19

MODELLING SURVIVAL DATA

OBJECTIVES

After reading this chapter, you should be able to:

1. Distinguish between non-parametric, semi-parametric, and parametric analyses of survival time data.

2. Carry out non-parametric analyses using either actuarial or Kaplan-Meier lifetables and compare the survival experiences of groups of individuals using a variety of statistical tests.

3. Generate survivor and cumulative hazard function graphs to display survival data.

4. Understand the relationships among survivor functions $S(t)$, failure functions $F(t)$, probability density functions $f(t)$, hazard functions $h(t)$, and cumulative hazard functions $H(t)$.

5. Carry out a semi-parametric analysis of survival data using a Cox proportional hazards model.
   (a) Evaluate the model to:
      i. assess the validity of the assumption of proportional hazards,
      ii. assess the validity of the assumption of independent censoring,
      iii. evaluate other aspects of the model such as its overall fit, the functional form of the predictors in the model, and check for outliers and influential points.
   (b) Incorporate time-varying effects into the model to evaluate or account for non-proportional hazards.

6. Carry out a parametric analysis of survival data based on an assumption that the survival times have an exponential, Weibull, or log-normal distribution.

7. Incorporate frailty effects into a model to account for unmeasured covariates at the individual or group level.

8. Analyse multiple failure-type (recurrence) data.

9. Fit discrete time survival models when appropriate.

19.1 INTRODUCTION

In previous chapters, we have looked at statistical models for evaluating how much of an outcome occurred (linear regression), whether or not an event occurred (logistic regression), which category of event occurred (multinomial models), and the number of events that occurred (or the rate of event occurrence) (Poisson regression). However, we are often interested in how long it takes for an event to occur (time-to-event data). These data are often referred to as 'survival' data because the outcome of interest is often the time until death (*eg* time to death following heart transplantation (Ganesh *et al*, 2005)). However, the analytical approaches discussed in this chapter apply equally to any time-to-event data (*eg* time to pregnancy following an ectopic pregnancy (Ego *et al*, 2001)). The unit of analysis is often the individual, but it may be a group of people (*eg* time to occurrence of an outbreak of an infectious disease in a nursing home) or cells or organs (*eg* sperm survival (Aitken and Baker, 2006)). In general, we will present the discussion in terms of individuals. The occurrence of the event of interest is often referred to as a 'failure', even though in some cases the outcome is desirable (*eg* time to conception in women in a fertility program). Some relatively recently published texts which cover the analysis of survival data include Cleves *et al* (2008); Collett (2003); Hosmer and Lemeshow (2008); and Therneau and Grambsch (2000).

There are specific issues that affect how we quantify and express time to occurrence of an event and how we evaluate the effects of factors (predictors) on that time. However, before discussing these issues, let's look at a simple hypothetical example (Example 19.1).

19.1.1 Features of survival data

Throughout this chapter, data from the Worcester Heart Attack Study (WHAS) (Goldberg *et al*, 2007; Goldberg *et al*, 2010; McManus *et al*, 2011) are used for the examples. The primary outcome of interest is a patient's survival time following admission to hospital with a heart attack. A distribution of those survival times (for patients for whom the time of death was known) is shown in Fig. 19.1.

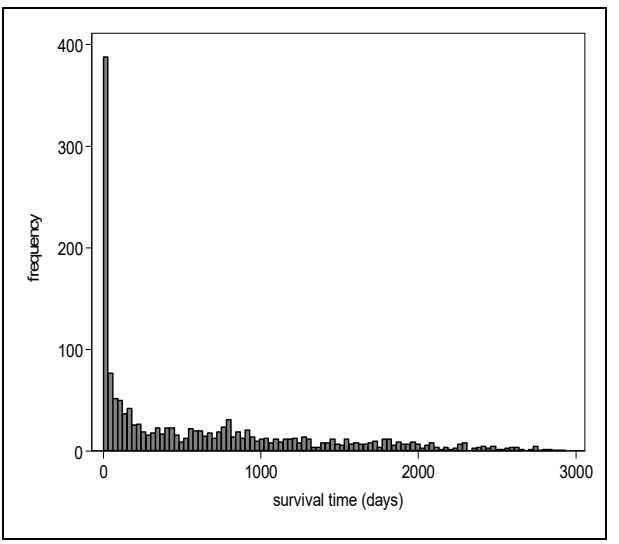

Fig. 19.1 Survival times in heart attack dataset

Here are three common features of survival data.

1. There is **strict left truncation**; there are no values <0.
2. Survival data often have a highly **right-skewed distribution**, with many individuals 'failing' early and a small number having long times to 'failure'.
3. Survival data are often **censored** (*ie* the individual is lost to follow-up before the event of interest (failure) is observed (see Example 19.1).

Example 19.1 Hypothetical survival data
data = brstcan_hyp

Fig. 19.2 shows the time from first occurrence of breast cancer to recurrence in 12 patients. The study was carried out over a 5.5-year period, with patients entering the study as they were diagnosed and treated for the first occurrence. Once enrolled, not all patients were followed for the rest of the study period because some died (from other diseases) or they moved away from the study location. For convenience, all patients were assumed to have had the initial diagnosis and treatment at the start of a year and events (recurrence or loss to study) occurred at the mid-point of a year. In reality, this would not normally be the case.

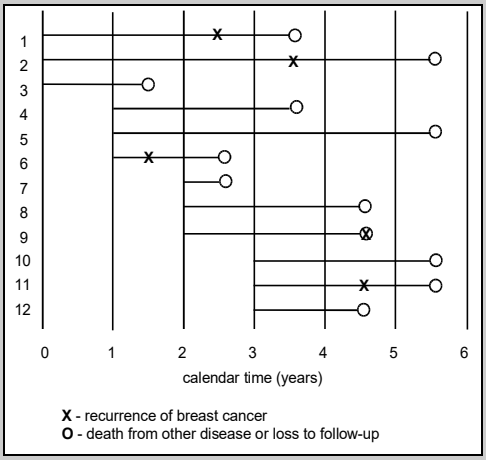

Fig. 19.2 First diagnosis to recurrence

One way to simplify the graphic representation of these 12 people would be to express all recurrence times as being relative to the time of first diagnosis (Fig. 19.3).

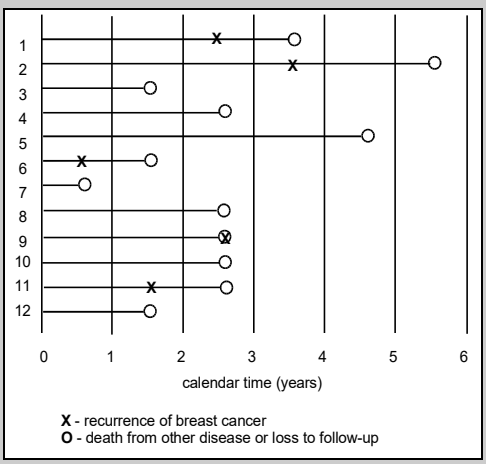

Fig. 19.3 Recurrence relative to first diagnosis

19.1.2 Quantifying survival time

How should the time to recurrence (*ie* time after initial diagnosis) of breast cancer among patients who have been treated be quantified and expressed (Example 19.1)? For many individuals, we do not know the time to recurrence; all we know is that it did not recur in the time period for which the individuals were followed. These 'non-failures' are called censored observations and are a unique feature of time-to-event data.

Possible ways of quantifying and expressing time to recurrence follow (Example 19.1 data).

1. **Mean time to recurrence** The mean time to recurrence can only be computed using

data from individuals in which recurrence has been observed. Consequently, we can only use data from 5 women (mean survival=2.1 years). The estimate will have a downward bias because recurrence in women who had a long time to recurrence are less likely to be observed. On the other hand, if the follow-up observation period is long, the mean suffers from the fact that it might be heavily influenced by a few individuals with very long survival times. Time-to-event data often have an asymmetrical distribution with a long right tail (*ie* right skew).

2. **Median time to recurrence** This can only be computed directly if at least 50% of the individuals are observed to have the event of interest and if none of the censored observations were censored before the failure of the median individual (*ie* if they were going to fail, they had a failure time at least as large as the median). It cannot be computed for the data in Example 19.1. However, if it can be computed, the median is not influenced by a few individuals having long times to recurrence in the same way that the mean is.

3. **Overall probability of recurrence** The proportion of women having a recurrence of the tumour could be computed, but it is not at all clear what constitutes a 'negative' individual (*ie* one who does not have a recurrence). Should the woman be required to have some minimum number of years of follow-up to be considered eligible to contribute to the denominator of the proportion?

4. **n-year survival risk** This expresses the number of women who have not had a recurrence by the n^{th} year. For each year (*eg* first, second) it can be computed based on the women that were observed for that number of years. This approach is often used to quantify survival of people diagnosed with various forms of cancer (*eg* 5-year survival of breast cancer patients). The 2-year 'survival' in Example 19.1 is 0.78 (2 recurrences among 9 women with either 2 complete years of follow-up or a failure at <2 years).

5. **Incidence rate** The number of recurrences relative to the accumulated person-years at risk would be one way to use all of the available data. In some cases, the average time to recurrence could be estimated from the incidence rate (see Section 4.5). However, this approach assumes that the incidence rate of recurrence remains constant throughout the follow-up period and this is often not the case with time-to-event data. The incidence rate in Example 19.1 is 0.19 cases per person-year (5 cases in 26 person-years of follow-up time—women who no longer contribute to the pool of person-years once they have experienced a recurrence).

The approaches outlined above identify 2 key problems to be considered when analysing time-to-event data. First, many observations are censored; that is, the individual is not followed for a length of time sufficient to observe the event of interest if it were to occur. Second, the distribution of survival times is often not symmetrical, and might not even be unimodal. For example, tumour recurrence might be common in the first year after first diagnosis and then relatively rare for several years before becoming more common again as the person ages. These issues are important when evaluating the effect of predictors on the time-to-event occurrence.

19.1.3 Censoring

Censoring is defined as the occurrence (or possible occurrence) of a failure when the individual is not under observation. Censoring can arise in different ways as is summarised in Fig. 19.4.

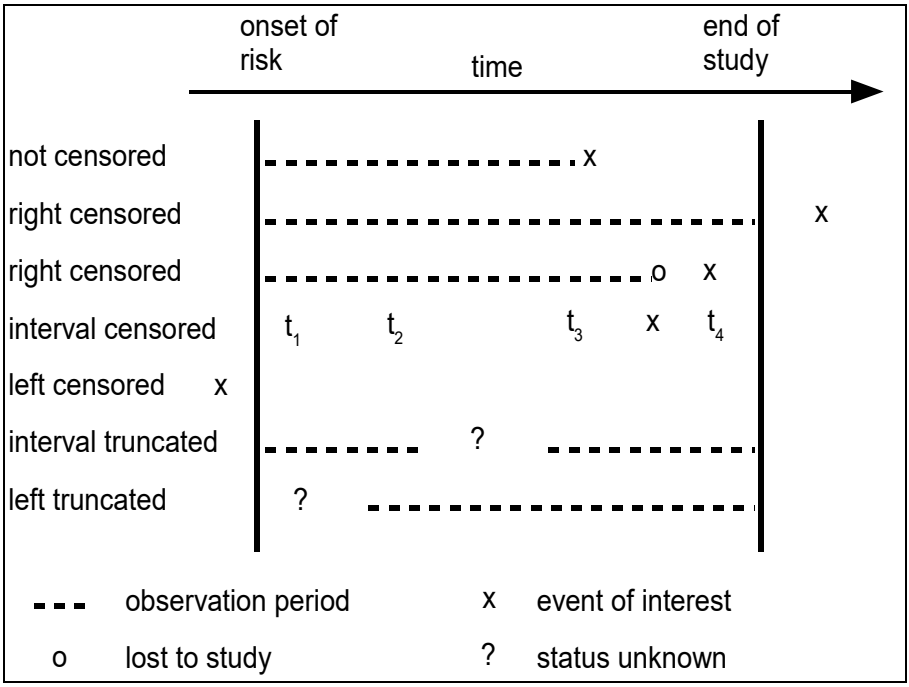

Fig. 19.4 Summary of censoring

Right censoring occurs when a person is lost to a study, before the outcome of interest has occurred. This might arise because the study ends before the event occurs or because it is lost to follow-up during the study (*eg* the person moves to another city). Right censoring is the most common form of censoring that needs to be dealt with in survival analyses.

Interval censoring might arise when an individual is only observed periodically throughout a study period and the event of interest can only be determined by some form of testing carried out at that examination. If examinations are conducted approximately every 6 months, and at one examination (t_4 in Fig. 19.4), it is determined that the event had happened in the preceding 6 months, all that is known is that the event occurred sometime between t_3 and t_4. The precise time the event occurred is not known.

Left censoring is similar to interval censoring except that the 'interval' occurs at the start of the study (*ie* the event occurred in the individual before they entered the period of observation). Consequently, the individual is not put in the study. Left censoring usually arises if the onset of risk occurs before the start of the study. For example, if a study of tumour recurrence started 6 months after surgical removal of an initial tumour but the patient was found to have a recurrence at the 6-month examination, the individual would be left censored. (**Note** If multiple failures are possible, the person might be put in the study and the left censoring then becomes left truncation (see below).)

A related concept is that of **truncation**. While censoring relates to the possible occurrence of outcome events during periods when the individual was not observed, truncation refers to periods of time in which nothing is known about the individual in terms of whether or not the outcome occurred or what the predictors were. These periods of time might be referred to as

gaps. In cases where multiple events are possible (*eg* myocardial infarctions), you have no knowledge of how many cases occurred during the gap. For events which can only occur once (*eg* death), all that is known is that the event did not occur during the gap (or the person would not have come back into the study). Truncation can occur throughout a study (**interval truncation**) or at the start of a study (**left truncation**—also known as delayed entry). Right truncation is the same as right censoring.

As noted above, the most common problem is with right censoring and it will be the only type of censoring or truncation that we deal with in examples in this chapter. A more complete discussion of censoring and how the various forms are dealt with can be found in Chapter 4 of Cleves *et al* (2008).

19.1.4 Evaluating the effect of factors on survival times

Because time-to-event data are continuous, it would be tempting to evaluate the effects of factors on the time to the occurrence of an event using linear regression models. However, as noted above, time-to-event distributions are often not symmetrical, and might not even be unimodal. The assumption of normally distributed errors required for a linear regression model would often be violated in these cases. (In extreme cases, a linear model might predict negative survival times, which are impossible.)

Even if the distribution of the errors is (or can be made) approximately normal, the problem of censored observations remains. In the case of cancer recurrence data, some individuals will be lost to the study, either by withdrawing from the study (*eg* relocation to another city) or through death from an unrelated cause. However, linear models may be used to analyse time-to-event data, provided there are few censored observations and the distribution of survival times is either normal or can be transformed to make them normal. Box-Cox transformation may help with the latter problem.

19.1.5 General approaches to analysing survival data

There are 3 general approaches to analysing survival data:

1. non-parametric analyses.
2. semi-parametric models.
3. parametric models.

These are discussed in much more detail later, but the essential features of each approach are summarised here.

In a **non-parametric** analysis, we make no assumptions about the distribution of survival times, nor about the functional form of the relationship between a factor (predictor) and the survival time. Consequently, they are appropriate only for evaluating the effect of qualitative (categorical) predictors.

In a **semi-parametric** analysis, we make no assumption about the distribution of the survival time, but merely use the survival time to order the observations in terms of time of occurrence of the event. We then evaluate the probability of the event occurring at each of those time points as a function of the predictors of interest. Because the time variable is used only to order the observations, it makes no difference whether there was a large time interval or a small time interval between successive events.

In a **parametric** analysis, we replace the distributional assumption that the errors are normally distributed (as required in a linear model) with a more appropriate distribution that reflects the pattern of survival times. Because we specify a distribution for the survival times, the length of the interval between events is relevant for the analysis. Consequently, if the assumed distribution is correct, a parametric model may be more efficient than a semi-parametric model (*ie* it makes better use of the available data).

19.2 NON-PARAMETRIC ANALYSES

As noted above, in a non-parametric analysis of survival data, we make no assumption about either the distribution of survival times or the functional form of the relationship between a predictor and survival. Hence, they can be used to compare survival experiences of groups of individuals, but not to evaluate the effect of a continuous predictor on survival times. We will look at 3 non-parametric methods for analysing survival data:

- actuarial life tables
- Kaplan-Meier estimator of the survivor function
- Nelson-Aalen estimator of the cumulative hazard function.

In the following section, we introduce the concepts of survivor and hazard functions. These will be described more formally in Section 19.7.

19.3 ACTUARIAL LIFE TABLES

Life tables were originally developed to summarise long-term survival data by dividing the lifespan into short intervals in which the probability of dying was reasonably constant over the time interval. (It certainly is not constant over an entire lifespan.)

The requirements to create an actuarial life table are as follows.

- A clearly demarcated starting point to the period of risk (*eg* birth, first diagnosis of a disease, first exposure to a risk factor *etc*)
- A well-defined study outcome (death, seroconversion, pregnancy diagnosis, tumour recurrence *etc*)
- Only one episode or event per individual (not multiple remissions or relapses)
- Losses to follow-up should be independent of the study outcome (individuals lost from the study should have the same future experience as those that remain under observation)
- The risk of the outcome remains constant over calendar time (no secular (long-term) changes in risk). This does not imply that risk stays the same in an individual over time. Secular changes in survival rates for cancers (*eg* due to better therapies), might for example, affect validity of studies of survivorship
- The risk of outcome must remain constant within the intervals used for constructing a life table. Intervals of any length could be calculated to meet this requirement. In fact, the intervals need not be of the same length.

19.3.1 Steps in constructing the actuarial life table

Table 19.1 shows the columns required to build an actuarial life table, based on the data from Example 19.1.

Table 19.1 Actuarial life table

j	t_{j-1} , t_j	l_j	w_j	r_j	d_j	q_j	p_j	S_j
1	0 < 1	12	1	11.5	1	0.087	0.913	0.913
2	1 < 2	10	2	9.0	1	0.111	0.889	0.812
3	2 < 3	7	3	5.5	2	0.364	0.636	0.516
4	3 < 4	2	0	2.0	1	0.500	0.500	0.258
5	4 < 5	1	1	0.5	0	0.000	1.000	0.258

where ...

j listing of time intervals (time intervals should be established *a priori*)

t_{j-1}, t_j time span covered in the interval

l_j subjects at risk of failure at the start of the time interval
$l_j = l_{j-1} - (w_{j-1} + d_{j-1})$

w_j subjects withdrawn during interval (censored observations)
Individuals who died of causes other than the condition under study or were otherwise lost to follow up during that interval. Individuals who were still free of the outcome when the study ended are counted as withdrawals in the last interval

r_j average number of subjects at risk during the current time interval
$r_j = l_j - (w_j / 2)$
Calculation is based on the assumption that the censored observations were withdrawn, on average, at the mid-point of the interval

d_j outcomes (failures) during the interval
Number experiencing the outcome during the time interval (death, seroconversion, relapse *etc*)

q_j risk of event during interval
$q_j = d_j / r_j$
Probability that the subject will develop the study outcome during the given interval, conditional upon surviving without the outcome up to the start of the interval

p_j probability of surviving the interval
$p_j = 1 - q_j$
Conditional probability of surviving the time interval, given survival to the beginning of the interval

S_j cumulative survival probability to the end of the interval
$S_j = (p_1)(p_2)(p_3)....(p_j)$
Probability of surviving without experiencing the event of interest from the start of follow-up through the end of the current interval in the life table

The risk of an individual experiencing the event of interest during the interval (q_j) divided by the length of the interval is also known as the **hazard**. The cumulative survival probability (S_j) is also known as the **survivor function**. These 2 quantities are key elements of all survival analyses. An actuarial life table (based on 1-year intervals) for individuals in the heart attack dataset who had had previous coronary artery bypass surgery is shown in Example 19.2.

Example 19.2 Actuarial estimates of survival
data = mi (restricted to previous coronary artery bypass patients)

Data on survival of individuals who had previously had coronary artery coronary angioplasty and then experienced a heart attack are shown below. An actuarial life table estimate of the cumulative survivor function based on 1-year intervals is presented here.

Actuarial life table

Interval		Beg. total	Deaths	Lost	Cum. Survival	SE	95% CI	
0	365	391	118	0	0.698	0.023	0.650	0.741
365	730	273	50	0	0.570	0.025	0.520	0.618
730	1095	223	38	0	0.473	0.025	0.423	0.522
1095	1460	185	18	49	0.420	0.025	0.370	0.469
1460	1825	118	21	3	0.344	0.026	0.295	0.395
1825	2190	94	11	38	0.294	0.026	0.244	0.345
2190	2555	45	6	0	0.255	0.027	0.204	0.309
2555	2920	39	3	34	0.220	0.030	0.165	0.281
2920	3285	2	0	2	0.220	0.030	0.165	0.281

The estimate of the cumulative probability of survival up to day 3,285 (22%) is very close to the estimate of the cumulative survival probability at the right-hand end of the graph (based on the Kaplan-Meier survivor function) shown in Fig. 19.5.

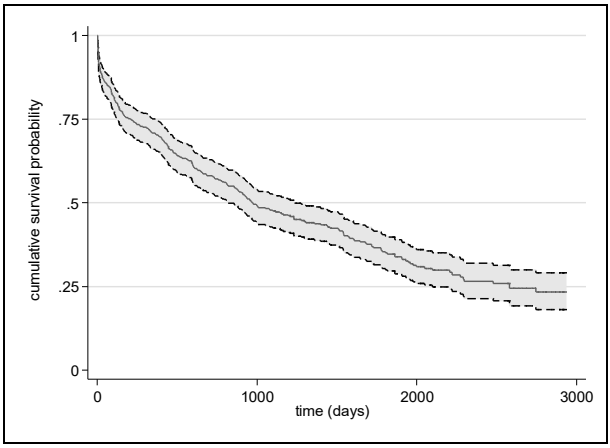

Fig. 19.5 Kaplan-Meier survivor function (95% CI)

19.4 KAPLAN-MEIER ESTIMATE OF SURVIVOR FUNCTION

19.4.1 Overview and comparison to actuarial method

The Kaplan-Meier (K-M) (Kaplan and Meier, 1958) estimate of the survivor function is also known as the **product-limit estimate**. It has 2 important differences from the actuarial estimate described above.

1. The K-M method does not depend on discrete time intervals constructed by the investigator. Each row in the table (hence each time interval) is defined by the time at which the next subject (or subjects, in the case of 2 events happening at the same time) experiences the event of interest.
2. Censored observations (losses to follow up *etc*) between 2 events are counted as individuals at risk only up to the time of the earlier of the 2 events.

The K-M method has the advantage that it avoids the assumption that withdrawals occurred uniformly throughout the interval (*ie* the actuarial assumption) and that the risk is constant over the arbitrarily selected interval. (The only remaining assumption about withdrawals is that they have the same future experiences as those remaining under observation.)

19.4.2 Construction of the K-M life table

An ordered list of the event times is constructed from the sample, with subjects ranked in ascending order of the time of the event of interest. Based on these, Table 19.2 can be filled out (using the data from Example 19.1).

Table 19.2 Kaplan-Meier life table

j	t_j	r_j	d_j	w_j	q_j	p_j	S_j
1	0.5	12	1	1	0.083	0.917	0.917
2	1.5	10	1	2	0.100	0.900	0.825
3	2.5	7	2	3	0.286	0.714	0.589
4	3.5	2	1	0	0.500	0.500	0.295
5	4.5	1	0	1	0.000	1.000	0.295

where:

j listing of time points

t_j time of event

r_j subjects at risk of event at time t_j

$r_j = r_{j-1} - (d_{j-1} + w_{j-1})$

Includes all subjects known to be alive and in the study (not censored) at the time of the event at time t_j, plus the number experiencing the event at time t_j. When censored times are tied with event times, the event is usually assumed to have occurred first

d_j number of events at time t_j

w_j number of censored observations at time t_j

Censoring between time t_j and t_{j+1} is assumed to have happened at t_j so the individuals

will not be considered at risk at time t_{j+1}

q_j risk of event at time t_j
$q_j = d_j/r_j$
Also known as the instantaneous hazard, this is the probability of the event at time t_j, conditional upon survival to time t_j

p_j probability of survival at time t_j
$p_j = 1–q_j$

S_j cumulative probability of surviving up to and including time t_j
$S_j = (p_1)(p_2)(p_3)....(p_j)$

Survivor functions are usually presented graphically as step functions of the cumulative survival over time. They start at one and monotonically descend (*ie* they never go up) as time proceeds. Fig. 19.5 shows a Kaplan-Meier survivor function (and its 95% confidence intervals) for individuals in the heart attack data study who had had previous coronary artery bypass surgery. A tabular listing of the survivor function would be far too long to be useful. Some issues related to the presentation of survival plots have been presented (Pocock *et al*, 2002), including a suggestion that plots of failure functions (see Section 19.7) might be more useful.

19.4.3 The Kaplan-Meier function and estimator

The Kaplan-Meier estimator plays an important role in many procedures used for the analysis of survival data. This section describes the estimator and resulting function in slightly more technical detail.

Assume the following:

t_j j=1,...,n are failure times
t^* is the final failure time = max(t_j)
d_j the number of failures at time t_j
r_j the number of subjects at risk at time t_j
I_k time (0, t^*) is divided into many small intervals (I_k)
p_k the probability of surviving through I_k if alive at the start of I_k

As I_k gets very small, then p_k=1 if there is no failure during the interval and p_k=(r_j-d_j)/r_j if t_j falls in the interval I_k. If there are ties between failures and censored observations, it is assumed that the failures occurred first (*ie* the censored observations are included in the 'at-risk' group).

The Kaplan-Meier estimator of survival $S(t)$ at time t is defined as:

$$S(t) = \prod_{j:t_j \le t} (r_j - d_j)/r_j \qquad \text{for } 0 \le t \le t^* \qquad \textit{Eq 19.1}$$

The Kaplan-Meier function is therefore a **piecewise constant** (*ie* remains constant over time intervals), **non-increasing** (*ie* it can be flat or go down, but never up) and **right-continuous** (*ie* after an event, it remains constant up until, but not including, the next event) function on the interval (0, t^*). It only changes value at failure times (t_j).

The most commonly used standard error (SE) of $S(t)$ is attributed to Greenwood (reported in Collett (2003)). Because survival probabilities often have a very skewed distribution, it is not usual to compute a confidence interval as an estimate ±1.96(SE). Consequently, confidence

intervals are computed by estimating $S(t)$ and its SE on either a natural log (ln) scale or a ln(-ln) scale, and then back-transforming the estimates to the original time scale. (**Note** the ln(-ln) transformation maps probabilities (0,1) onto $(-\infty, \infty)$.)

19.5 NELSON-AALEN ESTIMATE OF CUMULATIVE HAZARD

In the 2 sections above, we introduced the concept of 'hazard' being the probability of failure at a point in time, given that the individual had survived up to that time point. This is discussed more formally in Section 19.7, but for now, we note that a **cumulative hazard** (Nelson-Aalen estimate) can also be computed. The cumulative hazard is the expected number of outcomes for one subject occurring up to a point in time (assuming that the outcome could occur multiple times in an individual). For example, in the heart attack data, the cumulative hazard at day 3,000 would be the sum of all the individual hazards (computed at failure times) up to day 3,000.

The cumulative hazard can range from 0 to infinity (as the time period gets longer, the expected number of outcomes keeps going up with no upper bound). A graph of the cumulative hazard is, like a graph of the survivor function, a way of expressing the overall failure (survival) experience of the population. Fig. 19.6 shows the cumulative hazard (and 95% CI) for the heart attack data (once again restricted to patients who had previous bypass surgery).

Using the notation from Section 19.4.3, the Nelson-Aalen estimator of the cumulative hazard $H(t)$ at time t is computed as:

$$H(t) = \sum_{j:t_j \le t} d_j / r_j \qquad \text{for } 0 \le t \le t^* \qquad \textit{Eq 19.2}$$

As with the Kaplan-Meier estimator of $S(t)$, SE can be determined and confidence intervals are computed on a ln scale and back transformed.

19.6 STATISTICAL INFERENCE IN NON-PARAMETRIC ANALYSES

19.6.1 Confidence intervals and 'point-in-time' comparisons

Although the formulae have not been shown, SEs of the cumulative survival estimates can be computed from actuarial or Kaplan-Meier survivor functions at any point in time. These SEs can be used to test the difference between survivor functions (usually on a log scale) for 2 (or more) populations at any point using a standard normal Z-test. However, there are potentially an infinite number of points at which the cumulative survival probabilities could be computed.

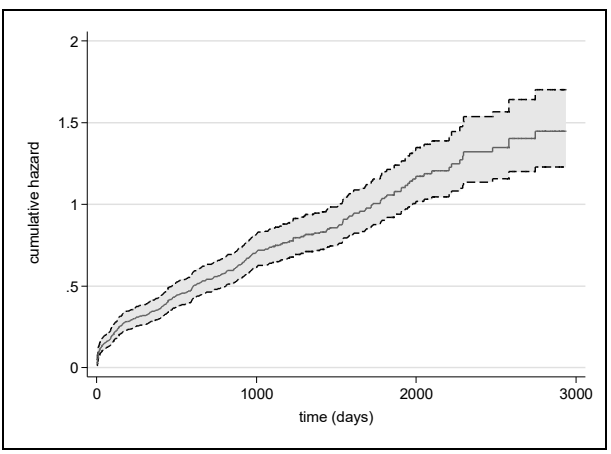

Fig. 19.6 Nelson-Aalen cumulative hazard function (with 95% confidence interval)

This could lead to a serious problem of 'data snooping' or multiple comparisons and consequently, 'point-in-time' comparisons are only valid if it is possible to identify specific times at which the comparison of survival probabilities is warranted. These should be specified *a priori* (*ie* before the data are collected) and if multiple time points are evaluated, some adjustment for multiple comparisons must be made.

19.6.2 Tests of the overall survival curve

There are several tests that can be used to determine whether the overall survivor functions in 2 (or more) groups are equal. They are all based on a series of contingency tables of observed and expected events for each group at each time point at which an event occurred (assuming the test is based on a Kaplan-Meier survivor function). The observed number of events at each time point is compared with the expected number and a χ^2 test computed. (Under the H_0 that there is no difference between the 2 groups, the expected number of events is a function of the amount of follow-up time in each group.) Consequently, the tests can be viewed as the survival analysis equivalent of the Mantel-Haenszel test for stratified data.

All of the tests assume that the ratio of risks of the event of interest for the 2 groups is constant across all strata (equivalent to the no-interaction assumption in a Mantel-Haenszel test). This assumption is known as the 'proportional hazards' assumption (you will see more of this later). If the survivor functions cross over, then it is clear that this assumption is violated. The differences among the tests depend on the weights used to combine the estimates derived at each point in time.

Log-rank test
The log-rank test is the simplest as it assigns equal weight to each point estimate (weights $w(t_j)$=1). Consequently, it is equivalent to doing a standard Mantel-Haenszel procedure to combine the estimates (Example 19.3).

Wilcoxon test
This test weights the intervals according to the sample size ($w(t_j)=n_j$). Consequently, it is more sensitive to differences early in the time period when the sample size is larger. Some advocate

Example 19.3 Log-rank test
data = mi (restricted to previous coronary artery bypass patients)

A log-rank test for equality of survivor functions was used to compare the survival experience of married and not-married individuals. Based on the follow-up time in each of the 2 groups, one would expect to see 94.3 and 165.7 events in the not-married and married groups, respectively. However, we observed 114 and 146 respectively, indicating an excess of deaths in the not-married group. The resultant P-value for this comparison was 0.011.

Married	Events observed	Events expected
no	114	94.34
yes	146	165.66
Total	260	260.00

using both Wilcoxon and the log-rank test to see if differences in the survival curves occur early or late in the time period studied. The Wilcoxon test is less sensitive than the log-rank test to the assumption of proportional hazards, but will be unreliable if the censoring patterns vary across the groups being compared.

Other tests

Other non-parametric tests include the Cox test, the Tarone-Ware test, and the Peto-Peto-Prentice test. The first is based on a Cox regression procedure (see Section 19.8), while the Tarone-Ware weights the stratum-specific estimates by the square root of the population at risk at each time point. The Peto-Peto-Prentice test weights the stratum-specific estimates by the overall survival experience (an estimate of $S(t)$ just before the time point of interest), and consequently reduces the influence of different censoring patterns between the groups.

Example 19.4 shows separate survivor functions for married and not-married individuals who had previously had bypass surgery. The results from several of the tests for the overall equality of the survivor functions are also presented.

19.7 SURVIVOR, FAILURE AND HAZARD FUNCTIONS

The concepts of survivor and hazard functions were introduced when we looked at non-parametric methods of analysis of survival data. Before proceeding with semi-parametric and parametric analyses, we need to develop a more complete understanding of these and related functions.

Example 19.4 Comparing survivor functions

data = mi (restricted to previous coronary artery bypass patients)

Fig. 19.7 shows the Kaplan-Meier survivor functions for married and not-married individuals.

Not-married individuals appeared to have shorter survival times (lower cumulative survival probabilities). The statistical significance of the test results for the difference between these 2 survivor functions are shown below. All statistical tests provide comparable results.

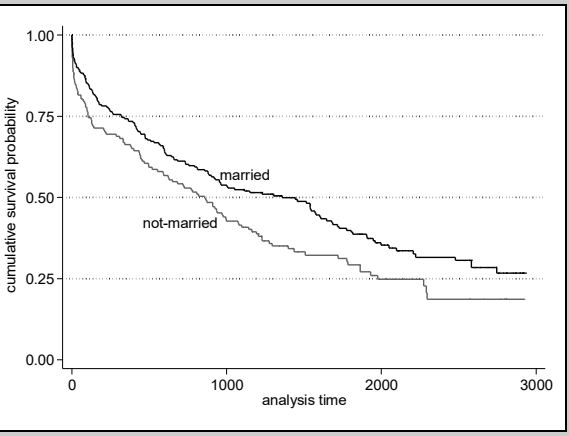

Fig. 19.7 K-M survival curves, by marital status

Test	P-value
log-rank	0.011
Wilcoxon	0.010
Cox	0.012
Tarone-Ware	0.009
Peto-Peto-Prentice	0.011

19.7.1 Survivor function

The survivor function ($S(t)$) is the probability that an individual's survival time (T) (or more generally, the time-to-event occurrence) will exceed some specified time t. It can be written as:

$$S(t) = p(T \geq t) \qquad \textit{Eq 19.3}$$

As noted, survivor functions are non-increasing. They start at 1 and drop to 0 if all individuals ultimately experience the event of interest. **Note** By convention, cumulative functions will be designated by upper-case letters and density functions by lower-case letters. The survivor function is a cumulative function in that it represents the cumulative probability of surviving up to a point in time t.

19.7.2 Failure function

The failure function ($F(t)$) is the probability of not surviving past time t. Consequently, it is:

$$F(t) = 1 - S(t) \qquad \textit{Eq 19.4}$$

19.7.3 Probability density function

The probability density function ($f(t)$) describes the distribution of survival times and is the slope (derivative) of the failure function. Consequently, it represents the instantaneous rate at which failures are occurring in the study population at a point in time. It is estimated by taking the derivative of a smoothed estimate of the failure function with respect to time (Fig. 19.8).

19.7.4 Hazard function

The hazard function ($h(t)$) is the probability of an event occurring at time t given that it had not occurred up to time t. With time divided into discrete intervals (as in a life table), it can be expressed as:

$$h(t) = p(T = t | T \geq t) \qquad \textit{Eq 19.5}$$

With time on a continuous scale, the hazard function describes the instantaneous probability of an event occurring at a point in time given that it did not occur previously. The hazard function is:

$$h(t) = \lim_{\Delta t \to 0} \frac{p(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \qquad \textit{Eq 19.6}$$

The hazard function can also be computed as the ratio of the probability density function (which represents the rate at which failures are occurring at a point in time) and the survivor function (which represents the probability of surviving up to that point in time). It can further be expressed as:

$$h(t) = \frac{f(t)}{S(t)} = \left[\frac{-\frac{d\,S(t)}{dt}}{S(t)} \right] = -\left[\frac{d}{dt} (\ln S(t)) \right] \qquad \textit{Eq 19.7}$$

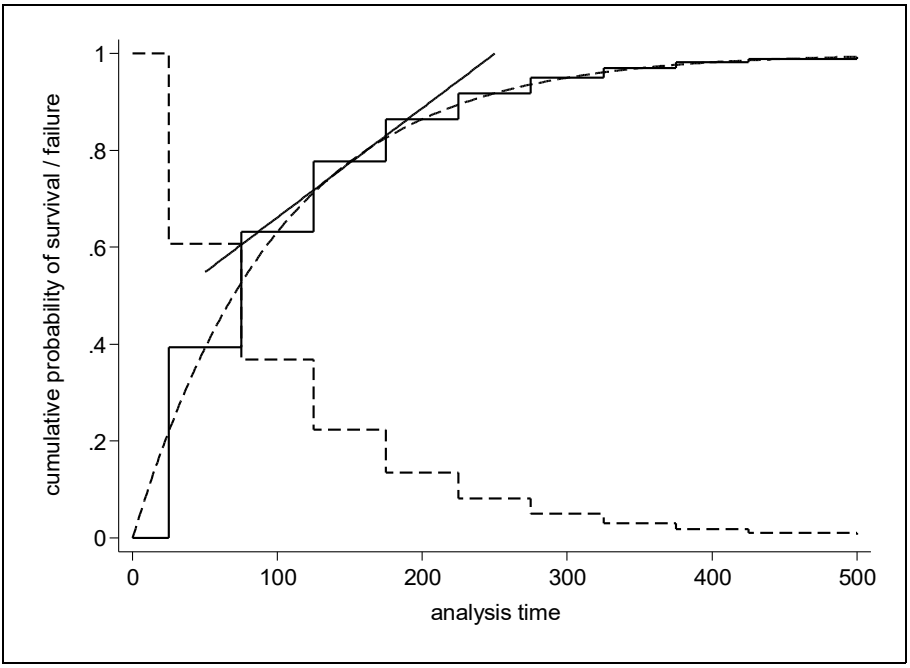

Fig. 19.8 Survivor function (dashed stepped line). Failure function (solid stepped line). Smoothed failure function (dashed curved line) and tangent of smoothed failure function (short solid line) giving the slope at a single point

Hazard functions are always non-negative (*ie* greater than or equal to zero) and have no upper bound (their value will change with the time scale used).

19.7.5 Cumulative hazard function

The cumulative hazard ($H(t)$), also known as the **integrated hazard**, represents the accumulation of hazard over time. It can be computed as the integral of the hazard function but is more conveniently found using the following equation.

$$H(t) = -\ln S(t) \qquad \textit{Eq 19.8}$$

As noted, the cumulative hazard represents the expected number of outcomes of interest that would occur in an individual (assuming that repeat occurrences were possible). For example, if you were studying the survival of lung cancer patients who already have metastases, and at 3 years you find that the cumulative hazard=2.7 [$H(t_3)$=2.7], then that would suggest that in 3 years after initial diagnosis, we would expect to see 2.7 deaths. Obviously, only one death is possible, but it provides an indication that the probability of the individual surviving to 3 years post-diagnosis is very low.

19.7.6 Relationships among survivor, failure, and hazard functions

Some of the relationships between the survivor, failure, and hazard functions have already been shown in previous sections. As each of these functions determines the survival time

distribution, if one of them is known, the others can all be computed.

$$f(t)=\frac{dF(t)}{dt} \quad h(t)=\frac{dH(t)}{dt} \quad h(t)=\frac{f(t)}{S(t)} \qquad \textit{Eq 19.9}$$

Note $f(t)$ and $h(t)$ are derivatives of $F(t)$ and $H(t)$ which are step functions. Consequently, the step functions are smoothed before the derivative is taken.

$$F(t)=1-S(t) \quad H(t)=-\ln S(t) \quad S(t)=e^{-H(t)} \qquad \textit{Eq 19.10}$$

Note The last expression for $S(t)$ (above) gives the Flemming-Harrington estimate of the survivor function when the Nelson-Aalen estimate is used for $H(t)$. This estimate will be larger than the Kaplan-Meier estimate of $S(t)$ computed directly, but will be close if the number of failures is small relative to the number of individuals at risk.

While survival experiences for groups of individuals are usually shown by plotting the survivor function, the hazard function plays a key role in semi-parametric and parametric analyses.

19.7.7 Examples of hazard functions

A wide variety of hazard functions has been studied, but constant and Weibull functions are the 2 most commonly encountered in survival analyses. Other forms used include the log-normal, log-logistic, gamma, and Gompertz. The names of these functions refer to the corresponding survival time distributions (see Section 19.9).

Constant hazard

A constant hazard is one which does not change over time. With a constant hazard (λ), the survivor function drops exponentially and survival times will have an exponential distribution. The hazard $h(t)$, density $f(t)$ and survivor $S(t)$ functions are:

$$h(t)=\lambda \qquad f(t)=\lambda e^{-\lambda t} \qquad S(t)=e^{-\lambda t} \qquad \textit{Eq 19.11}$$

The appropriateness of an exponential model can be assessed by plotting the cumulative hazard $H(t)$ (or equivalently $-\ln S(t)$) against t. If the exponential model is appropriate, the line will be straight. Fig. 19.9 shows a survivor function derived from a constant hazard of λ=0.01 per day.

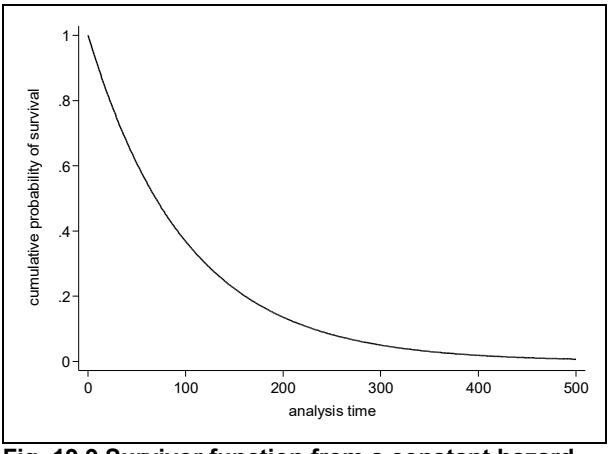

Fig. 19.9 Survivor function from a constant hazard (h(t)=0.01)

Weibull hazard

A Weibull hazard function depends on 2 non-negative parameters: a scale parameter (λ) and a shape parameter (p). If p=1, the resulting survival time distribution is the exponential distribution. If p<1, then the hazard function decreases monotonically. If p>1, then the function is monotonically increasing with a value between 1 and 2 producing a curve that increases at a decreasing rate, p=2

produces a hazard function that increases linearly with time and $p>2$ produces a function that increases at an ever-increasing rate. The hazard and survivor functions are:

$$h(t)=\lambda\, pt^{(p-1)} \qquad S(t)=\exp(-\lambda t^{p}) \qquad \textit{Eq 19.12}$$

Fig. 19.10 shows Weibull hazard functions for several values of p. An increasing Weibull hazard function ($p>2$) might be appropriate for the hazard of death for people over 60 years of age—the hazard increases and does so at an increasing rate. A decreasing Weibull hazard function ($p<1$) might be appropriate for the survival of individuals after cardiac surgery when the hazard is highest right after surgery and then decreases (at least in the short term).

The suitability of the Weibull distribution or hazard can be assessed by evaluating the log-cumulative hazard plot [$\ln(H(t))$ versus $\ln(t)$]. If the data fit a Weibull distribution, the line on the graph should be approximately straight. The intercept and the slope of the line will be $\ln(\lambda)$ and p, respectively. Parametric survival models based on exponential and Weibull hazard functions are described in Section 19.9.

Other hazard functions

One of the limitations of the Weibull hazard function is that the hazard can only increase or decrease over time. **Gamma**, **log-normal** and **log-logistic** hazards can be used to deal with the situation in which the risk first increases and then decreases (or vice versa). Such a function would be appropriate in a situation where the risk of death was high early in an illness, drops to a lower level and then increases again over time. For example, a new infection with *Mycobacterium tuberculosis* might produce a high risk of death early after infection, followed by a sharp reduction in the risk and then a gradually increasing risk of death from chronic tuberculosis. Detailed descriptions can be found in survival analysis texts (Cleves *et al*, 2008; Collett, 2003; Hosmer and Lemeshow, 2008; Therneau and Grambsch, 2000).

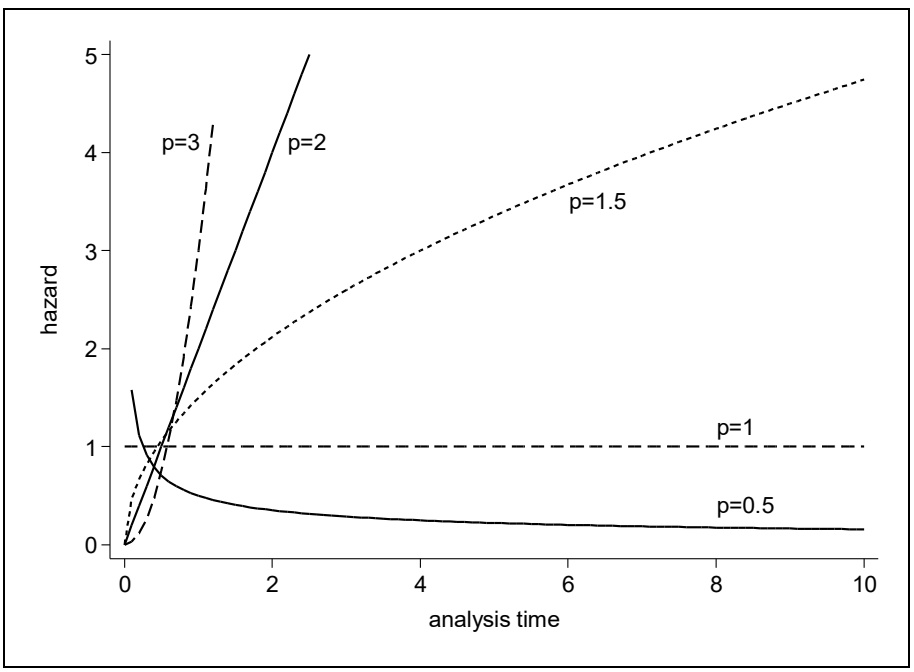

Fig. 19.10 Weibull hazard functions for various values of shape parameter (p)

19.8 SEMI-PARAMETRIC ANALYSES

Non-parametric analyses are limited to evaluating the effect of one, or a small number of, qualitative variable(s) on survival times. However, we often want to simultaneously evaluate the effects of multiple continuous or categorical explanatory variables. This requires that we model the survival data using a multivariable technique. The most commonly used form of multivariable analysis for survival data is the **proportional hazards model** (also known as the **Cox regression model**) (Cox, 1972). It is a semi-parametric model in that we do not have to assume any specific functional form for the hazard, but we do model the ratio of hazards as a linear function of the predictors.

19.8.1 Cox proportional hazards model

The proportional hazards model is based on the assumption that the hazard for an individual is a product of a baseline hazard (h_0) and an exponential function of a series of explanatory variables.

$$h(t) = h_0(t)\mathrm{e}^{\beta X} \qquad \textit{Eq 19.13}$$

where $\beta X = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$. Equivalently, it can be expressed as:

$$HR = \frac{h(t)}{h_0(t)} = \mathrm{e}^{\beta X} \qquad \textit{Eq 19.14}$$

where *HR* is the hazard ratio. The first formulation emphasises that the hazard for an individual is always a multiple ($e^{\beta X}$) of a baseline hazard (see Fig. 19.11; left panel), while the second formulation shows that it is the ratio of the hazards which is assumed to be constant over time. On the log scale, the log hazard is a constant (βX) above or below the baseline log hazard as shown below and in Fig. 19.11 (right panel).

$$\ln h(t) = \ln h_0(t) + \beta X \qquad \textit{Eq 19.15}$$

Two important features of this model are that no assumption is made about the shape of the baseline hazard (h_0) and that the model has no intercept. In fact, the intercept (which in most regression models reflects the value of the outcome when all covariates (predictors) are zero) is subsumed into the baseline hazard which represents the hazard when all covariates are zero.

19.8.2 Hazard ratios

Based on Eq 19.14, the $\ln HR = \beta X$. Consequently, exponentiating the coefficient from a proportional hazards model produces a hazard ratio. Hazard ratios have interpretations similar to odds ratios and risk ratios. They represent the effect of a unit change in the predictor on the frequency of the outcome (which in this case is measured as a hazard). **Note** You will sometimes encounter hazard ratios referred to as relative risks (or risk ratios), but this is not a correct use of the term and should be avoided. For example, if factor X_1 has an $HR=2$, then a unit change in X_1 will double the hazard of the outcome. If X_1 is a dichotomous variable and, because we are assuming that this *HR* is constant (over time), this means that, at any point during the risk period, 'failures' will be occurring at twice the rate in individuals with $X_1=1$ than in individuals with $X_1=0$. This is not equivalent to a doubling of the risk over the full study period.

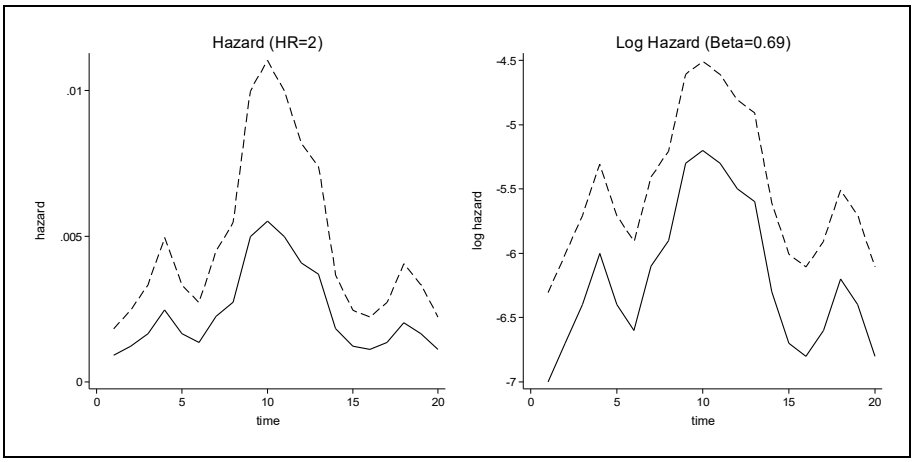

Fig. 19.11 Effect of a hypothetical factor on a baseline hazard shown on 2 scales—hazard scale in left panel and on log hazard scale in right

Example 19.5 provides some examples of *HR*s derived from the Worcester Heart Attack Study (WHAS). Patients admitted to 1 of 10 hospitals in the greater Worcester, Massachusetts area with an acute myocardial infarction were followed for many years after discharge (the longest follow-up period in the data provided was slightly over 8 years). A few factors that might affect survival following the myocardial infarction, and which are used in examples in this chapter, are described in Table 19.3. (**Note** The variables listed are just a small subset of the variables recorded in the WHAS database.) The dataset (-mi-) is described more fully in Chapter 31.

Table 19.3 Key variables in -mi- dataset

Variable	Description
id	patient id (1–2965)
hosp	hospital id (1–10)
surv_mi	survival time from day of admission to hospital (days)
died	observation time ended in death or censoring (1=died, 0=censored)
sex	gender (1=male, 0=female)
age	age at admission (years)
white	race (1=white, 0=other)
mar_c2	married (1=yes, 0=no)
card	cardiac arrest during hospitalisation (1=yes, 0=no)
ptca	coronary angioplasty during hospitalisation (1=yes, 0=no)

Example 19.5 Cox proportional hazards model
data = mi

A Cox proportional hazards model was fit to the heart attack data with sex, age, marital status, cardiac arrest, and coronary angioplasty as predictors. The first table presents the model in terms of coefficients.

No. of subjects = 2851
No. of failures = 1480 Number of obs = 2851
Time at risk = 3669144 LR $\chi^2_{(5)}$ = 1298.51
Log likelihood = -10452.582 Prob > χ^2 = 0.0000

Predictor	Coef	SE	Z	P	95% CI	
sex	0.200	0.058	3.45	0.001	0.087	0.314
age	0.051	0.002	20.42	0.000	0.046	0.055
mar_c2	-0.208	0.057	-3.65	0.000	-0.320	-0.096
card	1.724	0.090	19.22	0.000	1.548	1.900
ptca	-0.919	0.063	-14.67	0.000	-1.042	-0.796

Being male, and having a cardiac arrest while in the hospital both increased the log hazard of death (by 0.200 and 1.724 units, respectively). Being married reduced the log hazard by 0.208 units while having coronary angioplasty while in the hospital reduced it by 0.919 units. For each additional year of age (at admission), the log hazard went up by 0.051 units. As we rarely think in terms of log hazards, it is more common to present the results as *HR*s.

Predictor	HR	SE	Z	P	95% CI	
sex	1.222	0.071	3.45	0.001	1.091	1.369
age	1.052	0.003	20.42	0.000	1.047	1.057
mar_c2	0.812	0.046	-3.65	0.000	0.727	0.908
card	5.606	0.503	19.22	0.000	4.702	6.684
ptca	0.399					

Here it appears that being male increased the hazard of death by 1.2 times. The interpretation of this effect is that at any point after admission to hospital, deaths were occurring at a 20% higher rate in males than in females. Similarly, for each additional year of age, the hazard of death rose by approximately 5%. A pair of Kaplan-Meier survivor functions (one for each gender) shows that males had a higher hazard of death and that this effect remained relatively constant over the whole study period. Graphs for cardiac arrest (yes/no) or coronary angioplasty (yes/no) would not show such a constant effect (discussed in Section 19.8.10).

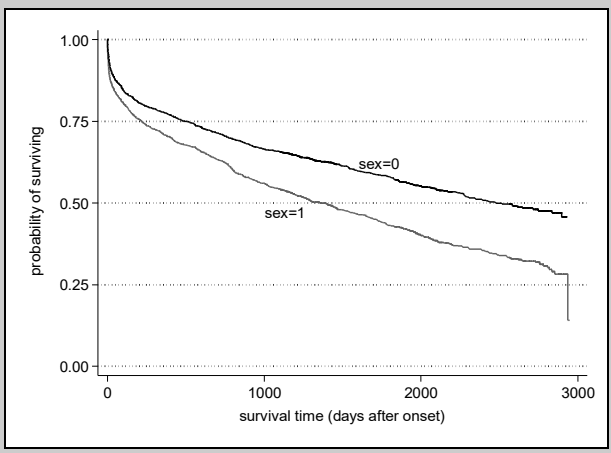

Fig. 19.12 K-M survival estimates, by sex

19.8.3 Fitting the Cox proportional hazards model

Obtaining partial maximum likelihood estimates of the parameters in a Cox proportional hazards model requires an iterative estimation procedure (the Newton-Raphson procedure is most commonly used). As with a non-parametric Kaplan-Meier estimation procedure, a Cox model is only evaluated at the times at which failures occur. In fact, fitting a Cox model with no predictors produces exactly the same survival curve as a Kaplan-Meier estimation does. In both procedures, it is not the actual times at which failures occur which are important, but the order in which they occur.

The estimation is based on the partial (profile, conditional) likelihood function, which has a different interpretation than the usual likelihood function (as described in Example 19.6), but is used in the same way for statistical inference.

19.8.4 Handling of ties

Because the order in which failures occur is critical for conducting the analysis, there must be a way of handling the problem of 2 (or more) failures being recorded at the same time. Details of various methods of dealing with ties can be found in texts on survival analysis but they fall into

Example 19.6 Partial likelihoods for a Cox model
data = hypothetical time to death following diagnosis of lung cancer in 20 people

Assume that you are studying the effect of age on survival in lung cancer patients and the data, sorted by time to death, were as follows:

Person	Time to death (mo)	Age at diagnosis (yrs)
1	3	59.6
2	8	68.1
...	...	...
20	63	65.7

For the first person, a maximum likelihood procedure would ask the question 'What was the probability of this person dying at 3 months, given that s/he was 59.6 years old at the time of diagnosis?' In contrast, a partial likelihood procedure asks the question 'Given that a death occurred at 3 months, what was the probability that it was person #1 (given the age of the person)?' This likelihood can be written as follows.

$$L_1 = \frac{h_1(3)}{h_1(3)+h_2(3)+...+h_{20}(3)} = \frac{h_0(3)e^{\beta*59.6}}{h_0(3)e^{\beta*59.6}+...+h_0(3)e^{\beta*65.7}} = \frac{e^{\beta*59.6}}{e^{\beta*59.6}+...+e^{\beta*65.7}}$$

The partial likelihood of the first failure being person #1 is that person's likelihood relative to the sum of all of the likelihoods. For the second person, the partial likelihood is

$$L_2 = \frac{h_2(8)}{h_2(8)+h_3(8)+...+h_{20}(8)} = \frac{e^{\beta*68.1}}{e^{\beta*68.1}+...+e^{\beta*65.7}}$$

The product of the partial likelihoods is the likelihood of the model. **Note** The analysis only depends on the sequence of events (not the actual time) and that the baseline hazard has no effect because it is common to all individuals.

2 general approaches. The first is called a **marginal calculation** or continuous-time calculation and is based on the assumption that the timing of the events was not really tied, but simply due to the fact that the timing of the failure was not recorded with sufficient precision to differentiate among 'tied' observations. The second is called the **partial calculation** and is based on the assumption that the events were actually tied and treats the problem as a multinomial problem.

Exact calculation of the likelihood function under either assumption is computationally demanding (particularly for partial calculations) and may be slow in large datasets with many ties. Two approximate methods have been developed for marginal calculations. The **Breslow** method is simplest and is adequate if there are not a lot of ties. The **Efron** method provides a closer approximation to the exact calculation. An approximation attributed to **Cox** can be used for partial calculations. However, for a moderate dataset such as -mi-, exact calculation methods (marginal or partial) are feasible. In this case, the exact methods and the Breslow and Efron approximations all produce very similar results (data not shown). Because of the general superiority of the Efron method compared with the Breslow method, this approach will be used in all subsequent Cox models in this chapter.

19.8.5 Baseline hazard

Although, as noted above, no assumption is made about the baseline hazard (h_0) and the Cox model does not estimate it directly, an estimate of it can be derived conditional on the set of coefficients in the estimated model. This baseline hazard represents the hazard in an individual for whom all predictors equal zero. For it to be meaningful, it is important that X=0 is reasonable for all predictors. If computed from the -mi- data using the model shown in Example 19.5 (with age rescaled to a baseline of 60 years), this would represent the daily hazard of death in a 60-year-old, not-married female who had neither coronary angioplasty nor a cardiac arrest while in hospital.

The baseline hazard can only be estimated on days on which failures occur, and the estimate will bounce around quite a lot from day to day (particularly once the surviving population at risk becomes small). Consequently, it is necessary to smooth the estimate of the baseline hazard and this is shown in Fig. 19.13. (**Note** relatively little smoothing was applied because using a wider bandwidth masks the initial high hazard.) The initial daily hazard of death in this 'baseline' person was about 0.0015 (0.15% per day) but, by approximately one year after initial hospitalisation, it had dropped to approximately 0.0003 (0.03% per day), where it remained. It is important to note that this reflects the probability of death among the pool of individuals remaining alive at any specific time. It does not necessarily indicate that the hazard

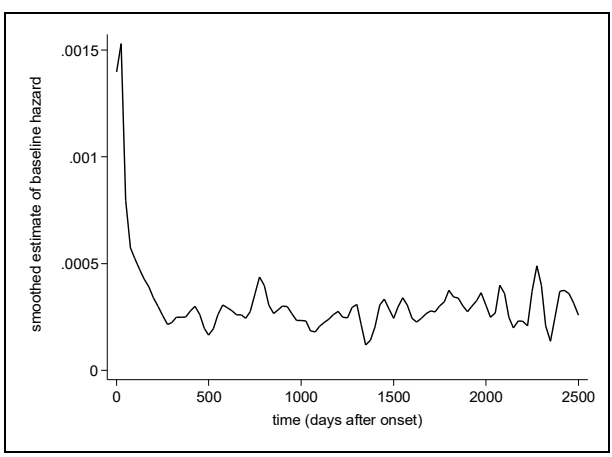

Fig. 19.13 Smoothed estimate of baseline hazard

for an individual remains constant after one year. This relatively constant hazard is being observed in a population that is increasingly made up of individuals unlikely to 'fail' (die). This issue of the nature of the population changing is discussed further in Section 19.11 (frailty models).

19.8.6 Model-building

In general, model-building procedures for Cox models are similar to those used for other regression-type models. Wald tests and likelihood ratio tests can be used to evaluate the significance of individual predictors or groups of predictors. Confounding and interaction can be assessed using methods presented for other regression-type models. Because the explanatory variables are related to the logarithm of the hazard ratio, it follows that interaction will be assessed on a multiplicative scale. There are, however, 2 issues that are specific to survival models: **stratified analysis** to allow for different baseline hazards in different groups of individuals in the study, and the possibility of including **time-varying covariates**.

19.8.7 Stratified analysis

Although we made no assumption about the shape of the baseline hazard, we have assumed that it is appropriate for an individual with all X_j=0. Let's consider the effect of gender on the hazard of death in the MI data. As we obtained a significant *HR* for -sex-, we would assume that it multiplies the h_0 by the *HR* and that this effect was constant over time. If we had reason to believe that the shape of the hazard was different in males compared with females, we could stratify the analysis on -sex- and allow for separate estimates of the baseline hazard in each group.

In a stratified Cox model, different baseline hazards ($h_{0j}(t)$) are assumed across groups of individuals to yield the following hazard function for the j^{th} group.

$$h_j(t) = h_{0j}(t)\, e^{\beta X} \qquad \textit{Eq. 19.16}$$

The difference from the unstratified model (Eq 19.13) is only in the baseline hazards whereas the regression term $e^{\beta X}$ is unchanged. Thus, the effects of predictors on *HR*s relative to the baseline hazard are assumed to be equal across all strata. Stratum-level predictors cannot be assessed in a stratified model because their effects will be absorbed in the baseline hazards. However, you can include interactions between a covariate (*eg* -mar_c2-) and a stratifying variable (*eg* -sex-). Example 19.7 shows a stratified (by sex) analysis of the -mi- data with an -mar_c2- by -sex- interaction included. (**Note** Stratified analyses provide one means of dealing with clustered data—by stratifying on the clustering variable. Dealing with clustered data is discussed further in Section 19.11.)

19.8.8 Time-varying covariates

Until now, we have focused on exposure factors that do not change their value over time and we have assumed that the effect of a factor was constant over time (proportional hazards assumption). However, survival analysis gives us the opportunity to relax both of these conditions. The terminology used with time-varying covariates may be confusing, so we will distinguish between **time-varying predictors** and **time-varying effects**.

Example 19.7 Stratified Cox proportional hazards model
data = mi

A stratified (by sex) model was fit with a sex by marital status (-mar_c2-) term included.

No. of subjects = 2851
No. of failures = 1480 Number of obs = 2851
Time at risk = 3669144 LR $\chi^2_{(5)}$ = 1252.78
Log likelihood = -9419.3796 Prob > χ^2 = 0.0000

Predictor	Coef	SE	Z	P	95% CI	
mar_c2	-0.109	0.082	-1.32	0.186	-0.270	0.052
mar*sex	-0.191	0.112	-1.70	0.089	-0.410	0.029
age	0.051	0.002	20.46	0.000	0.046	0.056
card	1.746	0.090	19.49	0.000	1.570	1.921
ptca	-0.925	0.063	-14.76	0.000	-1.047	-0.802

The main effect of -mar_c2-, which is now the effect of -mar_c2- in females (sex=0), is completely non-significant. The interaction terms produce a P-value of 0.089, which suggests that there is some evidence that marital status may have different effects in males compared with females.

Given the long-term nature of many survival studies, it is conceivable that the values of some of those predictors might change over time. These are time-varying predictors. For example, in the heart attack data, the location of the patient (in the hospital vs discharged) changed.

On the other hand, a predictor may remain constant, but its effect may change over time. These are time-varying effects. For example, having a cardiac arrest in the hospital might be constant for virtually the entire follow-up period (assuming that arrests occurred soon after admission), but the effect of that arrest on the probability of survival might change over time. It would be reasonable to assume that it would greatly increase the hazard of death in the time period immediately after its occurrence, but that this negative effect might be reduced over time. If this is true, the assumption of proportional hazards is violated.

Time-varying predictors
A time-varying predictor representing the location of the patient (-disch-=in hospital vs discharged) can be computed and added to data that have been restructured to allow for multiple time periods for each individual. (In this case there are 2 time periods: in the hospital vs discharged. Patients who die in hospital will have only 1 record.) This predictor can then be used to determine how a patient's hazard of dying changes when they are discharged from hospital. Example 19.8 shows both the data restructuring and the Cox model for the evaluation of -disch- as a time-varying predictor.

Events that occurred during hospitalisation (*eg* cardiac arrest, cardiac catheterisation, coronary angioplasty) could be analysed in a similar manner to study their impact on survival while hospitalised. For example, the effect of coronary angioplasty on survival (until discharge) could be evaluated by splitting the time in the hospital into 2 periods (pre- and post-angioplasty) for patients who had the procedure. No change would be made to the records for patients who did not have angioplasty. The effect of angioplasty on time to death or discharge (which would now be the censoring event) could then be evaluated.

Example 19.8 Time-varying predictor
data = mi

Data from the first 3 patients in the heart attack dataset are as follows.

		observation period		
patient	time in hospital	start	end	outcome
1	4	0	2918	0 = censored
2	5	0	2911	0 = censored
3	2	0	921	1 = died

Patients 1, 2, and 3 were hospitalised for 4, 5, and 2 days, respectively. All were discharged. Patients 1 and 2 were then followed for 2,914 and 2,906 additional days without dying from a heart attack. Patient 3 died 919 days after discharge (921 after admission). The data were restructured so each patient had 2 records, one for the time period in the hospital and one for the period after discharge. These data are as follows.

		observation period		
patient	start	end	outcome	location
1	0	4	0 = censored	0 = hosp
1	4	2918	0 = censored	1 = discharged
2	0	5	0 = censored	0 = hosp
2	5	2911	0 = censored	1 = discharged
3	0	2	0 = censored	0 = hosp
3	2	921	1 = died	1 = discharged

A Cox model fit to these data with the single predictor -disch- (*ie* patient had been discharged) produces:

No. of subjects = 2931
No. of failures = 1522 Number of obs = 5582
Time at risk = 3789652 LR $\chi^2_{(1)}$ = 244.60
Log likelihood = -11336.554 Prob > χ^2 = 0.0000

Predictor	HR	SE	Z	P	95% CI	
disch	0.094	0.014	-15.88	0.000	0.070	0.126

Although it appears that there were 5,582 observations, the number of subjects is correctly identified as 2,931. Once a patient was discharged, their hazard of death from a heart attack was reduced by a factor of 0.09 (*ie* less than 10% of what it was prior to discharge). This is not surprising given that only stable patients would be discharged.

Time-varying effects

A time-varying effect represents an interaction between a predictor and time (the effect of the predictor depends on what time point you are looking at). Effects may change at discrete points in time or may change continuously over time. A continually changing effect may change in a

linear manner with time (*eg* effect drops by a given amount every 100 days), with ln time (*eg* effect drops by a given amount for every one ln unit increase in time (equivalent to every 2.72-fold increase in time), or with any other function of time. Evaluating how effects may or may not change over time is an important part of validating a Cox proportional hazards model and is discussed further in Section 19.8.10.

One approach to evaluating how effects of predictors change over time is to fit an Aalen's linear hazards model (Hosmer and Royston, 2002). This model plots a cumulative regression coefficient for a predictor against time. If the effect of the predictor remains constant, the cumulative predictor will be expected to increase (or decrease) in a straight line over time. In general this is true, although some curvature to this line has been observed even when hazards are proportional. Example 19.9 shows Aalen's linear hazards model applied to an evaluation of the effect of coronary angioplasty on survival in the heart attack data.

19.8.9 Validating the model

Validation of a Cox proportional hazards model will be covered in the following 6 sections. The components in the validation process include:

- evaluating the proportional hazards assumption (Section 19.8.10)
- evaluating the assumption of independent censoring (Section 19.8.11)
- evaluating the overall fit of the model (Section 19.8.12)
- evaluating the functional form of predictors (Section 19.8.13)
- checking for outliers (Section 19.8.14)
- detecting influential points (Section 19.8.15).

19.8.10 Evaluating the assumption of proportional hazards

There are 3 general ways of evaluating the assumption of proportional hazards:

- graphical assessment
- the use of time-varying effects
- statistical assessment using Schoenfeld residuals.

Example 19.9 Aalen's linear hazards model
data = mi

Aalen's linear hazards model was fit to the heart attack data with coronary angioplasty (-ptca-) as the sole predictor, and the cumulative coefficient for -ptca- was plotted against time (up to 2,000 days).

There was initially a very strong beneficial effect of surgery (negative coefficient indicates reduced hazard of death). This beneficial effect persists, but becomes much less pronounced (the change in the cumulative coefficient between 1,000 and 1,500 days is much less than between 0 and 500 days).

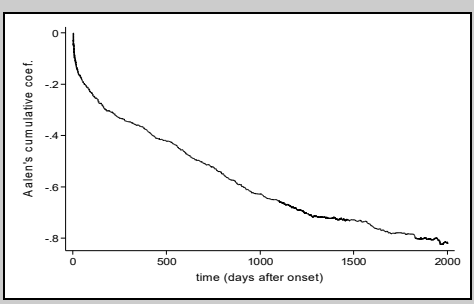

Fig. 19.14 Aalen's linear hazards model for coronary angioplasty (-ptca-)

Graphical assessment

For a categorical predictor, the assumption of proportional hazards can be tested by examining the log-cumulative hazard plot (ln$H(t)$ vs ln$t$) to check if the curves for the 2 (or more) study groups are parallel. If they are not parallel, then the assumption has been violated.

Fig. 19.15 shows a log cumulative hazard plot for the effect of cardiac arrest in the heart attack data. (**Note** It is actually the -log of the cumulative hazard that has been plotted which explains why the curves slope down (instead of up) as the cumulative hazard rises.) It is clear that the curves are not parallel, suggesting that the proportional hazards assumption has been violated. This seems reasonable because we would expect having a cardiac arrest to have a pronounced effect in the short term, but a reduced effect as time passed.

Another approach to graphical assessment is to compare plots of predicted survival times from a Cox model (which assumes proportional hazards) with Kaplan-Meier survivor function plots (which make no such assumption). If the 2 sets of curves are close together, it suggests that the proportional hazards assumption has not been violated. Fig. 19.16 shows such a plot. Clearly, the predicted values from the Cox model for patients who had a cardiac arrest are not at all close to the observed values (K-M curve). A limitation to graphical assessment is that it is limited to evaluating unconditional associations or situations in which the predictor being evaluated is clearly the strongest predictor in a multivariable setting.

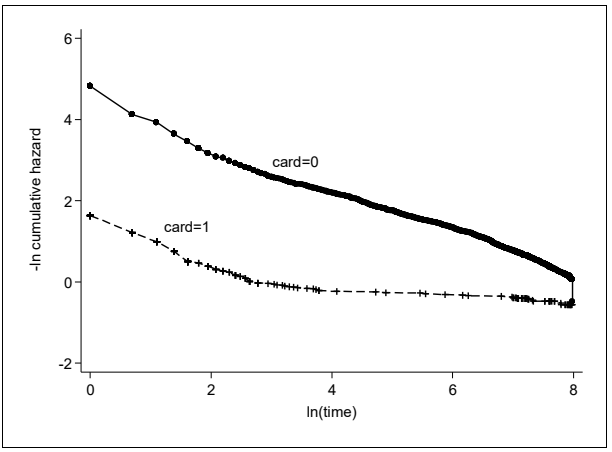

Fig. 19.15 Log cumulative hazard plot

Note It is actually the -log of the cumulative hazard that has been plotted

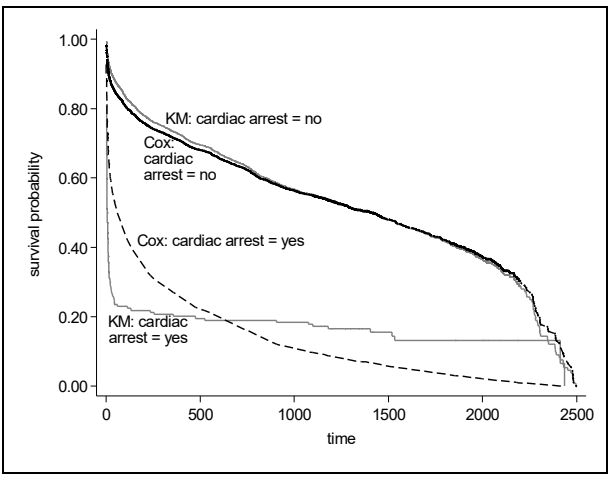

Fig. 19.16 Kaplan-Meier Cox plot

Time-varying effects

A term for the interaction between the surgery and time (or the log of the survival time) can be added to the model. The effect of surgery can be allowed to interact with time in a linear fashion or with ln(time) (or any other function of time for that matter). The advantage of adding a predictor*time interaction term is that if the assumption of proportional hazards is violated, the addition of the interaction term can solve the problem (provided the change in effect over time can be appropriately modelled).

In Example 19.10, a Cox model has been fit in which the effect of cardiac arrest (-card-) is allowed to vary with ln(time). The negative effect of having a cardiac arrest disappears by about day 400 and the effect then becomes beneficial (HR<1). No beneficial effect was expected beyond day 400 so allowing the effect to decay linearly with ln(time) may not be adequate. Some special procedures for integrating the use of fractional polynomials into the fitting of time-varying effects are discussed in Royston and Sauerbrei (2008).

Example 19.10 Assessing proportional hazards assumption—time-varying covariates
data = mi

A Cox model with a single predictor (-card-) was fit but the effect of having a cardiac arrest was allowed to interact with time on a natural log scale. This was chosen because it was assumed that the effect of an arrest would drop off rapidly soon after the attack and then more slowly as time went on (instead of a linear, or straight-line, decay in effect).

No. of subjects = 2931
No. of failures = 1522 Number of obs = 2931
Time at risk = 3789652 LR $\chi^2_{(2)}$ = 387.09
Log likelihood = -11265.31 Prob > χ^2 = 0.0000

Predictor	HR	SE	Z	P	95% CI	
main effect						
card	3.437	0.164	21.00	0.000	3.116	3.758
ln(time) interaction effect						
card	-0.572	0.049	-11.79	0.000	-0.668	-0.477

Both the initial effect (labelled 'main effect') of -card- and the time interaction were highly significant predictors of time to death. The significance of the ln(time) interaction term confirms that the effect of a cardiac arrest does vary with time (*ie* the proportional hazards assumption does not hold). In the presence of interaction, the effect of a cardiac arrest can be better understood by computing the HR at a number of time points. The HR at time t is $\exp\{3.437\text{-}(0.572*\ln(t))\}$.

Time (days)	ln(time)	HR
1	0	31.1
7	2	9.9
55	4	3.1
403	6	1.0
2981	8	0.3

While a cardiac arrest dramatically increases the hazard of death when it occurs, the effect drops off with time and is completely gone by 403 days (*ie* if you had a cardiac arrest while in hospital, the apparent negative effect of that attack is gone approximately one year after discharge). (**Note** a model with -card- interacting with ln(time) had a much higher log likelihood (indicating a better fit) than one with an interaction with time on a linear scale (data not shown).) Other functions of time might also be considered, although a model with -card- interacting with both log(time and log($time^2$) did not have significantly better fit (data not shown).

Schoenfeld residuals

Schoenfeld and **scaled Schoenfeld** residuals are based on the contribution that an observation makes to the partial derivative of the log partial likelihood. Hence they are also sometimes called 'partial residuals'. There is a separate set of residuals for each regression coefficient in the model, each set corresponding to the partial derivative for that parameter. These residuals are only computed at observed survival times. Scaled Schoenfeld residuals are adjusted using an estimate of the variance of the residual and these are better for detecting departures from the assumed model.

A graph of the scaled Schoenfeld residuals for a given predictor, when plotted against time (or ln(time)) can provide a graphical assessment of the proportional hazards assumption. This is particularly useful for continuous predictors because the log cumulative hazard plot is not useful for those variables. This graphical assessment can be enhanced by adding a smoothing line to indicate the overall trend. The residuals should hover around the 'zero' line, indicating no trend in the residuals over time. If the residuals trend up or down, it suggests that the effect of the predictor is varying over time. Fig. 19.17 (in Example 19.11) shows a plot of the scaled Schoenfeld residuals for age against ln(time). The assumption of proportional hazards appears to be reasonable for this predictor.

Schoenfeld residuals also form the basis of a statistical test of the assumption of proportional hazards. The test checks for a non-zero slope of the scaled Schoenfeld residuals against time (or a function of time) using a generalised linear regression. It provides an overall assessment and a test for each predictor separately. Results of this test for the heart attack data are presented in Example 19.11. The results clearly indicate that time-varying effects need to be added to the model. (Despite this, for the sake of simplicity, all future models will not include time-varying effects.)

19.8.11 Evaluating the assumption of independent censoring

One of the fundamental assumptions of survival models is that censoring is independent of the outcome of interest. This means that censored individuals should have the same future survival expectation as non-censored individuals (*ie* if the individuals were not censored, they would have the same survival distribution as the non-censored individuals). There are no specific tests to evaluate the independence of censoring and the event of interest. However, sensitivity analyses can be used to look at the extreme situations of complete positive or negative correlations between censoring and the event of interest.

Complete **positive correlation** would mean that every individual that was censored would have experienced the event of interest immediately if it had not been censored. This could be evaluated by refitting the model after recoding all of the censored observations so that they had the event of interest instead of being censored (at the time of censoring).

Complete **negative correlation** would mean that every individual that was censored would be guaranteed a long 'event-free' existence if it had not been censored. This could be evaluated by refitting the model after changing each censored individual's time at risk to a large, but plausible, value.

The above 2 analyses would provide the possible range of values that the coefficients of the factors of interest could possibly take if the assumption of independent censoring was badly violated. If gross violation of this assumption does not drastically alter the estimates of the

Example 19.11 Assessing the proportional hazards assumption—Schoenfeld residuals
data = mi

A Cox model with sex, age, marital status, cardiac arrest, and coronary angioplasty as predictors was fit to the heart attack data (without any time-varying covariates). Schoenfeld and scaled Schoenfeld residuals were obtained.

Fig. 19.17 shows a smoothed plot of scaled Schoenfeld residuals for age plotted against time on a log scale. It is difficult to detect any evidence of non-proportionality from this graph.

The statistical test for non-zero slope for each of the predictors (against ln(time)) resulted in the following.

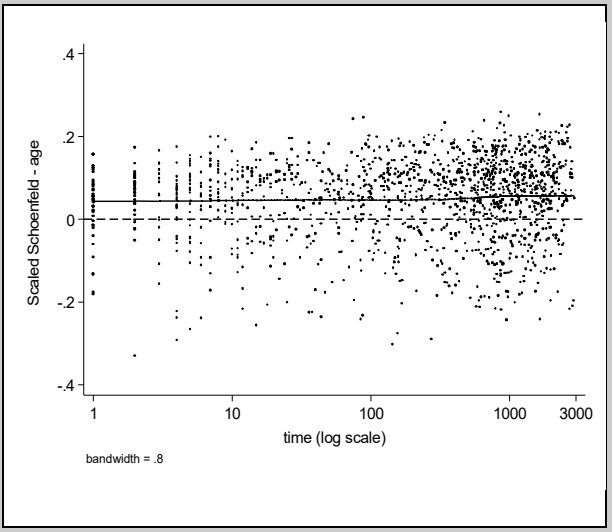

Fig. 19.17 Scaled Schoenfeld residuals for age

Variable	χ^2	df	Prob >χ^2
sex	0.97	1	0.3237
age	9.93	1	0.0016
mar_c2	3.86	1	0.0495
card	192.37	1	0.0000
ptca	25.59	1	0.0000
global test	211.98	5	0.0000

The global test was highly significant, as were the tests for three of the predictors (-age-, -card-, and -ptca-). It is clear that the assumption of proportional hazards was violated. Given the relatively large sample size in these analyses, statistical tests are more able to detect departures from non-proportionality than graphical procedures.

Note Although there is strong evidence of violation of the assumption of non-proportionality, for the sake of simplicity, models without time-varying effects will be used in all subsequent examples.

parameters of interest, you can be confident that the actual bias in the parameter estimates will be small.

Example 19.12 presents the results of a sensitivity analysis designed to evaluate this assumption in the heart attack data.

Example 19.12 Evaluating the assumption of independence of censoring
data = mi

A Cox model with sex, age, marital status, cardiac arrest, and coronary angioplasty as predictors was fit to the heart attack data. The model was then refit assuming complete positive and complete negative correlations between censoring and death (see text for description of method). Negative correlation was based on assigning 3,000 days as the time to death for all censored individuals. The results are summarised in the table below.

Variable	Original estimate	Assuming complete positive correlation	Assuming complete negative correlation
sex	0.200	0.092	0.107
age	0.051	0.023	0.023
mar_c2	-0.209	-0.131	-0.145
card	1.751	1.094	1.049
ptca	-0.924	-0.404	-0.497

Both sensitivity analyses resulted in substantial attenuation of all coefficients. However, all effects remained in the same direction and remained significant (P-values not shown). Given that the sensitivity analysis represents 2 extreme (and completely improbable) situations, the effect of censoring on the estimates is not a large concern, although methods to follow up some censored observations to determine their reason for censoring would be warranted.

19.8.12 Evaluating the overall fit of the model

Four approaches to evaluating the overall fit and predictive ability of the model are:

- to evaluate the distribution of the Cox-Snell residuals graphically
- to use a goodness-of-fit test similar to the Hosmer-Lemeshow test used for logistic regression
- to evaluate concordance between the predicted and observed sequence of pairs of events
- to compute an overall r^2 statistic.

Cox-Snell residuals are the estimated cumulative hazards for individuals at their failure (or censoring) times. If the model is appropriate, these residuals are a censored sample from a unit exponential distribution (*ie* an exponential distribution with a mean of one and variance of 1). Consequently, the range of these residuals is zero to $+\infty$. Cox-Snell (CS) residuals can be used to assess the overall fit of a proportional hazards model by graphically assessing how close these residuals are to having a unit exponential distribution. To do this, you:

- compute the CS residual
- fit a new proportional hazards model with the CS residuals used as the 'time' variable (along with the original censoring variable)
- derive an estimate of the cumulative hazard function ($H(t)$) from this new model
- plot $H(t)$ against the CS residuals.

If the residuals have a unit exponential distribution, the cumulative hazard should be a straight line with an intercept of 0 and a slope of 1. In practise, these graphs have been of limited value. Assessment of the linearity of the graph is a subjective procedure and substantial apparent

departures from the 45° line can result from a few observations with long survival times (when most of the observations are clustered at the lower left end of the line) (see Example 19.13).

For censored observations, the estimated cumulative hazard is an underestimate of the true cumulative hazard for an individual (by virtue of the fact that we don't observe them for the full period until they have the outcome of interest). Consequently, Cox-Snell residuals are sometimes modified by the addition of a constant (either 1 or ln(2)=0.693) for censored observations. There is no evident rationale for choosing one adjustment over the other, but this is only important if a substantial proportion of the observations are censored.

Several **goodness-of-fit** tests similar to a Hosmer-Lemeshow test for logistic regression models can be computed (May and Hosmer, 2004a). These tests all divide the data into groups and add indicator variables for these groups to the models and assess the overall significance of the indicator variables, with significance indicating evidence of lack of fit. An omnibus test designed to detect all causes of lack of fit was proposed by Grønnesby and Borgan (1996). The observed number of failures in groups defined by quantiles of risk from the fitted model are compared with the expected number of failures which are based on martingale residuals (see

Example 19.13 Evaluating overall fit of a model
data = mi

A Cox proportional hazards model was fit to the data with fixed effects for sex, age, marital status, cardiac arrest, and coronary angioplasty.

Cox-Snell residuals were computed and plotted (Fig. 19.18), as described in the text. It appears that there is relatively good agreement between the plotted values and the expected (45°) line.

Goodness-of-fit tests
The Grønnesby and Borgan omnibus goodness-of-fit test produces a P-value of 0.20 (no evidence of lack of fit), despite the fact that we know that the assumption of proportional hazards has been seriously violated. The Moreau, O'Quigley and Mesbah test, designed specifically for detecting non-proportional hazards, fails to produce a result with this model as a result of collinearity among predictors.

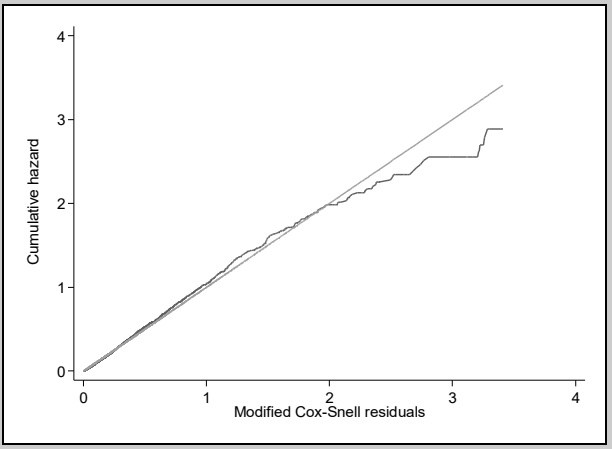

Fig. 19.18 Plot of Cox-Snell residuals

Concordance
Harrell's C statistic was 0.76 indicating that the model correctly predicts the sequence of 2 observed failures 76% of the time (*ie* reasonable predictive ability).

***r*²**
The r^2 for the Cox model with sex, age, marital status, cardiac arrest, and coronary angioplasty as predictors produced an estimated r^2 of 0.46 (46%) with a bootstrapped 95% confidence interval of (0.42, 0.51). Collectively these predictors have moderate ability to predict time to death for heart attack patients.

Section 19.8.13). However, the validity of the test depends on choosing an appropriate number of groups (May and Hosmer, 2004b). The number of groups should roughly equal the number of failure events in the data divided by 40, with a minimum of 2 and a maximum of 10. Using this strategy the test has reasonable power provided the sample size is greater than 200 with no more than 50% censoring (in smaller samples, the power is low). However, this test fails to identify the problem of non-proportional hazards in the heart attack data (Example 19.13). An alternative test designed specifically to evaluate the proportional hazards assumption was proposed by Moreau *et al* (1985). It requires the computation of time-dependent indicator variables and it successfully detects the problem of non-proportional hazards. These tests should not be used in situations in which there are time-varying covariates in the model.

Closely related to the issue of evaluating overall fit is the question of evaluating overall predictive ability. **Harrell's C concordance statistic** computes the proportion of all pairs of subjects in which the model correctly predicts the sequence of events (*ie* which one would have come first). It ranges from 0 to 1 with a value of 0.5 indicating no predictive ability at all (you would expect to get 50% correct by chance alone).

For a linear regression model, we would use $\boldsymbol{r^2}$ as a measure of predictive ability. Recently, Royston (2006) described several possible measures of explained variation for survival models and proposed an r^2 statistic for proportional hazard models. Comparable to the adjusted r^2 from linear regression, it is also possible to adjust the proposed r^2 for the number of predictors in a survival model. The r^2 compares a fitted model with a null model and provides an estimate of the amount of variation in survival times that is explained by the predictors. However, it cannot be used to compare models with different hazard structures (*eg* a semi-parametric Cox model with a parametric Weibull model—see Section 19.9 for hazard structures in parametric models) because the null models are different. An estimate of the r^2 for the heart attack data, and bootstrap 95% confidence intervals are shown in Example 19.13.

19.8.13 Evaluating the functional form of predictors

Martingale residuals can be used to evaluate the functional form of the relationship between a continuous predictor and the survival expectation for individuals. These residuals represent the difference between the observed final outcome for an individual and the cumulative hazard for that individual at the final point in time. (As such, they are more like typical residuals which represent a difference between an observed and a predicted value.) Because they are based on the estimated cumulative hazard, these residuals are similar to Cox-Snell residuals except their range is from -∞ to 1. The values of these martingale residuals are:

- uncensored observation i: $1-\hat{H}_i(t_i)$
- censored observation i: $0-\hat{H}_i(t_i)$

Consequently, residuals will be negative for all censored observations and for observations in which $H(t_i)>1$ (equivalent to $S(t_i)<0.37$).

To check for the functional form of continuous predictors, martingale residuals should be computed from a model that does not include the continuous predictor of interest. These residuals are then plotted against the predictor. A smoothing function (*eg* lowess smoothing) can be used to better visualise the relationship. If the relationship is linear, the smoothed martingale residual line should be approximately straight. Fig. 19.19 (Example 19.14) shows a lowess smoothed graph of martingale residuals against age.

Example 19.14 Evaluating functional form of predictors
data = mi

Fig. 19.19 shows a lowess smoothed graph of martingale residuals against age. It appears that a linear relationship may not be appropriate. Fractional polynomials (see Section 15.6.4) were used to further evaluate this possibility, and it was determined that a log-transformed version of age would provide a better fit. A model using log(age) had a significantly higher log likelihood (-10442.75) than one based on age (log L=-10446.45). Consequently, -age- will be included as log(age) in all subsequent models.

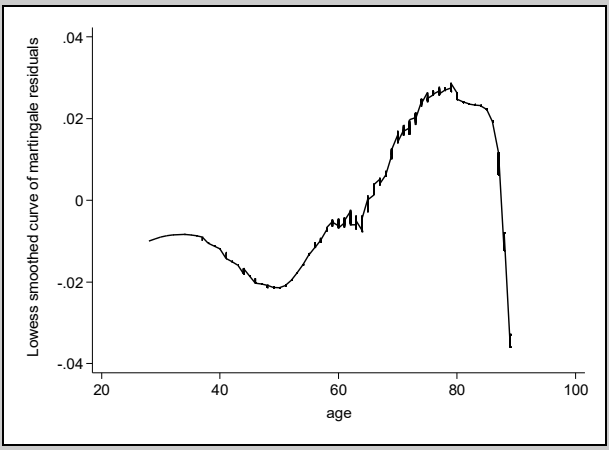

Fig. 19.19 Martingale residuals vs lactation number

The apparent large drop in the residuals at the right side of the graph is probably an aberrant 'end effect' resulting from a small number of data points with high age values. (There is no reason to believe that the hazard of death drops dramatically as age approaches 90.)

19.8.14 Checking for outliers

Deviance residuals can be used to identify outliers (*ie* points that are not well fit by the model). Deviance residuals are martingale residuals that have been rescaled so they are symmetric around 0 (if the fitted model is appropriate). The sum of the squared deviance residuals is the deviance (D) of the model.

If plotted with an observation number as the plotting symbol, they can be used to identify outlying observations. Fig. 19.20 is a plot of deviance residuals from the model with sex, log(age), marital status, cardiac arrest, and coronary angioplasty as predictors. The residuals with the largest absolute value (>3.5 or <-3.5) are identified on the graph. These were individuals who 'beat the odds'. They were all elderly, had experienced a cardiac arrest when hospitalised, but had not had coronary angioplasty. Despite this, they had relatively long survival periods.

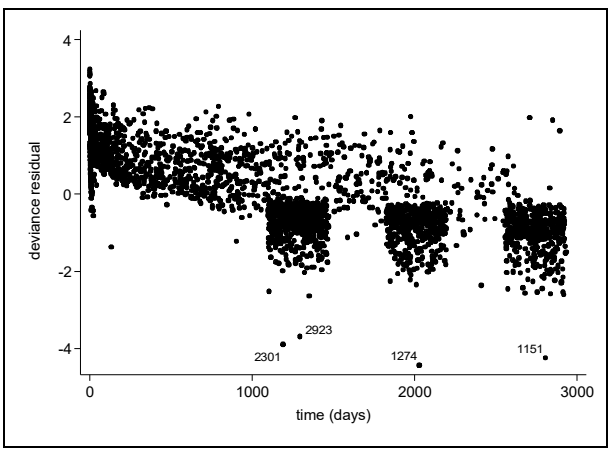

Fig. 19.20 Deviance residuals

19.8.15 Detecting influential points

Score residuals and scaled score residuals can be used to identify influential observations. The former have a 'leverage-like' property while the latter measure the impact of observations on coefficients in the model.

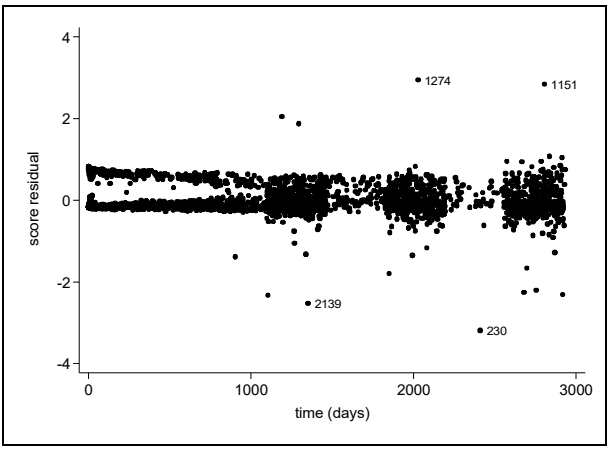

Fig. 19.21 Score residuals

Score residuals are a variation of martingale residuals but are computed for each predictor (covariate) in the model. They have a 'leverage-like' property in that observations that are far from the mean of the predictor have larger (positive or negative) residuals. When plotted against time, they typically form a 'fan-shaped' pattern (with the centre of the fan at the mean of the time variable) and observations lying outside this fan should be considered as potentially influential. Fig. 19.21 shows score residuals for coronary angioplasty (-ptca-).

Score residuals can be modified to compute a **delta-beta** like parameter for coefficients in the model. This modification involves multiplying the score residual by the estimated variance of the coefficient (from the variance-covariance matrix of the coefficients) and produces what is called a **scaled score residual**.

Fig. 19.22 shows a plot of the scaled score residuals for coronary angioplasty (-ptca-) against time. Elimination of the 4 residuals with the largest absolute values produced a substantial (14%) increase in the coefficient for -card- suggesting that the estimate for this parameter may not be very robust.

19.9 PARAMETRIC MODELS

As noted previously, Cox proportional hazards models make no assumption about the shape of the baseline hazard, which can be a real advantage if you have no idea what that shape might be, or if it has a very irregular form. However, these models achieve this flexibility at a price. Because they only use information about the observations at times at which one or more of the subjects fail, they do not efficiently use all of the information you have about the

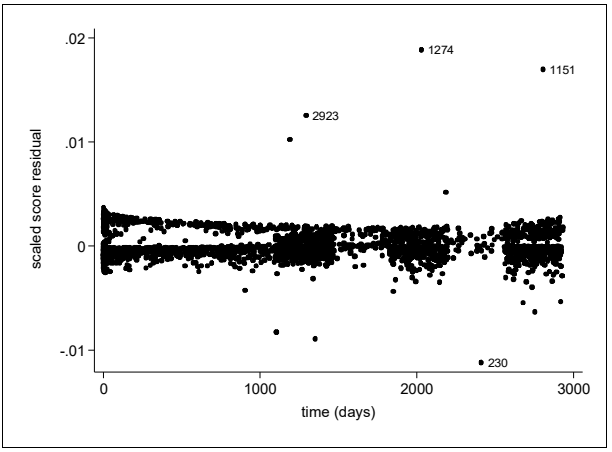

Fig. 19.22 Scaled score residuals (delta-beta)

observations. For example, because the Cox model is based solely on the rank ordering of the observations, it makes no difference if 2 successive failures are one day apart or one year apart. The length of the interval, which provides some valuable information in terms of survival times, is ignored. Consequently, if you can correctly specify the form of the baseline hazard, a parametric model will be more efficient (*ie* use more of the available information).

A parametric model satisfying the proportional hazards assumption could be written in the same way as a semi-parametric model:

$$h(t) = h_0(t)\mathrm{e}^{\beta X}$$

but $h_0(t)$ is assumed to have a specified functional form. The major difference between parametric and semi-parametric models is that βX now includes an intercept term (β_0). (An alternative method of writing these models is described in Section 19.10.)

Not all parametric models are proportional hazards models. (Models which are not are discussed in Section 19.10.) Three parametric models which are also proportional hazards models are the exponential, Weibull, and Gompertz. Each of these will be discussed briefly. When using any of these models, it must be kept in mind that in addition to specifying a correct function for the baseline hazard, the assumption of proportional hazards must also be evaluated and met.

19.9.1 Exponential model

An exponential model is the simplest form of parametric model in that it assumes that $h_0(t)$ is constant over time (*ie* in the baseline group, the rate at which failures are occurring remains constant). Consequently

$$h(t) = \lambda = c\left(\mathrm{e}^{\beta X}\right) \qquad \textit{Eq 19.17}$$

where c is the constant baseline hazard and λ is the time-constant value of $h(t)$ for any given set of predictor values. The density and survivor functions of the exponential distribution were given in Eq 19.11. As noted previously, the survival times will have a decreasing exponential distribution.

Interpretation of coefficients

Coefficients for predictors in an exponential model can be interpreted the same way as coefficients from a Cox model. The exponentiated coefficient is the **hazard ratio** (Section 19.8.2). The intercept in the model is the estimate of the log of the (constant) baseline hazard. In Example 19.15, an exponential model is fit to the heart attack data. If this model was appropriate (which it isn't—more on that later), the baseline hazard would be estimated to be $e^{-8.61}$=0.00018. That is, on any given day, a person in the baseline group (55-year-old, not-married female who had not had a cardiac arrest or coronary angioplasty) had approximately a 0.02% (1 in 5000) chance of dying on that day.

Evaluating the assumption of constant hazard

The assumption that the baseline hazard is constant over time can be evaluated in several ways. The first is to generate an estimate of the baseline hazard from a Cox model and plot it to see if it approximately follows a straight, horizontal line. Fig. 19.13 showed that the baseline hazard fell rapidly up to about one year and then levelled off. A second approach is to fit a model with a piecewise-constant baseline hazard (Cleves *et al*, 2008; Dohoo *et al*, 2003). In this case, the baseline hazard is allowed to vary across time intervals by including indicator variables for each

Example 19.15 Exponential regression
data = mi

An exponential survival model was fit to the heart attack data after log transforming -age- (and subtracting 4 so the 'baseline' individual is someone 55 years of age).

No. of subjects = 2851
No. of failures = 1480 — Number of obs = 2851
Time at risk = 3669144 — LR $\chi^2_{(5)}$ = 1663.70
Log likelihood = -4576.9189 — Prob > χ^2 = 0.0000

Predictor	Coef	SE	Z	P	95% CI	
sex	0.255	0.058	4.39	0.000	0.141	0.369
age_ln	4.093	0.185	22.18	0.000	3.731	4.454
mar_c2	-0.261	0.057	-4.59	0.000	-0.372	-0.149
card	2.061	0.089	23.16	0.000	1.887	2.236
ptca	-1.069	0.062	-17.18	0.000	-1.191	-0.947
constant	-8.612	0.088	-98.05	0.000	-8.784	-8.439

The *HR* for being married (mar_c2=1) would be $e^{-0.261} = 0.77$, suggesting that, at any given point in time, a married person was 0.77 times as likely to die as a not-married person (if this model was correct).

of the time intervals in the model. The baseline hazard is assumed to be constant within each time period, but can vary between time periods. This produces the results and step graph shown in Example 19.16. It is clear that the hazard falls rapidly over the first 30–60 days, and is relatively flat after day 120. However, a model which assumed that the hazard declined in a curved manner, might be a reasonable approximation. A third approach to evaluating the assumption of constant hazard is to evaluate the shape parameter from a Weibull model (see Section 19.9.2).

19.9.2 Weibull model

In a Weibull model, it is assumed that the baseline hazard function has a shape which gives rise to a Weibull distribution of survival times. The Weibull hazard was discussed in Section 19.7.7 and shown graphically in Fig. 19.10. In addition, Eq 19.12 gives the formulae for the hazard and survivor functions.

If a vector of covariates (predictors) is added to a Weibull model, the formula for the hazard function becomes:

$$h(t) = \lambda\, p t^{p-1} e^{\beta X} \qquad \textit{Eq 19.18}$$

where βX does not include an intercept term (β_0). Example 19.17 shows a Weibull model fit to the heart attack data. The estimate of the shape parameter (p) is 0.56 (95% CI: 0.53, 0.58) suggesting that the hazard is decreasing over time.

Evaluating the Weibull distribution
As was noted, the suitability of the assumption that the survival times follow a Weibull distribution can be assessed by generating a log-cumulative hazard plot. If the distribution is

Example 19.16 Piecewise constant exponential regression model
data = mi

A model which allows the baseline hazard to vary between time periods but forces it to remain constant within time periods is called a piecewise constant exponential model. Results from such a model and a graph of the resulting baseline hazard are shown below.

No. of subjects = 2851
No. of failures = 1480 Number of obs = 18955
Time at risk = 3669144 LR $\chi^2_{(13)}$ = 2766.26
Log likelihood = -4025.6408 Prob > χ^2 = 0.0000

Predictor	Coef	SE	Z	P	95% CI	
sex	0.195	0.058	3.37	0.001	0.082	0.308
age_ln	3.630	0.183	19.88	0.000	3.272	3.988
mar_c2	-0.223	0.057	-3.92	0.000	-0.335	-0.112
card	1.801	0.089	20.26	0.000	1.627	1.975
ptca	-0.936	0.062	-15.00	0.000	-1.058	-0.814
day0_15	2.708	0.112	24.12	0.000	2.488	2.928
day16_29	1.528	0.151	10.15	0.000	1.233	1.824
day30_59	0.825	0.154	5.36	0.000	0.523	1.126
day60_119	0.536	0.139	3.85	0.000	0.263	0.808
day120_479	-0.036	0.112	-0.32	0.748	-0.255	0.183
day480_839	-0.097	0.116	-0.84	0.402	-0.325	0.130
day840_1199	-0.221	0.124	-1.78	0.074	-0.463	0.022
day1200_1919	-0.126	0.117	-1.08	0.281	-0.356	0.103
constant	-8.685	0.121	-72.05	0.000	-8.922	-8.449

The coefficients for predictors day 0_15 through day 1200_1919 show how the log hazard changes relative to the value for the baseline time period (days 1920+). There was little evidence of any statistical difference between any of the periods from day 120 to 1919 compared with the baseline time period.

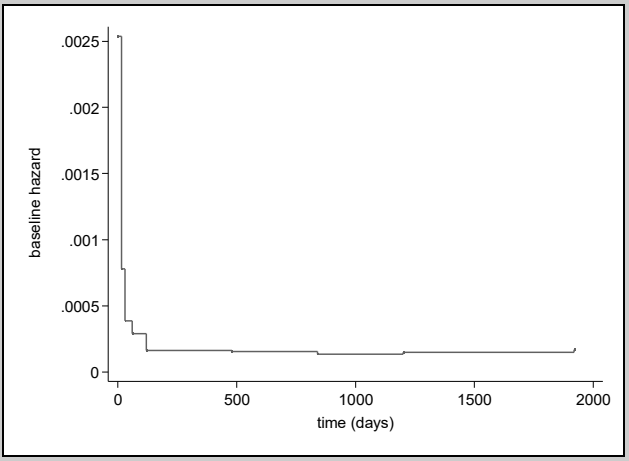

Fig. 19.23 Piecewise constant estimate of baseline hazard

Example 19.17 Weibull model
data = mi

A Weibull model was fit to the heart attack data.

No. of subjects = 2851
No. of failures = 1480 Number of obs = 2851
Time at risk = 3669144 LR $\chi^2_{(5)}$ = 1293.74
Log likelihood = -4128.7058 Prob > χ^2 = 0.0000

Predictor	Coef	SE	Z	P	95% CI	
sex	0.196	0.058	3.38	0.001	0.082	0.309
age_ln	3.489	0.180	19.36	0.000	3.136	3.843
mar_c2	-0.215	0.057	-3.77	0.000	-0.326	-0.103
card	1.835	0.089	20.55	0.000	1.660	2.010
ptca	-0.933	0.063	-14.90	0.000	-1.055	-0.810
constant	-5.228	0.126	-41.42	0.000	-5.475	-4.981
ln_p	-0.586	0.022	-26.68	0.000	-0.629	-0.543
p	0.556	0.012			0.533	0.581
1/p	1.797	0.039			1.721	1.876

With the exception of the exponential model (which is clearly inappropriate), the effects of the predictors are similar among the Cox, piecewise exponential, and Weibull models (and also similar in terms of statistical significance). The shape parameter (p) from the Weibull distribution indicates that the hazard is falling with time (*ie* p<1).

Weibull, this graph will show a straight line. A rough evaluation can be obtained by generating a simple plot of ln$H(t)$ vs ln(t) for all of the data. Fig. 19.15 shows a plot of -ln$H(t)$ vs ln(t) for each of the 2 cardiac arrest groups in the heart attack data. The baseline hazard will be included in the non-arrest group and that line was approximately straight suggesting that the Weibull model might be appropriate. The step graph of the baseline hazard (Fig. 19.23 in Example 19.16) also suggests that the Weibull model may be appropriate. A Weibull model with a shape parameter of 0.56 assumes the hazard monotonically decreases over time.

19.9.3 Gompertz model

The Gompertz model is used less frequently than the exponential and Weibull models but has been used to model mortality data. In a Gompertz model, the log of the baseline hazard varies linearly with time so the baseline hazard is as follows.

$$h_0(t) = \lambda e^{pt} \qquad \textit{Eq 19.19}$$

The baseline hazard increases exponentially if p>0 and decreases exponentially if p<0. If p=0, the hazard is constant at λ (exponential model). A Gompertz model fit to the heart attack data (results not shown) produces an estimate of p of -0.00061 (95% CI: -0.0007, -0.0005) which also suggests that the hazard is falling with time.

19.10 ACCELERATED FAILURE TIME MODELS

As noted above, not all parametric models are proportional hazards models. However, those that are can be written in 1 of 2 ways: as a proportional hazards model (which is what has been presented thus far), or as an accelerated failure time model (AFT). Other parametric models (discussed below) can only be written in the AFT metric, because the predictors in these models do not necessarily multiply the baseline hazard by a constant amount.

The general form of an AFT model is:

$$\ln t = \beta X + \ln \tau \qquad \text{or} \qquad t = e^{\beta X} \tau \qquad \textit{Eq 19.20}$$

where ln$t$ is the natural log of the time to the failure event, βX is a linear combination of explanatory variables and lnτ is an error term with an appropriate distribution. **Note** The values of the βs in this representation will not be the same as the βs in a proportional hazards representation.

From Eq 19.20 it can be seen that τ is the distribution of survival times when βX=0 (*ie* $e^{\beta X}$=1). τ is assumed to have a specific distribution (*eg* Weibull, log-normal). If τ has a log-normal distribution, then the log of survival times will have a normal distribution which is equivalent to fitting a linear model to ln(survival times) (assuming you can ignore the problem of dealing with censored observations). Three specific distributions of survival times (log-logistic, log-normal and generalised gamma) are discussed in Section 19.10.2.

Eq 19. 20 can be rearranged as follows:

$$\tau = e^{-\beta X} t \qquad \text{or} \qquad \ln(\tau) = -\beta X + \ln(t) \qquad \textit{Eq 19.21}$$

The linear combination of predictors in the model (βX) acts additively on log(time) or multiplicatively on time (*ie* they accelerate or decelerate the passage of time by a multiplicative factor) where $e^{-\beta X}$ is called the **acceleration parameter** because if:

- $e^{-\beta X}>1$, then $t<\tau$ so time passes more quickly (*ie* failures expected sooner)
- $e^{-\beta X}=1$, then $t=\tau$ so time passes at a 'normal' rate (*ie* no effect of predictors)
- $e^{-\beta X}<1$, then $t>\tau$ so time passes more slowly (*ie* failures expected later).

As indicated above, the exponential and Weibull models can be written either as proportional hazards models or as AFT models. The relationship between the coefficients from a proportional hazards expression (β_{ph}) of a Weibull model and an AFT expression (β_{aft}) is:

$$\beta_{aft} = \frac{-\beta_{ph}}{p} \qquad \textit{Eq 19.22}$$

where p is the shape parameter from the Weibull model.

19.10.1 Coefficients in AFT models

A coefficient in an AFT model represents the expected change in the ln(survival time) for a 1-unit change in the predictor. For example, assume you have a dichotomous predictor (X with a coefficient of 2). If, in the absence of X, a study subject is expected to fail at t=5 days (ln(t)=1.61), the presence of X would increase the expected ln(survival time) to 1.61+2=3.61 or the survival time to 37 days. The presence of X in a subject who was expected to survive 30

days would result in an increase expected survival time from 30 to 222 days. As you can see, in absolute time, factors have a greater impact at longer expected survival times.

An alternative interpretation is to exponentiate the coefficient to compute a **time ratio** (*TR*). A coefficient of 2 produces a *TR* of e^2=7.4 which means that the presence of *X* increases the expected survival time by a factor of approximately 7 times.

19.10.2 Specific survival time distributions

Log-logistic model

A log-logistic model assumes that survival times follow a log-logistic distribution, or alternatively, log survival times follow a logistic distribution (a symmetric distribution similar to a normal distribution). The hazard function for a log-logistic distribution is as follows.

$$h(t)=\frac{e^{\theta}}{\gamma t(t^{-1/\gamma}+e^{\theta})} \qquad \textit{Eq 19.23}$$

where γ>0 is a scale parameter. $h(t)$ decreases as a function of t if γ>1, otherwise it is increasing and then decreasing with a peak at:

$$t=\left(\frac{(1/\gamma)-1}{e^{\theta}}\right)^{\gamma}=e^{-\theta\gamma}((1/\gamma)-1)^{\gamma} \qquad \textit{Eq 19.24}$$

In the log-logistic model, $-\theta\gamma$ is modelled as a function of the predictors (*ie* $-\theta\gamma=\beta X$). The p^{th} percentile (and median) of a log-logistic distribution are the following.

$$t_p=\left(\frac{p}{100-p}\right)^{\gamma}e^{-\theta\gamma} \qquad t_{50}=e^{-\theta\gamma} \qquad \textit{Eq 19.25}$$

Fig. 19.24 shows hazard functions for various values of γ (left panel) and a histogram of log-logistic distributed survival times (based on 2000 simulated observations) when γ=0.25. (In all cases the median survival time is set to 20 days.)

A log-logistic survival model expressed in AFT terms is shown in Example 19.18. Because log-normal or log-logistic models can rise and then fall, one of these might be most appropriate if there is an initial rise in hazard immediately after admission to hospital. A simple life table based on the data from days 1–10 does not provide any clear evidence of this.

Log-normal model

In a log-normal model, the survival times are distributed normally on a log time scale, or alternatively, log times are distributed normally. The survivor function is:

$$S(t)=1-\Phi\left(\frac{\ln t-\mu}{\sigma}\right) \qquad \textit{Eq 19.26}$$

where Φ is the cumulative distribution function of a standard normal (Gaussian) distribution and μ and σ are the mean and standard deviation of log survival times. (The formulae for the hazard functions for the log-normal and generalised gamma distributions can be derived from $f(t)$ and $S(t)$, but are complex and beyond the scope of this text. See Cleves *et al* (2008); Collett (2003) for details.

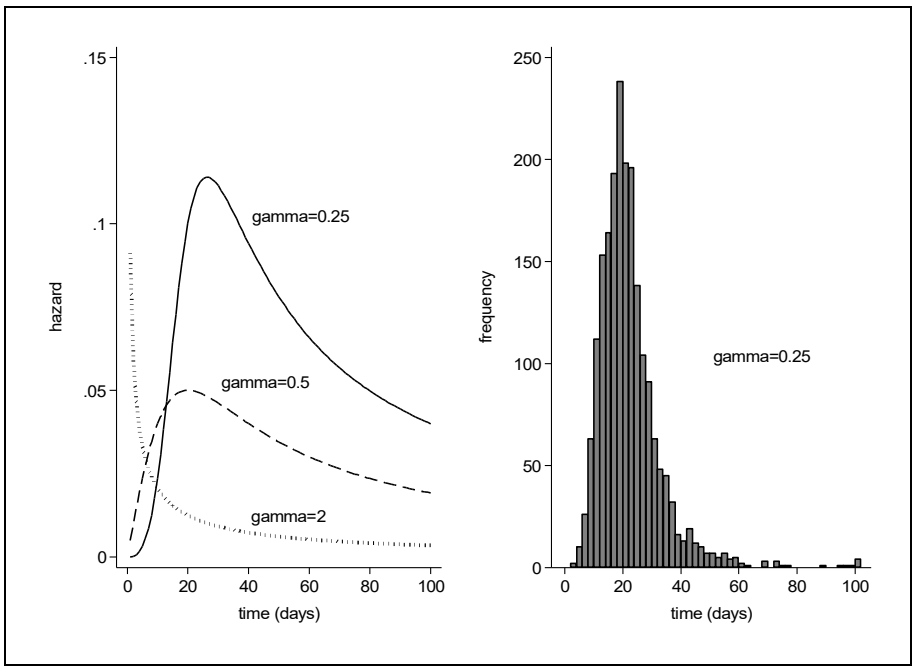

Fig. 19.24 Hazard functions (left) and survival times (right) for log-logistic distribution

Example 19.18 Log-logistic model of heart attack data
data = mi

A log-logistic model was fit to the heart attack data and produced the following.

Predictor	Coef	SE	Z	P	95% CI	
sex	-0.279	0.110	-2.53	0.011	-0.495	-0.063
age_ln	-6.002	0.323	-18.56	0.000	-6.635	-5.368
mar_c2	0.382	0.108	3.53	0.000	0.170	0.594
card	-4.465	0.187	-23.85	0.000	-4.832	-4.098
ptca	1.673	0.112	14.94	0.000	1.453	1.892
constant	8.559	0.153	56.03	0.000	8.259	8.858
ln_gamma	0.310	0.022	13.94	0.000	0.266	0.354
gamma	1.364	0.030			1.305	1.424

The time ratio for -mar_c2- was $e^{0.382}$=1.47 which suggests that, on average, time to death in married people was approximately 50% longer than in not-married people. Using Eq 19.25, the median survival time of the baseline group is:

$$t_{50} = e^{\text{constant}} = e^{8.559} = 5213 \text{ days}$$

while the corresponding value for similar individuals who had experienced a cardiac arrest was:

$$t_{50} = e^{8.559-4.465} = 60 \text{ days}$$

Generalised gamma model
A generalised gamma distribution is a 3-parameter (μ, κ, σ) distribution for which the hazard function can take a wide variety of shapes which include the Weibull, log-normal and gamma distributions. Consequently, it is particularly useful for evaluating the shape of the hazard function (see Section 19.10.3).

19.10.3 Choosing a parametric model

Selecting an appropriate parametric model involves both biological and statistical procedures. The selection should be guided by knowledge of how failures arise and insights into what would be expected in terms of a hazard function.

As noted, the generalised gamma distribution provides some insight into what might be an appropriate distribution.

- if κ=1, the distribution is Weibull and σ=1/p is the inverse shape parameter
- if κ=1 and σ=1, the distribution is exponential
- if κ=0, the distribution is log-normal.

For the heart attack data, a generalised gamma model produces estimates of κ=0.61 (95% CI: 0.47, 0.74,) and σ=2.07 (95% CI: 1.95, 2.20). Both are substantially different from 1, and their confidence intervals do not include 1. This suggests that a Weibull model may be inadequate. Example 19.19 shows the log-likelihood for each of the 5 parametric models, along with the number of distribution parameters and the point estimate of the effect of -mar_c2-.

Example 19.19 Comparison of parametric models
data = mi

Five parametric models were fit to the heart attack data and compared.

		# Parameters				Time ratio for
Model	Log L	Distribution	Predictors	df	AIC	-mar_c2-
exponential	-4576.92	1	5	6	9165.8	1.30
Weibull	-4128.71	2	5	7	8271.4	1.47
log-logistic	-4117.91	2	5	7	8249.8	1.47
log-normal	-4151.66	2	5	7	8317.3	1.40
generalised gamma	-4114.59	3	5	8	8245.2	1.47

While the generalised gamma model fits the best (largest log L and smallest AIC) the log-logistic may be a suitable alternative. The Gompertz model is not shown because a time ratio cannot be computed from this model, but it had a log L of -4479.2 and an AIC of 8972.4, suggesting it had a very poor fit.

The estimates of the time ratio for -mar_c2- were all quite similar except for the exponential model (which had the worst fit of all the models shown).

19.11 FRAILTY MODELS AND CLUSTERING

As noted in previous sections, predictors in survival models (semi-parametric and parametric) act multiplicatively on the baseline hazard (*ie* the hazard for an individual is a multiple of the baseline function). In a frailty model, an additional latent (unobserved) effect (*ie* the frailty) acts multiplicatively on the hazard. The frailty is not measured directly, but is assumed to have a specified distribution and the variance of the distribution is estimated from the data.

There are 2 general types of frailty model: individual frailty and shared frailty (Gutierrez, 2002). In an individual frailty model, the additional variance is unique to individuals and serves to account for additional variability in the hazard among individuals in much the same way that the negative binomial model accounts for more variability than a Poisson model. Shared frailty models constitute one approach to dealing with clustered data, and are discussed starting in Section 19.11.3.

19.11.1 Individual frailty models

Within a population described by an average hazard $h(t)$, some individuals fail early and some fail late. This variation in survival time may be attributed to 3 components. Part may be due to differences among individuals in terms of measured covariates and this variability will be removed by including those covariates in the model. Part may be due to unmeasured covariates which make some individuals more prone to fail early (*ie* 'frail' individuals). The final part is that attributable to random variation and is explained by the survival time distribution that is selected. The effect of frailty (unmeasured covariates) can be thought of as overdispersion—more variability in the survival times than would be expected based on the chosen distribution.

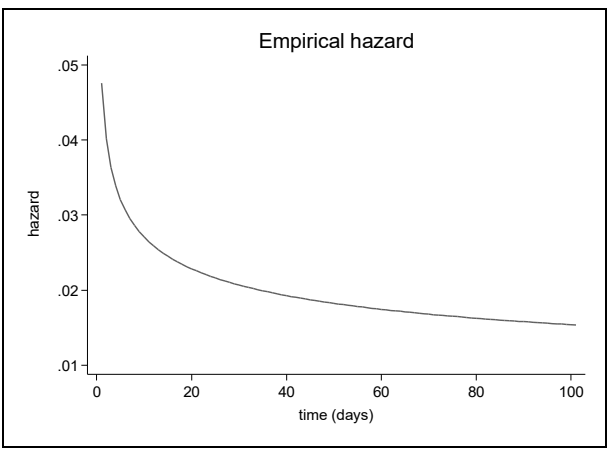

Fig. 19.25 Effects of individual frailties (see text)

The effect of individual frailty can be seen in Fig. 19.25 which shows the empirical hazard for a population of 2,000 individuals. All individuals in this population had a constant hazard set to 0.05 (exponential model), but individuals were assigned individual frailties (gamma distribution, μ=1, σ=1) which made some individuals more prone to fail than others. The graph shows the estimated hazard from a Weibull regression model (shape parameter p=0.75). Although every individual had a constant hazard, the average hazard for the population clearly falls as the frail individuals fail and the remaining population increasingly consists of more robust individuals.

An individual frailty model can be written as follows:

$$h(t|\alpha)=\alpha\, h(t) \qquad \textit{Eq 19.27}$$

Conditional on the frailty, the hazard at any point is multiplied by a factor (variable) α, which is

assumed to have a distribution with a mean of 1 and a variance of θ. Two commonly assumed distributions of α are the gamma and the inverse Gaussian.

A frailty effect can account for apparent changes in the hazard in a population over time. A Weibull model fit to the data used to create Fig. 19.25 has a shape parameter of 0.75 suggesting that the hazard is falling over time. If the model is refit with a gamma frailty added, the shape parameter changes to 1.3, suggesting that, for individuals with the same frailty, the hazard is actually rising. It is impossible to separate individual frailty effects from the distributional assumptions of the model, so in practice, individual frailties have limited applicability unless the expected distribution of survival times is known with certainty (O'Quigley and Stare, 2002).

Example 19.20 shows the addition of a gamma frailty to the Weibull model of the heart attack data.

The concept of individual frailty does not apply to Cox (semi-parametric) models because the frailty effect represents variation in survival times in excess of what would be expected from the assumed distribution of survival times. However, in a Cox model, there is no assumed

Example 19.20 Individual frailty model
data = mi

A Weibull model with a gamma individual frailty was fit to the heart attack data.

No. of subjects = 2851
No. of failures = 1480 Number of obs = 2851
Time at risk = 3669144 LR $\chi^2_{(5)}$ = 1338.68
Log likelihood = -4106.2367 Prob > χ^2 = 0.0000

Predictor	Coef	SE	Z	P	95% CI	
sex	0.198	0.070	2.83	0.005	0.061	0.336
age_ln	4.032	0.224	18.02	0.000	3.594	4.471
mar_c2	-0.250	0.069	-3.61	0.000	-0.385	-0.114
card	2.729	0.174	15.64	0.000	2.387	3.071
ptca	-1.096	0.076	-14.37	0.000	-1.245	-0.946
constant	-5.825	0.173	-33.74	0.000	-6.163	-5.486
ln_p	-0.430	0.033	-13.06	0.000	-0.495	-0.365
ln_theta	-0.747	0.193	-3.88	0.000	-1.124	-0.369
p	0.650	0.021			0.610	0.694
1/p	1.537	0.051			1.441	1.640
theta	0.474	0.091			0.325	0.691

The shape parameter for the Weibull model was 0.65, indicating the hazard fell over time. However, this estimate was somewhat higher than when individual frailties are not incorporated into the model (p=0.56). The variance of the gamma frailty (theta) was estimated to be 0.474. Part of the decline in the population frailty has been attributed to individual frailties (*ie* the population consisting of increasingly robust individuals as time moves on), so less is reflected in the shape parameter of the Weibull distribution.

distribution of survival times. Any 'overdispersion' would be incorporated into the baseline hazard (h_0) which has no specified form.

19.11.2 Clustering in survival data

Individuals within a group or cluster (*eg* patients within a hospital) have features in common (*eg* procedures for handling specific conditions, level of nursing care) that lead to a lack of independence among individuals within a cluster and could result in more similar survival times (*eg* patients treated at one hospital may have better survival probabilities than those from another). The general problem of clustering is covered in Chapters 20 through 24. However, in terms of survival models, there are several approaches that can be used to deal with clustered data. If the number of clusters are limited, fixed effects representing the clusters can be included in the model. Stratified models (Section 19.8.7), in which the strata are the clusters, can also be used to address the issue of clustering, but like the use of fixed effects models, preclude the evaluation of cluster level predictors. Robust standard errors (Chapter 20) are a general approach that can be used to address the problem of lack of independence in many types of model, but have some limitations (Lin and Wei, 1989). Shared frailty models are based on the assumption that groups of individuals within a cluster have a common frailty and model that frailty so they are analogous to random effects models (see Chapters 21 and 22).

19.11.3 Shared frailty models—introduction

Just as individual frailties can be considered to represent the effects of unmeasured covariates, shared frailties represent the effects of unmeasured covariates that a group of individuals have in common. These can represent the random effect of a grouping variable such as hospital. (See Chapters 20–24 for more discussion of random effects.) A shared frailty would be an appropriate way of dealing with the lack of independence observed when we have multiple failure times in an individual. (The frailty would represent the common characteristics of the individual that affect time to each event occurrence.)

A shared frailty model can be written as follows:

$$h_i(t|\alpha_i) = \alpha_i h(t) \qquad \textit{Eq 19.28}$$

where α_i represents the frailty for the i^{th} group (and $h_i(t)$ and $h(t)$ incorporate the effects of the predictors). The survival probability, conditional upon the frailty, is written:

$$S_i(t|\alpha_i) = S(t)^{\alpha_i} \qquad \textit{Eq 19.29}$$

Frailties can take on a variety of distributions, but the most commonly used ones are gamma, inverse Gaussian and positive stable distributions. The statistical significance of a frailty can be assessed with a likelihood ratio test, but the usual χ^2 reference statistic is not correct because variances cannot be less than 0, so the P-value should be cut in half.

Example 19.21 shows the results of fitting a Weibull model with a gamma-distributed shared frailty for hospital to the heart attack data.

Example 19.21 Shared frailty Weibull model
data = mi

A shared frailty model (Weibull distribution with a gamma distributed frailty common to all patients in a hospital) was fit to the -mi- data.

No. of subjects = 2851 — Obs per group min = 12
No. of groups = 10 — avg = 285.1
No. of failures = 1480 — max = 1118
Time at risk = 3669144 — LR $\chi^2_{(5)}$ = 1220.99
Log likelihood = -4122.4119 — Prob > chi2 = 0.0000

Predictor	Coef	SE	Z	P	95% CI	
sex	0.213	0.058	3.67	0.000	0.099	0.327
age_ln	3.384	0.182	18.63	0.000	3.028	3.741
mar_c2	-0.214	0.057	-3.76	0.000	-0.326	-0.102
card	1.897	0.090	21.16	0.000	1.722	2.073
ptca	-0.951	0.064	-14.88	0.000	-1.076	-0.826
constant	-5.271	0.135	-38.90	0.000	-5.536	-5.005
ln_p	-0.581	0.021	-27.25	0.000	-0.623	-0.539
ln_theta	-3.894	1.032	-3.77	0.000	-5.917	-1.872
p	0.559	0.012			0.537	0.583
1/p	1.787	0.038			1.714	1.864
theta	0.020	0.021			0.003	0.154

The estimated variance of the gamma frailty distribution was very small (0.02) suggesting little variation in survival times between hospitals. Although small, it was highly significant ($LRT\ \chi^2$=12.6, P<0.001) which indicates that some hospitals had higher hazards of death than other hospitals.

The coefficients for all of the fixed effects were very similar to those obtained from the Weibull model without a shared frailty (see Example 19.17).

19.11.4 Shared frailty models—Cox models

A Cox model with a frailty term added can be written either as:

$$h_i(t|\alpha_i) = h_0(t)e^{\beta X}\alpha_i \qquad \textit{Eq 19.30}$$

with the α_i being the frailty on the hazard scale (frailties on the hazard scale are often assumed to have a gamma distribution), or as:

$$h_i(t|\delta_i) = h_0(t)e^{\beta X+\delta_i} \qquad \textit{Eq 19.31}$$

with the δ_i (the shared frailty for the i^{th} group) on the log-hazard scale.

Estimating shared frailties in a Cox model is not straightforward. Four possible approaches are available: using a penalised likelihood function (see Example 19.22), using an expectation maximisation (EM) algorithm, fitting a random effects Poisson model (see below), or using Bayesian methods. With the exception of the Poisson model approach, these methods will not

Example 19.22 Shared frailty Cox model
data = mi

A shared frailty Cox model (with a gamma distributed frailty common to all patients in a hospital) was fit to the heart attack data.

No. of subjects = 2851 — Obs per group min = 12
No. of groups = 10 — avg = 285.1
No. of failures = 1480 — max = 1118
Time at risk = 3669144 — Wald $\chi^2_{(5)}$ = 1135.96
Log likelihood = -10437.755 — Prob > chi2 = 0.0000

Predictor	Coef	SE	Z	P	95% CI	
sex	0.212	0.058	3.65	0.000	0.098	0.325
age_ln	3.520	0.184	19.09	0.000	3.159	3.882
mar_c2	-0.221	0.057	-3.87	0.000	-0.332	-0.109
card	1.802	0.091	19.79	0.000	1.624	1.981
ptca	-0.942	0.064	-14.76	0.000	-1.067	-0.817
theta	0.020	0.022				

Once again, the estimate of the frailty variance is quite small (0.02), but highly significant (*LRT* χ^2=9.99, P=0.001)

The coefficients for the fixed effects are very similar to those obtained from the Weibull model with shared frailty (Example 19.21) and also with the Cox model without a shared frailty (model used in Example 19.14 although model not presented).

be discussed further except to state that the penalised likelihood approach is the computationally simplest and most commonly used method.

Shared frailty Cox model—Poisson regression
Poisson regression methods can be used to fit a standard Cox proportional hazards model and doing so produces exactly the same results. While this is not necessary (or practical) for fitting a standard Cox model, it has an advantage for shared frailty models in that random effects (equivalent to frailties) can be added to the Poisson model. This allows for the possibility of having more than one level of random effect and those effects can take on either gamma or log-normal distributions.

The procedure for fitting a Poisson model to survival data follows.

- Split each observation into multiple records according to the complete set of failure times in the dataset (*ie* each record will represent the time interval between the 2 consecutive failures). (**Note** This may create a long dataset and cause numerical problems.)
- Compute the length of time represented by each record (*ie* the interval between the 2 failure times) and log transform it.
- Fit a Poisson model which includes fixed effects for each time interval represented in the dataset and the log of the interval length as an offset.

To avoid fitting the large number of fixed effects for the time periods, create a set of orthogonal polynomials (see Section 15.6.3) for time and use them instead of the set of fixed effects.

To fit a shared frailty model, include time as the set of polynomials (as described above) and add a random effect for the group variable (*eg* hospital). Example 19.23 compares the results of fitting a proportional hazards model using standard Cox regression and Poisson regression procedures.

19.11.5 Frailty models—interpretation of coefficients

In a frailty model, the effects of predictors on the hazard or survival are 'conditional' on the frailty; that is they represent the effect of the predictor compared with an individual without the factor, but from the same group.

As noted above, the effects of predictors (*eg HR*) are the effects 'conditional' on the frailty. For proportional hazards models (*eg* Weibull), the *HR* at any time *t* represents the shift in the hazard due to a unit change in the predictor, conditional on the frailty (*ie* assuming a comparable frailty). For a dichotomous predictor, it represents the effect of the factor being present compared with an individual with exactly the same frailty but with the predictor absent. This is analogous to a 'subject-specific' effect—see Section 22.4.1.

Example 19.23 Cox model fit by Poisson regression
data = mi

Several Cox proportional hazards models were fit to the heart attack data with the same set of predictors as used in previous examples. The models were:

- Cox proportional hazards model (no frailty)
- Poisson regression model with time intervals as fixed effects (no frailty)
- Poisson regression model with time as a 4th order polynomial (no frailty)
- Cox proportional hazards model with shared frailty (gamma distribution)
- Poisson regression model with time as a 4th order polynomial and random effect (gamma distribution).

The coefficient for -mar_c2-, its Wald test P-value, and the estimate of the variance of the gamma distributions are presented.

Model	Coef	SE	P	Variance
Cox - no frailty	-0.221	0.057	0.000	
Poisson - time as fixed effects - no frailty	-0.220	0.057	0.000	
Poisson - time as polynomial - no frailty	-0.224	0.057	0.000	
Cox - shared frailty	-0.221	0.057	0.000	0.0198
Poisson - time as polynomial - shared frailty	-0.225	0.057	0.000	0.0233

As can be seen, the standard Cox model and Poisson models produce very similar results. (The minor difference is due to the fact that a very small number (18 of 880) of the risk sets generated by splitting the data at failure times were dropped because they contained only censored observations—presumably from the record of the failure being dropped due to a missing predictor value.) Expressing time as a 4th-order polynomial instead of a set of fixed effects produces results which are quite close and would probably be considered acceptable. The 2 approaches to fitting the shared frailty model produce slightly different, but close results.

In gamma frailty models, the population hazards (analogous to marginal effects—see Section 22.4.1) are not proportional over time and the hazard ratio only represents the population effect of the predictor at time 0. In general, the effect of the predictor on the population hazard will diminish over time in favour of the frailty effect. In simple terms, the frailty of the individual (or group) accounts for the fact that, over time, the population is increasingly 'robust' and the predictor has less and less influence on the hazard. With gamma frailties, the population *HR* tends to 1 as time approaches infinity, while, for an inverse Gaussian frailty, the *HR* tends toward the square root of the *HR*. This problem in interpreting the marginal effects of predictors is not present if the model is expressed as an AFT model—time ratios remain the same.

19.12 MULTIPLE OUTCOME EVENT DATA

In all of the material presented in this chapter so far, we have assumed that there was only one possible occurrence of the outcome of interest (*ie* death). However, in some instances, multiple outcome events are possible, and these fall into 3 general classes.

- **Multiple different failure events**—These arise in situations where you want to evaluate the effect of a predictor on multiple possible outcomes such as an evaluation of the effect of smoking cessation on the time to a heart attack, the time to a stroke, and the time to onset of chronic obstructive pulmonary disease. These are sometimes referred to as competing risks data
- **Multiple 'same' endpoints (not ordered)**—These arise in situations where multiple possible outcomes of the same event are possible, but there is not necessarily any ordering to them (*eg* time to onset of glaucoma in each eye). One way of dealing with these is to change the unit of observation to the eye, but in many cases, most of the risk factors will be at the person level
- **Multiple 'same' endpoints (ordered)**—These are also called **recurrence data**. They arise when it is possible for the outcome event to occur multiple times in the same individual (*eg* multiple heart attacks). The key feature to these is that there is a natural ordering to them (*ie* the second case cannot happen before the first). The lack of independence among episodes must be accounted for (see below). This type of data is the focus of this section.

19.12.1 Models for recurrence data

Event times within an individual are often correlated for 2 reasons. First, there is likely to be heterogeneity among individuals, with some individuals more likely to experience the outcome than others, leading to clustering of events within the individual. As a result, observations within an individual are not independent. The second is that the probability of occurrence of one event may increase or decrease the probability of subsequent events (called event dependence).

There are 2 general approaches to dealing with the problem of heterogeneity. One is to adjust the variance estimate using robust standard errors (see Sections 19.11.2 and 20.5.4). An alternative is to fit a shared frailty model with the frailty representing the intrinsic susceptibility of the individuals (Therneau and Grambsch, 2000). The former approach produces population averaged estimates of effect while the latter generates subject-specific estimates (Cain and Cole, 2006; Kelly and Lim, 2000).

The problem of event dependence can be dealt with either by including a covariate for the number of previous events in the model (see Anderson-Gill model below) or by stratifying the data according to the number of events (see Prentice-William-Peterson model below). The former approach assumes there is a common baseline hazard function for all events. The latter allows for the baseline hazard to vary for each event (*ie* a different baseline hazard for first events compared with second *etc*). Models for repeated events data have been reviewed recently along with a proposal for a conditional frailty model which addresses both the issues of heterogeneity and event dependence (Box-Steffensmeier and De Boef, 2006) (beyond the scope of this text).

Three approaches to modelling recurrence data have been reviewed (Wei and Glidden, 1997). Two of these will be summarised below, but the third (a marginal model (Wei *et al*, 1989)) is no longer recommended (Hosmer and Lemeshow, 2008) and will not be described. Details of structuring data appropriately for these analyses is presented in Cleves (2000) and is summarised below. The 2 approaches are shown in Example 19.24 using data from patients in the WHAS dataset who had one, or multiple, heart attacks. The number of people with 1, 2, 3, 4, 5, and 6 heart attacks were 2,230, 2,229, 361, 76, 24, and 8, respectively. Factors of interest that were investigated were gender (-sex-), race (-white-), and number of previous heart attacks (-prev_mi-).

Anderson-Gill model
This model is a generalised proportional hazard model and is the simplest approach to analysing recurrence data. The risk of recurrence is assumed to be independent of previous events, although the assumption of independence can be relaxed by including a time-varying predictor for the number of previous occurrences. The model is fit by assuming each subject's 'at-risk' time starts over again after each outcome is observed. If an individual is not considered to be at risk for a defined period after the occurrence of a case, then the time not at risk can be excluded (interval censored or gap). For example, after a heart attack, a person would presumably have to be free of symptoms for some minimum period of time for a recurrence to be considered a new attack.

Prentice-William-Peterson model—conditional risk sets model
This model is a proportional hazards model that is conditional on previous occurrences. It is equivalent to carrying out a stratified analysis with the strata defined by the number of previous outcome events. All first occurrences would be in the first stratum, the second stratum would consist of second cases, but only individuals that had experienced a first case would be at risk *etc.* Time at risk for each outcome can be measured either from the start of the study period or from the time of the previous event. The choice of approach depends on whether you feel that there is reason to 'reset the clock' each time an event occurs. An example of the former approach is shown in Example 19.24. As noted above, this approach allows for a different baseline hazard in each risk set (stratum).

19.13 DISCRETE-TIME SURVIVAL ANALYSIS

Up to this point, we have assumed that failure times were recorded on a continuous basis, that is, we knew exactly when each failure time occurred (at least to the unit of time measurement—which was 'days' in the heart attack datasets). However, we are often faced with the situation in which failures are known to occur in an interval, but the exact time is not available. These are called interval-censored data. Such a situation would arise if we measured time to seroconversion in individuals and they were tested only every 6 months. At some point, we

Example 19.24 Multiple failure event models
data = mi_mult

The structure of the multiple heart attack data for the Anderson-Gill and Prentice-William-Peterson models is shown in the table below. The Prentice-William-Peterson model assumed that the clock 'restarted' after each MI. Patient 19 had 1 MI at age 44 and was censored at age 52. Patient 20 had MIs at 76, 77, and 78 years of age.

						Anderson-Gill		Prentice-Williams-Peterson		
id	sex	white	agemi	mi	prev. mi	Start	End	Start	End	Risk set
19	1	1	44	1	0	0	44	0	44	1
19	1	1	52	0	1	44	52	0	8	2
20	0	0	76	1	0	0	76	0	76	1
20	0	0	77	1	1	76	77	0	1	2
20	0	0	78	1	2	77	78	0	1	3

The results for -sex- and -prev_mi- from fitting a variety of models to the data are shown below.

	sex		prev. mi	
	Coef	SE	Coef	SE
Anderson-Gill model				
Cox – robust SE	0.365	0.039	0.443	0.042
Weibull – robust SE	0.344	0.034	0.443	0.033
Weibull – gamma frailty	0.358	0.036	0.445	0.021
Prentice-William-Peterson (PWP) model				
Cox – robust SE	0.164	0.026	na	na
Weibull – robust SE	0.133	0.025	na	na

The Weibull model which used robust SE (as opposed to shared (gamma) frailty) to deal with heterogeneity produce lower estimates for the effect of -sex- and -prev_mi- because they produce population averaged estimates instead of subject-specific estimates. Being male (sex=1) generally increased the risk of having multiple heart attacks. However, the effect was not as pronounced if the clock was 'reset' so time at risk started over at 0 (PWP model) compared with when risk was evaluated as years from birth (AG model). The hazard increased with each previous MI.

would observe a serologic response and would know that seroconversion had occurred some time during the preceding 6 months. Discrete-time survival analysis can be used to analyse such data. In some cases, failure times may have been recorded on a continuous basis, but actual failure times are uncertain and grouping them into intervals may improve data quality.

Discrete-time models may also be used for continuous time data if:

- the dataset is very large and not amenable to standard survival analysis methods, or
- there are many time-varying predictors, or
- there are time-varying effects which are not easily modelled as some function of time.

This last situation was evident in the heart attack data in which a cardiac arrest in the hospital appeared to result in a high hazard of death immediately after its occurrence, but a rather rapid elimination of that effect thereafter (see Example 19.9).

Discrete time—basis for analysis

Time is divided into intervals, denoted I_j. In each interval, the number of subjects at risk is n_j and the number of failures is d_j. The probability of failure during the interval (or discrete time hazard) is then

$$h_j = d_j / n_j \qquad \textit{Eq 19.32}$$

Intervals can either be chosen to reflect the underlying biology of the situation or at convenient points which balance the width of the interval and the number of failures in each interval. (**Note** A general guide is to have a minimum of 5 failures in each interval and it is important to avoid choosing intervals based on the observed data—*ie* data snooping.) Observations which are censored during the interval may be considered to have been censored at the start of the interval (*ie* not included in n_j), censored at the end (*ie* included in n_j) or counted for 1/2 of the time interval (as was done in an actuarial life table analysis). Table 19.4 shows the heart attack data divided into intervals based on the expected effect of a cardiac arrest (*ie* short intervals to start and much longer intervals later). It appears that the increased hazard associated with a cardiac arrest has disappeared by day 60 and the two groups have comparable hazard thereafter.

Table 19.4 Heart attack data divided into intervals (by cardiac arrest category)

		no cardiac arrest			cardiac arrest		
t_{j-1}	t_j	n_j[1]	d_j	h_j[2]	n_j[1]	d_j	h_j[2]
0	10	2757	140	0.508	206	119	5.777
10	20	2617	66	0.252	87	21	2.414
20	30	2551	36	0.141	66	6	0.909
30	45	2515	39	0.103	60	6	0.667
45	60	2476	31	0.083	54	1	0.123
60	90	2445	52	0.071	53	0	0.000
90	120	2393	49	0.068	53	1	0.063
120	180	2344	78	0.055	52	1	0.032
180	360	2266	126	0.031	51	3	0.033
360	720	2140	207	0.027	48	2	0.012
720	1440	1895	309	0.023	46	5	0.015
1440	2942	1078	252	0.016	25	4	0.011

[1] observations censored during the interval considered censored at the start of the interval

[2] estimated daily hazards multiplied by 100 for ease of reading

If the data are truly discrete-time data (*ie* collected only at specific times), the time periods are defined by the data-collection periods, so intervals do not need to be created. If there are many time periods (intervals), you may want to replace the fixed effects for each time period with some form of polynomial model of time (*eg* orthogonal polynomials as was done in the Poisson

model in Section 19.11.4). See Singer and Willett (1993; 2003) for a review of discrete-time methods.

Discrete-time—logistic regression

Once the data have been structured as described above, they can be analysed using logistic regression according to the following model.

$$\text{logit}(h_j) = \beta_0 + \alpha_j + \beta X \qquad \textit{Eq 19.33}$$

where logit(h_j) is the probability (hazard) of failing in interval I_j given being present at the start of the interval, and β_0 is the logit(hazard) in the baseline time period for a baseline individual, α_j is the effect of the j^{th} time period (compared with the baseline period) and βX represents the predictors in the model. This model assumes additivity on the logit(hazard) scale or proportional odds for the hazard probabilities. (**Note** This corresponds to a continuation-ratio model (see Section 17.2.4) for the multinomial probabilities across all intervals.)

Discrete-time logistic regression models can easily be extended to include one or more random effects (shared frailties) using procedures for modelling multilevel data (see Chapter 22). Specialised software is required to fit individual frailty models to discrete time data (Jenkins, 1995). Example 19.25 shows the results of such a model with the time intervals as shown in Table 19.4 and a -card- by time interaction included.

Discrete-time—complementary-log-log regression

As noted above, the logistic model assumes that the log-odds of the outcome are additive, or alternatively that the odds are proportional. This is the same as saying that the *OR* for a predictor is constant across all time intervals (although this assumption can be relaxed by including interaction terms with the predictor). An alternative to logistic regression is to use a complementary log-log model which is based on the assumption of proportional hazards (not proportional odds) and consequently is a more natural fit with models such as the Cox proportional hazards model.

The complementary log-log function transforms a probability as follows

$$\text{cloglog}(p) = \ln[-\ln(1-p)] \qquad \textit{Eq 19.34}$$

Fig. 19.26 shows the relationship between probability and both the complementary log-log and logit functions. At $p<0.2$ the 2 functions are very close, but become substantially different at large values of p (and may produce substantially different results in binary regression analyses). As noted, the main advantage of a complementary log-log model is that it is based on the proportional hazards assumption and consequently, exponentiated coefficients can be interpreted as hazard ratios (as opposed to odds ratios).

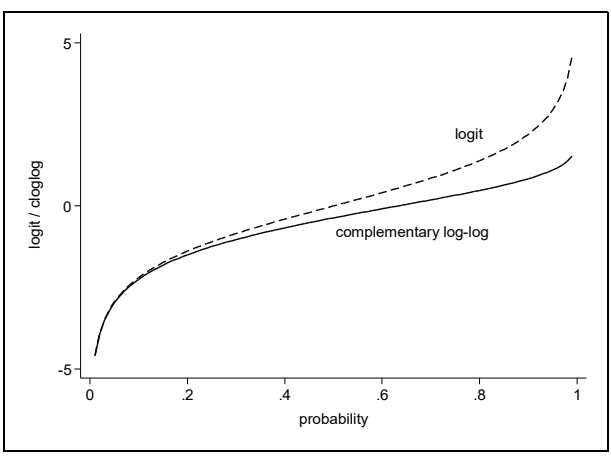

Fig. 19.26 Complementary log-log and logit functions

Example 19.25 Discrete-time analysis—logistic regression
data = mi

A discrete-time analysis using logistic regression was carried out on the heart attack data. Cardiac arrest by time period interaction terms were included. Not all time period and time period interaction coefficients are shown.

Logistic regression — Number of obs = 25902
Log likelihood = -3856.1672 — LR $\chi^2_{(24)}$ = 2393.55
Prob >χ^2 = 0.0000

Predictor	Coef	SE	Z	P	95% CI	
sex	0.236	0.070	3.37	0.001	0.099	0.373
age_ln	3.602	0.215	16.72	0.000	3.180	4.024
mar_c2	-0.236	0.069	-3.42	0.001	-0.371	-0.101
ptca	-1.124	0.077	-14.63	0.000	-1.275	-0.974
card	3.495	0.184	18.96	0.000	3.133	3.856
period 0–9	-0.717	0.155	-4.62	0.000	-1.020	-0.413
period 10–19	-1.282	0.191	-6.72	0.000	-1.656	-0.908
period 20–29	-1.225	0.189	-6.48	0.000	-1.595	-0.854
... some output omitted						
period 360–719	0.937	0.118	7.93	0.000	0.705	1.169
period 720–1439	2.038	0.115	17.73	0.000	1.813	2.264
per 0–9 * card	-0.378	0.355	-1.06	0.287	-1.074	0.318
per 10–19 * card	-0.889	0.516	-1.72	0.085	-1.900	0.122
per 20–29 * card	-0.865	0.549	-1.58	0.115	-1.941	0.211
... some output omitted						
per 360–719 * card	-3.772	0.767	-4.92	0.000	-5.275	-2.269
per 720–1439 * card	-3.111	0.553	-5.63	0.000	-4.194	-2.027
constant	-3.740	0.134	-27.92	0.000	-4.002	-3.477

The coefficient for cardiac arrest (β=3.495) represents the effect of a cardiac arrest in the first time period (0–10 days after admission to hospital). The -card- by time period interaction term for the period 120–180 days (not shown in table) was -3.408 indicating that the effect of a cardiac arrest had dissipated by 120 days. The interaction terms remained at approximately that level for all subsequent time periods.

Example 19.26 shows the results from a complementary log-log model of the heart attack data.

Example 19.26 Discrete-time analysis—complementary log-log regression
data = mi

A complementary log-log model of the heart attack data that were used in Example 19.25 produced the following results.

Complementary log-log regression
Zero outcomes = 24637 Number of obs = 25902
Non-zero outcomes = 1265 LR X^2 (26) = 2428.52
Log likelihood = -3858.094 Prob >X^2 = 0.0000

Predictor	Coef	SE	Z	P	95% CI	
sex	0.211	0.063	3.34	0.001	0.087	0.335
age_ln	3.314	0.199	16.64	0.000	2.923	3.704
mar_c2	-0.216	0.062	-3.47	0.001	-0.338	-0.094
ptca	-1.013	0.071	-14.34	0.000	-1.151	-0.874
card	2.860	0.133	21.59	0.000	2.601	3.120
period 0–9	-0.693	0.150	-4.61	0.000	-0.988	-0.399
period 10–19	-1.248	0.187	-6.67	0.000	-1.615	-0.882
period 20–29	-1.194	0.185	-6.45	0.000	-1.556	-0.831
... some output omitted						
period 360–719	0.877	0.111	7.93	0.000	0.660	1.094
period 720–1439	1.836	0.103	17.75	0.000	1.633	2.039
per 0–9 * card	-0.065	0.287	-0.23	0.822	-0.626	0.497
per 10–19 * card	-0.378	0.462	-0.82	0.413	-1.284	0.528
per 20–29 * card	-0.448	0.495	-0.90	0.366	-1.418	0.523
... some output omitted						
per 360–719 * card	-3.177	0.723	-4.39	0.000	-4.594	-1.759
per 720–1439 * card	-2.562	0.472	-5.43	0.000	-3.486	-1.638
constant	-3.692	0.126	-29.29	0.000	-3.940	-3.445

The results are similar to those from the logistic regression analysis. However, the effect of -card- may be underestimated in the first time period when the probability of death in cardiac arrest patients was 58%. A comparison of the log likelihoods suggests that the logistic model fits the data better.

19.14 SAMPLE SIZES FOR SURVIVAL ANALYSES

Computation of sample sizes for studies with survival time as the outcome can be a complex process. For studies where the primary focus is the comparison of survival times across 2 (or more) groups, as it often is in controlled trials, one approach is to compute the sample size required to have a desired power in an analysis based on an unweighted log-rank test. If an assumption of proportional hazards is likely not valid, basing the sample size on that required

for a weighted version of the test (*eg* Tarone-Ware or Harrington-Flemming tests) might be more appropriate.

However, there are many factors which will influence the required sample size. Some of the following have been discussed under sample size estimation in Chapter 2, and some are unique to studies of survival time.

1. Sample size might need to be increased to account for multiple predictors in the analysis, and/or to adjust for clustering of the data (*ie* non-independence among observations) (see Chapter 2).
2. As pointed out in Chapter 11, multiple comparisons (often arising from interim analyses), losses in the follow-up process, and subgroup analyses are common features of controlled trials which require adjustment to the sample size.

Example 19.27 Sample size calculations for a randomised controlled trial
data = hypothetical

Assume that you are about to start a randomised controlled trial of 2 drugs designed to prevent recurrence of a certain type of cancer, following initial treatment of the condition. Untreated controls will also be included in the study. Past experience has shown that as the risk of recurrence goes down, the longer the patient remains in remission. In the absence of treatment, you expect the cumulative probabilities of recurrence in each of 4 time periods to be as follows:

- end of year 1 30%
- end of year 2 50%
- end of year 4 60%
- later in life 65%.

Relative to untreated controls, you expect treatment A to have a *HR* of 0.75 and treatment B to have a *HR* of 0.5. You consider the following 5 scenarios.

(a) no loss to follow, no cross-over of treatments, equal allocation of subjects to the 3 groups

(b) same as (a) except cumulative loss to follow up of 5%, 15%, 30%, and 40% in the 4 time periods

(c) same as (a) except cumulative loss to follow up of 10%, 30%, 60%, and 80% in the 4 time periods

(d) same as (c) except 20% of control patients and 10% of treatment A patient's cross-over into treatment B

(e) same as (d) except you initially allocate patients in the following ratio: control=1, treatment A=1, treatment B=2.

For each scenario, you want to determine the sample size required to have an overall power of 80% for detecting a difference among treatment groups. The required sample sizes and expected number of recurrences are:

	Scenario				
	(a)	**(b)**	**(c)**	**(d)**	**(e)**
Total sample size	742	837	965	1379	1347
Expected number of recurrences	353	353	353	502	500

As expected, the sample size goes up with increasing loss to follow up (with no increase in the expected number of cases). Subjects switching treatments (d) increases the required sample size, while changing the allocation of subjects was able to slightly reduce the total sample size required.

3. The shape of the baseline hazard function might not be known in advance of the study, so a sample size estimate based on a non-parametric test (*eg* log-rank) would be appropriate.
4. The possibility of non-proportional hazards needs to be considered.
5. In controlled trials, cross-over might occur in which individuals could move from one treatment group to another (*eg* treated to non-treated if the patient fails to comply with treatment instructions).
6. Recruitment of individuals into the study may take place over time which might affect the length of follow-up period for individuals recruited.
7. Survival analyses are often used in randomised controlled trials. In non-randomised studies of therapeutic interventions, subjects with the new treatment are often matched to those receiving a standard treatment within strata defined by covariates of interest.

A general discussion of sample size issues can be found in Hosmer and Lemeshow (2008); Mazumdar *et al* (2006). A review of some of the issues identified above and a description of a software program for computing sample sizes for survival analysis studies has recently been published (Barthel *et al*, 2006; Royston and Babiker, 2002) (see Example 19.27).

REFERENCES

Aitken RJ, Baker MA. Oxidative stress, sperm survival and fertility control. Mol Cell Endocrinol. 2006 May 16;250(1-2):66-9.

Barthel FMS, Babiker A, Royston P, Parmar MKB. Evaluation of sample size and power for multi-arm survival trials allowing for non-uniform accrual, non-proportional hazards, loss to follow-up and cross-over. Stat Med. 2006;25(15):2521-42.

Box-Steffensmeier JM, De Boef S. Repeated events survival models: the conditional frailty model. Stat Med. 2006;25(20):3518-33.

Cain LE, Cole SR. Survival analysis for recurrent event data: an application to childhood infectious diseases. Stat Med. 2006;25(8):1431-3; author reply 3.

Cleves MA. Analysis of multiple failure-time data with Stata. Stata Tech Bull. 2000;49:30-9.

Cleves MA, Gould WW, Gutierrez RGRB. An Introduction to Survival Analysis Using Stata. 2nd Ed. College Station (TX): Stata Press; 2008.

Collett DB. Modelling Survival Data in Medical Research. 2nd Ed. New York: Chapman and Hall; 2003.

Cox DRJ. Regression models and life-tables (with discussion). Journal of the Royal Statistical Society, B. 1972;34:187-220.

Dohoo IR, Stryhn H, Sanchez JC, editors. Beyond Cox: 1. Hazard functions and time varying effects in parametric survival models 2003.

Ego A, Subtil D, Cosson M, Legoueff F, Houfflin-Debarge V, Querleu D. Survival analysis of fertility after ectopic pregnancy. Fertil Steril. 2001 Mar;75(3):560-6.

Ganesh JS, Rogers CA, Banner NR, Bonser RS. Donor cause of death and medium-term survival after heart transplantation: a United Kingdom national study. J Thorac Cardiovasc

Surg. 2005 May;129(5):1153-9.

Goldberg RJ, Ciampa J, Lessard D, Meyer TE, Spencer FA. Long-term survival after heart failure: a contemporary population-based perspective. Arch Intern Med. 2007 Mar 12;167(5):490-6.

Goldberg RJ, Spencer FA, Szklo-Coxe M, Tisminetzky M, Yarzebski J, Lessard D, et al. Symptom presentation in patients hospitalized with acute heart failure. Clin Cardiol. 2010 Jun;33(6):E73-80.

Grønnesby JK, Borgan O. A method for checking regression models in survival analysis based on the risk score. Lifetime Data Anal. 1996;2(4):315-28.

Gutierrez RGJ. Parametric frailty and shared frailty survival models. The Stata Journal. 2002;2:22-44.

Hosmer DW, Royston PJ. Using Aalen's linear hazards model to investigate time-varying effects in the proportional hazards regression model. The Stata Journal. 2002;2(4):331-50.

Hosmer DW, Lemeshow SB. Applied Survival Analysis. Regression modelling of time to event data. 2nd Ed. New York: John Wiley & Sons; 2008.

Jenkins SP. Easy estimation methods for discrete-time duration models. Oxford Bulletin of Economics and Statistics. 1995;57:129-38.

Kaplan EL, Meier PJ. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association. 1958;53:457-81.

Kelly PJ, Lim LL. Survival analysis for recurrent event data: an application to childhood infectious diseases. Stat Med. 2000;19(1):13-33.

Lin DY, Wei LJ. The robust inference for the Cox proportional hazards model. Journal of the American Statistical Association. 1989;84:1074-8.

May S, Hosmer DW. Hosmer and Lemeshow type Goodness-of-Fit statistics for the Cox proportional hazards model. Handbook of Statistics. 2004a;23(3):383-94.

May S, Hosmer DW. A cautionary note on the use of the Grønnesby and Borgan goodness-of-fit test for the Cox proportional hazards model. Lifetime Data Anal. 2004b;10 (3):283-91.

Mazumdar M, Tu D, Zhou XK. Some design issues of strata-matched non-randomized studies with survival outcomes. Stat Med. 2006;25(23):3949-59.

McManus DD, Piacentine SM, Lessard D, Gore JM, Yarzebski J, Spencer FA, et al. Thirty-year (1975 to 2005) trends in the incidence rates, clinical features, treatment practices, and short-term outcomes of patients <55 years of age hospitalized with an initial acute myocardial infarction. Am J Cardiol. 2011 Aug 15;108(4):477-82.

Moreau T, O'Quigley J, Mesbah M. A global goodness-of-fit statistic for the proportional hazards model. Applied Statistics. 1985;34:212-8.

O'Quigley J, Stare J. Proportional hazards models with frailties and random effects. Stat Med. 2002;21(21):3219-33.

Pocock SJ, Clayton TC, Altman DG. Survival plots of time-to-event outcomes in clinical trials:

good practice and pitfalls. Lancet. 2002;359(9318):1686-9.

Royston P, Babiker AJ. A menu driven facility for complex sample size calculation in randomized controlled trials with a survival or binary outcome. The Stata Journal. 2002;2(2):151-63.

Royston P. Explained variation for survival models. The Stata Journal. 2006;6:83-96.

Royston P, Sauerbrei W. Multivariable model-building. A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Chichester, England: John Wiley & Sons, Ltd; 2008.

Singer JD, Willett JB. It's about time: using discrete-time survival analysis to study duration and the timing of events. J Educational Stat. 1993;18:155-95.

Singer JD, Willett JB. Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. Oxford: Oxford University Press; 2003.

Therneau TM, Grambsch PM. Modelling survival data: extending the Cox model. New York: Springer-Verlag; 2000.

Wei LJ, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. J Am Stat Assoc. 1989;84:1065-73.

Wei LJ, Glidden DV. An overview of statistical methods for multiple failure time data in clinical trials. Stat Med. 1997;16(8):833-9.