R^*(t_j)} \exp\{\beta'(x_{l^*} - x_j)\} \right]^{-1},$$

it can be seen that the likelihood is expressed in terms of differences between the explanatory variables of the case and those of the controls. This form of the likelihood function is equivalent to that used in connection with conditional logistic regression. It therefore follows that computer software for analysing matched case-control studies using conditional logistic regression can be used to estimate the  $\beta$ -parameters in the hazard function,  $h_{RD}(t)$ .

#### 11.4.2 Further reading

A thorough review of multistate models is given by Andersen and Keiding (2002), and subsequent papers, in an issue of *Statistical Methods in Medical Research* that is devoted to this topic. Further examples of the use of multistate models in medical research have been given by Weiden *et al.* (1981), Hougaard and Madsen (1985) and Andersen (1988).

Another form of multistate model is the *competing risks model*. This model is relevant when there are several types of failure, corresponding to different causes of death, for example. The multistate model then has more than one death state. Crowder (2001) gives a comprehensive account of the methodology needed for the analysis of such data.

In order to use the methods presented in this section, the recurrence times must be known. Multistate models that do not rely on the recurrence times being known have been considered by many authors in connection with animal tumourigenicity experiments. In particular, see Dinse (1991), Kodell and Nelson (1980), and McKnight and Crowley (1984). A useful review of this literature is included in Lindsey and Ryan (1993).

#### 11.5\* Effect of covariate adjustment

In linear regression analysis, one effect of adding covariates to a model is to reduce the residual mean square, and hence increase the precision of estimates based on the model, such as a treatment effect. The estimated treatment effect, adjusted for explanatory variables, will then have a smaller standard error than the unadjusted effect. In modelling survival data, the inclusion of relevant explanatory variables often has a negligible effect on standard errors of parameter estimates. Indeed, the standard error of an estimated treatment effect, for example, may even be larger after adjusting for covariates. Essentially, this is because the treatment effect does not have the same interpretation in models with and without covariates, a point made by Ford *et al.* (1995). Nevertheless, it is important to include relevant explanatory variables in the model, and to check that the fitted model is appropriate, in order to ensure that a proper estimate of the treatment effect is obtained.

To illustrate this in a little more detail, suppose that  $t_i$  is the observed value of the random variable  $T_i$ , that is associated with the survival time of the  $i$ th of  $n$  individuals,  $i = 1, 2, \dots, n$ . We will consider the situation where there are two treatment groups, with  $n/2$  individuals in each group, and where there is a further explanatory variable whose values are available for each individual. The two explanatory variables will be labelled  $X_1, X_2$ , where  $X_1$  refers to the treatment effect, and takes the value 0 or 1. The values of  $X_1, X_2$  for the  $i$ th individual will be denoted  $x_{1i}, x_{2i}$ , respectively, and we will write  $z_{ji} = x_{ji} - \bar{x}_j$ , for  $j = 1, 2$ , where  $\bar{x}_j$  is the sample mean of the values of the explanatory variable  $X_j$ . A proportional hazards model will be adopted for the dependence of the hazard of death at time  $t$ , for the  $i$ th individual, on the values  $z_{1i}, z_{2i}$ , in which the baseline hazard is a constant value,  $\lambda$ . Consequently, the hazard function can be expressed in the form

$$h_i(t) = \lambda \exp(\beta_1 z_{1i} + \beta_2 z_{2i}), \quad (11.3)$$

and under this model, the survival times are exponentially distributed, with means  $\{\lambda \exp(\beta_1 z_{1i} + \beta_2 z_{2i})\}^{-1}$ . Using results given in Section 6.5.1 of Chapter 6, this model may also be expressed in accelerated failure time form as

$$\log T_i = \mu - \beta_1 z_{1i} - \beta_2 z_{2i} + \epsilon_i, \quad (11.4)$$

where  $\mu = -\log \lambda$  and  $\epsilon_i$  has a Gumbel distribution, that is  $\log \epsilon_i$  has a unit exponential distribution. The model represented in equations (11.3) or (11.4) will be referred to as Model (1).

Using the results for maximum likelihood estimation given in Appendix A, it can be shown that the approximate variance of the estimated treatment effect,  $\hat{\beta}_1$ , in Model (1), is

$$\text{var}(\hat{\beta}_1) = \frac{1}{[1 - \{\text{corr}(z_1, z_2)\}^2] \sum_{i=1}^n z_{1i}^2},$$

where  $\text{corr}(z_1, z_2)$  is the sample correlation between the values  $z_{1i}$  and  $z_{2i}$ . Since  $z_{1i}$  is either  $-0.5$  or  $0.5$ , and there are equal numbers of individuals in

each group,  $\sum_{i=1}^n z_{1i}^2 = n/4$ , and so

$$\text{var}(\hat{\beta}_1) = \frac{4}{n[1 - \{\text{corr}(z_1, z_2)\}^2]}. \quad (11.5)$$

Now consider the model that only includes the variable associated with the treatment effect,  $X_1$ , so that

$$h_i(t) = \lambda \exp(\beta_1 z_{1i}), \quad (11.6)$$

or equivalently,

$$\log T_i = \mu - \beta_1 z_{1i} + \epsilon_i, \quad (11.7)$$

where again  $\log \epsilon_i$  has a unit exponential distribution. The model described by equations (11.6) or (11.7) will be referred to as Model (2). In this model, the approximate variance of  $\hat{\beta}_1$  is given by

$$\text{var}(\hat{\beta}_1) = \frac{4}{n}.$$

Since the term  $1 - \{\text{corr}(z_1, z_2)\}^2$  in equation (11.5) is always less than or equal to unity, the variance of  $\hat{\beta}_1$  in Model (1) is at least equal to that of Model (2). The addition of the explanatory variable  $X_2$  to Model (1) cannot therefore decrease the variance of the estimated treatment effect.

The reason for this is that Model (1) and Model (2) cannot both be valid for the same data set. If Model (1) is correct, and Model (2) is actually fitted, the residual term in equation (11.7) is not  $\epsilon_i$  but  $\epsilon_i - \beta_2 z_{2i}$ . Similarly, if Model (2) is correct, but Model (1) is actually fitted, we cannot assume that the logarithm of  $\epsilon_i$  in equation (11.4) has a unit exponential distribution. Moreover, the parameter  $\beta_1$  is now estimated less precisely because a redundant parameter,  $\beta_2$ , is included in the model.

More detailed analytic and simulation studies are given in the paper by Ford *et al.* (1995), which confirm the general point that the inclusion of explanatory variables in models for survival data cannot be expected to increase the precision of an estimated treatment effect.

## 11.6 Measures of explained variation

The proportion of variation in a response variable that is explained by a fitted model is often used in statistical modelling to summarise the fit of the model. A number of measures of explained variation have been proposed for use in modelling survival data, which are analogues of the  $R^2$ -statistic that is widely used in linear regression analysis. Many of these have been reviewed by Schemper and Stare (1996), and Schemper and Henderson (2000) propose a new measure that has many favourable properties. The performance of measures of explained variation in survival analysis has also been studied by Schemper and Stare (1996), with the conclusion that no particular statistic can be recommended for general use. Moreover, the most satisfactory measures are more difficult to compute. For this reason, measures of explained variation are not routinely used in modelling survival data.

## 11.7 Modelling a cure probability

In survival analysis, it is generally assumed that all individuals will eventually experience the end-point of interest, if the follow-up period is long enough. This is certainly the case if the end-point is death from any cause. However, in some studies, a substantial proportion of individuals may not have experienced the end-point before the end of the study. This may be because the treatment has effectively cured the patient. For example, in a cancer trial, interest may centre on a comparison of two treatments, where the end-point is death from a particular type of cancer. If the treatment cures the individual, so that after a long period of follow-up, a number of patients have not died from the cancer, there will be a corresponding number of censored observations. This results in a larger proportion of censored observations than is usually encountered, referred to as *heavy censoring*. In this situation, the estimated survivor function will tend to level off at a value that may be considerably greater than zero. It may then be assumed that the population consists of a mixture of individuals, those who are susceptible to the end-point, and those who are not. The latter then correspond to the "cured" individuals. Standard methods of survival analysis then need to be adapted, so that the probability of cure is modelled simultaneously with the time to the event. Some possible models have been proposed by Farewell (1982), Kuk and Chen (1992), Taylor (1995), Sy and Taylor (2000) and Peng and Dear (2000).

## 11.8 Some other designs in survival analysis

One form of design that is often used in clinical trials and animal experimentation is the cross-over design. In this design, individuals receive each treatment in the course of an experiment. Use of this design in survival studies with non-fatal end-points then allows the effect of different treatments to be compared within the same individual. Jones and Kenward (2003) and Senn (2002) provide comprehensive accounts of the design and analysis of cross-over studies. France *et al.* (1991), Fiengold and Gillespie (1996), and Lindsey *et al.* (1996) discuss the analysis of failure time data from such studies.

The procedures described in this book are based on studies that have a fixed sample size. There is much interest in the development and implementation of sequential methods, particularly in clinical trials, where the number of individuals is not fixed in advance. Instead, the study continues until the sample size is sufficient to estimate a particular parameter with a prespecified precision, or to distinguish between two alternative hypotheses. See Whitehead (1997) for full details on sequential survival studies, which can be implemented using the PEST software, documented in MPS Research Unit (2000).

# Index

- Accelerated failure time model, 195, 209  
baseline hazard function, 206  
baseline survivor function, 207  
comparing, 216  
comparison with proportional hazards, 202  
fitting, 216  
using SAS, 334  
general form, 206  
influence diagnostics, 242  
log-linear form, 207  
log-logistic, 212  
hazard function, 213  
percentile, 213  
residuals, 235  
survivor function, 212  
lognormal, 213  
hazard function, 214  
percentile, 214  
residuals, 236  
survivor function, 213, 214  
model checking, 231  
using SAS, 350  
percentile  
standard error, 217  
residuals, 231, 236  
semi-parametric, 229  
summary, 214

models with time-dependent variables, 258  
 nested models, 75  
 three or more groups, 48  
 two groups, 37  
     AFT model, 200  
     model for, 56  
     Weibull model, 168  
     Weibull models, 183  
 Competing risks, 326  
 Complementary log-log transformation, 277, 282  
 Computer software, 331  
 Conditional logistic regression, 325  
 Conditional probability, 12, 65  
 Confidence bands, 53  
     for survivor function, 28  
 Confidence interval for  $\beta$ 's, 69  
     hazard ratio, 70  
     median, 35  
     percentile, 35, 36  
     survivor function, 25  
 Cox proportional hazards model, 56, 57  
     estimation of hazard function, 97, 98  
     estimation of survivor function, 97  
     fitting, 63  
     use of SAS, 333  
     partial likelihood, 63, 64  
     residuals, 111  
 Cox-Snell residuals, 112, 232  
     limitations of in Cox model, 123  
     log-cumulative hazard plot, 148, 236  
     plots of, 122, 236  
 Cross-over design, 329  
 Cumulative hazard function, 12  
     estimation of, 33  
     Kaplan-Meier estimate, 33  
     log-linear AFT model, 209  
     Nelson-Aalen estimate, 33  
 Cumulative hazard plot of residuals, 122, 123  
 Delta-beta, 133, 149, 243  
     standardised, 133  
 Design  
     cross-over, 329  
     sequential, 329  
 Deviance, 76, 116

Deviance residuals, 116, 233, 259  
     plots of, 125, 148, 237  
 Distribution  
     Bernoulli, 281  
     binomial, 24  
     chi-squared, 44  
     exponential, 152, 198  
     gamma, 197  
     generalised  $F$ , 229  
     generalised gamma, 198  
     Gompertz, 191  
     Gompertz-Makeham, 227  
     Gumbel, 210  
     hypergeometric, 42  
     inverse Gaussian, 198  
     log-logistic, 195  
     logistic, 196  
     lognormal, 196  
     standard normal, 44  
     Weibull, 151, 154  
 Distribution function, 11  
     empirical, 15  
 Distribution of function of a random variable, 112  
 Dummy variables, *see* Indicator variables  
 Efficient score, 69, 108, 353  
 Empirical distribution function, 15  
 Empirical survivor function, 15, 22  
 End point, 1  
 Estimation of  
     baseline hazard function, 97, 217  
     with time-dependent variables, 257  
     baseline survivor function, 97, 217  
     with time-dependent variables, 257  
     cumulative hazard function, 33  
     exponential parameters, 160  
     hazard function, 29  
     survivor function, 15  
     Weibull parameters, 165  
 Explained variation, 328  
 Explanatory variables, 55  
     adjusted effects, 74  
     effect of adjustment, 327  
 Exponential distribution, 152, 198  
     hazard function, 152  
     model fitting, 160  
     parameter estimates, 160

percentiles, 152  
     survivor function, 152  
 Factor, 58, 59  
     levels, 58  
 Force of mortality, 11  
 Forward selection, 82  
 Frailty model, 320  
     fitting, 322  
     fitting using SAS, 349  
     shared, 321  
 Gamma distribution, 197  
     hazard function, 198  
     survivor function, 198  
 Generalised  $F$ -distribution, 229  
 Generalised gamma distribution, 198  
     hazard function, 198  
     survivor function, 198  
 Gompertz distribution, 191  
     hazard function, 191  
     percentile, 191  
     survivor function, 191  
 Gompertz-Makeham distribution, 227  
 Greenwood's formula, 25  
 Gumbel distribution, 210  
 Hazard  
     rate, 11  
     ratio, 57, 89  
         confidence interval, 70  
         standard error, 70  
 Hazard function, 11  
     baseline, 57  
         with time-dependent variables, 257  
     bathtub, 228  
     Cox model, 97  
     cumulative, 12  
     estimation of, 29  
     exponential, 152  
     gamma distribution, 198  
     general exponential, 227  
     generalised gamma, 198  
     Gompertz, 191  
     integrated, 12  
     inverse Gaussian, 199  
     Kaplan-Meier estimate, 30  
     kernel smoothed estimate, 32  
 Index plots, 122, 237  
 Indicator variables, 57, 59–61  
 Influence diagnostics, 131, 242  
      $C_i$ -statistic, 244  
      $F_i$ -statistic, 244  
      $I_{\max}$ -statistic, 137  
     delta-beta, 133, 243  
     likelihood displacement, 136  
 Influence of observation on parameter estimate, 132, 243  
     set of parameter estimates, 136, 244  
 Influential observations  
     identification of, 131, 242  
     treatment of, 139  
 Information  
     function, 108, 354  
     matrix, 69, 355  
 Informative censoring, 4, 318  
 Instantaneous death rate, 11  
 Integrated hazard, 12  
 Intensity rate, 11  
 Interaction, 60

Interpretation of parameter estimates, 89  
 Interval censoring, 2, 273  
 Interval-censored data, 273  
     arbitrary censoring, 286  
     modelling, 273, 276, 279, 286  
     modelling using SAS, 348  
     parametric models, 296  
 Inverse Gaussian distribution, 198  
     hazard function, 199  
     survivor function, 198  
 Kaplan-Meier estimate  
     of cumulative hazard, 33  
     of hazard function, 30  
     of survivor function, 19  
     standard error, 23, 27  
     using SAS, 331  
 $-2 \log \hat{L}$  statistic, 75, 183  
 Lack of memory property, 114  
 Left censoring, 2  
 Level of significance, 40  
 Levels of a factor, 58  
 Life-table estimate  
     of hazard function, 29  
     of survivor function, 17  
     standard error, 25  
 Likelihood displacement, 136  
 Likelihood function, 63  
     exponential, 160  
     for parametric model, 159  
     for randomly censored data, 357  
     partial, 64  
     Weibull, 164, 176  
 Likelihood ratio statistic, 75, 76, 354, 356  
 Linear component, 58  
     of model, 58  
 Linear trend, 88  
 $t_{\max}$ -statistic, 137, 149  
 Log-cumulative hazard plot, 142, 157, 169, 179  
     of residuals, 122, 148, 236  
 Log-hazard ratio, 89, 191  
 Log-linear model, 178, 207  
     cumulative hazard function, 209  
     hazard function, 209  
     intercept parameter, 178, 207  
 percentile, 209  
 scale parameter, 178, 207  
 survivor function, 208  
 Log-logistic distribution, 195  
     AFT model, 212  
     hazard function, 196  
     percentile, 196  
     proportional odds model, 225  
     survivor function, 196  
 Log-odds plot, 199  
 Log-rank test, 42, 46  
     for trend, 51  
     relation to proportional hazards model, 106  
     stratified, 49  
     use of SAS, 332  
 Logistic distribution, 196  
 Lognormal distribution, 196  
     AFT model, 213  
     hazard function, 197  
     percentile, 197  
     survivor function, 196  
 LOWESS smoother, 127  
 Mantel-Cox test, *see* Log-rank test  
 Mantel-Haenszel procedure, 44  
 Martingale residuals, 115, 232, 259  
     for null model, 127  
     plots of, 124, 237  
 Maximum likelihood, 63, 158, 353  
     estimates, 63  
 Mean survival time  
     in SAS output, 340  
 Median survival time, 33  
     confidence intervals, 35  
     estimation of, 33  
 Mitscherlich hazard function, 227  
 Mixed term, 61  
 Model  
     accelerated failure time, 206  
     choice, 192  
     comparing, 73  
     Cox proportional hazards, 56  
     linear component of, 58  
     log-logistic, 212  
     lognormal, 213  
     nested, 74  
     null, 74  
     overparameterised, 62  
 Outliers, 125  
 Overfitting, 87  
 Overparameterised model, 62  
 P-value, 39

parametric, 151  
 semi-parametric, 56  
 Weibull, 209  
     with time-dependent variables, 252  
     comparing models, 258  
     model fitting, 253  
     residuals, 259  
 Model checking  
     AFT models, 231  
     Cox proportional hazards models, 111  
     form of covariates, 127  
     models with time-dependent variables, 259  
     parametric models, 231  
 Model selection, 83, 85  
     backward elimination, 82  
     forward selection, 82  
     stepwise procedure, 82  
     strategy, 80  
 Modified Cox-Snell residual, 113  
 Multistate models, 323  
     three-state model, 324  
     two-state model, 323  
 Nelson-Aalen estimate  
     of baseline cumulative hazard, 101  
     of cumulative hazard function, 33  
     of survivor function, 21  
     standard error, 25  
 Nested models, 74  
     comparing, 75  
 Newton-Raphson procedure, 64, 69, 165  
 Non-linearity  
     testing for, 87  
 Non-parametric methods, 15  
 Non-proportional hazards, 313  
     time-dependent variables, 318  
 Normal distribution, 44  
 Null distribution, 39  
 Null hypothesis, 38  
 Null model, 74  
 Probability  
     conditional, 12  
 Probability density function, 11  
 Probability distribution, 11  
 Probability of cure  
     modelling, 329  
 Probability value, *see* P-value  
 Probit, 197  
 Product-limit estimate of survivor function, 20  
 Prognostic index, *see* Risk score  
 Prognostic variables, 55  
 Proportional hazards  
     assumption, 46  
 interpretation of, 40, 70  
 Parameter estimates  
     interpretation of, 89  
 Parametric models, 151, 195  
     assessment of suitability, 155  
     fitting to single sample, 158  
     fitting using SAS, 334  
     residuals, 231  
 Partial likelihood function, 66  
 Partial residuals, 117  
 Patient time, 3  
 Percentage points, 26, 41  
 Percentile  
     AFT model  
         standard error of, 362  
         confidence interval, 35, 36  
         estimation of, 34  
         exponential, 152, 161  
         Gompertz, 191  
         log-linear AFT model, 209  
         log-logistic, 196  
         log-logistic AFT model, 213  
         lognormal, 197  
         lognormal AFT model, 214  
         of survival time, 33, 34  
         standard error of, 35  
         Weibull, 155, 165  
             standard error of, 177, 359  
         Weibull AFT model, 211  
         Weibull model, 177  
             standard error of, 165, 360  
 Percentile-percentile plot, 204  
 Peto-Mantel-Haenszel test, *see* Log-rank test  
 Piecewise Cox model, 318  
 Piecewise exponential model, 202, 229  
 Probability  
     conditional, 12  
 Probability density function, 11  
 Probability distribution, 11  
 Probability of cure  
     modelling, 329  
 Probability value, *see* P-value  
 Probit, 197  
 Product-limit estimate of survivor function, 20  
 Prognostic index, *see* Risk score  
 Prognostic variables, 55  
 Proportional hazards  
     assumption, 46

stratified, 315  
 testing in Cox model, 141, 144  
 testing in Weibull model, 247  
 testing using time-dependent variable, 146  
 Proportional hazards model, 56  
     fitting, 63  
     Gompertz, 190, 191  
     Weibull, 175  
         log-likelihood, 177  
         log-linear form, 178  
         model fitting, 176  
 Proportional hazards property, 169, 212  
 Proportional odds model, 195, 223  
     log-logistic distribution, 225  
 Proportional odds property, 225  
  
 Q-Q plot, 204  
 Quadratic hazard function, 227  
 Quartiles, 35  
  
 $R^2$ -measures, 328  
 Random effect, 320  
 Random variable, 11  
 Rate, 11  
     death, 11  
     hazard, 11  
 Relative hazard, 57  
 Residuals, 111, 231  
     AFT model, 231  
     analysis of, 121, 236  
     Cox proportional hazards model, 111  
     Cox-Snell, 112, 232  
         plots of, 122, 236  
     cumulative hazard plot, 122  
     deviance, 116, 233, 259  
         plots of, 125, 237  
     index plot, 122  
     log-cumulative hazard plot, 122, 123, 236  
     log-logistic model, 235  
     lognormal model, 236  
     martingale, 115, 232, 259  
         plots of, 124, 237  
     modified Cox-Snell, 113  
     partial, 117  
     Schoenfeld, 117, 144  
         scaled, 118, 149  
     score, 118, 133, 233  
  
 standardised, 231  
 Weibull model, 234  
 Right censoring, 2  
 Risk score, 58, 125, 241  
 Risk set, 63, 66  
  
 Sample size requirements, 299  
     computer software, 312  
     number of deaths, 300, 302  
     number of patients, 306, 307, 310  
 SAS, 64, 290, 331  
 Schoenfeld residuals, 117, 144  
     scaled, 118, 149  
 Score residuals, 118, 133, 233  
 Score test, 107, 354, 356  
 Semi-interquartile range, 34  
 Sequential design, 329  
 Shared frailty model, 321  
 Significance level, 40  
 Skewness, 1  
 Smoothing  
     kernel estimate, 32  
     LOWESS, 127  
     residual plots, 127  
     splines, 127  
     using SAS, 349  
 Standard error, 23  
     approximate, 24  
 Standard error of  
     difference in estimates, 92  
     estimated survivor function, 23  
     exponential estimates, 161  
     exponential percentile, 162  
     hazard function, 31  
     hazard ratio, 70, 178  
     Kaplan-Meier estimate, 23, 27  
     life-table estimate, 25, 29  
     Nelson-Aalen estimate, 25  
     percentile, 35  
     percentile of AFT model, 217, 362  
     Weibull hazard ratio, 174  
     Weibull percentile, 165, 177, 359, 360  
 Standardised residuals, 231  
 Stepwise procedure, 82  
 Stratified  
     log-rank test, 49  
 Stratified models, 315  
 Stratified tests, 49  
 Study time, 3

Survival analysis, 1  
 Survival data  
     modelling, 56  
     summarising, 15  
 Survivor function, 11  
     actuarial estimate, 17  
     Altshuler's estimate, 22  
     baseline  
         with time-dependent variables, 257  
     comparing observed and fitted, 240  
     confidence interval, 25  
     Cox model, 97  
     empirical, 15, 22  
     estimation of, 15  
     exponential, 152  
     gamma, 198  
     generalised gamma, 198  
     Gompertz, 191  
     inverse Gaussian, 198  
     Kaplan-Meier estimate, 19  
     life-table estimate, 17  
     log-linear model, 208  
     log-logistic, 196  
     log-logistic AFT model, 212  
     lognormal, 196  
     lognormal AFT model, 213, 214  
     Nelson-Aalen estimate, 21  
     product-limit estimate, 20  
     standard error of estimate, 23  
     Weibull, 154  
     Weibull AFT model, 210  
     with time-dependent variables, 257  
  
 Taylor series approximation, 24  
 Test statistic, 38, 39  
 Tied observations, 67  
     in SAS, 333  
 Time origin, 1  
 Time-dependent variables, 251  
     applications, 260  
  
 Wald test, 69, 345, 354, 356  
 Weibull distribution, 151, 154  
     AFT model, 209  
     baseline hazard function, 175  
     hazard function, 154  
     model fitting, 164  
     parameter estimates, 165  
     percentile, 155  
     proportional hazards model, 175  
     scale parameter, 154  
     shape parameter, 154  
     survivor function, 154  
 Wilcoxon test, 45, 46  
     use of SAS, 332