

Event History Analysis

Fiona Steele

ESRC National Centre for Research Methods

NCRM Methods Review Papers

NCRM/004

Event History Analysis

Fiona Steele

Centre for Multilevel Modelling, Graduate School of Education, University of Bristol

September 2005

EVENT HISTORY ANALYSIS

A National Centre for Research Methods Briefing Paper

Fiona Steele

Centre for Multilevel Modelling
Graduate School of Education
University of Bristol
35 Berkeley Square
Bristol BS8 1JA

September 2005

1. Event History Data

1.1 Introduction

An event history is a longitudinal record of the timing of the occurrence of one or more types of event. Examples include employment histories which typically include dates of any changes in job or employment status, and partnership histories which usually include the start and end dates of co-residential relationships. In an analysis of employment histories events of interest might be the end of an employment or unemployment spell, while a study of partnership histories might examine entry into marriage and marital dissolution. Event history analysis is used to study the duration until the occurrence of the event of interest, where the duration is measured from the time at which an individual becomes exposed to the 'risk' of experiencing the event. For instance, an individual is at risk of marital dissolution from the time they marry. In some cases, it is less obvious when the risk period starts and the researcher must choose an appropriate origin. One such example arises in the study of women's age at first birth, referred to as the first birth interval. In societies where pre-marital births are rare, the age at marriage would be a natural origin, but if pre-marital births are common we might take the age of first menstruation as the start of the risk period.

The techniques described in this review are also commonly known as *survival analysis*, *duration analysis* or *hazard modelling*. Although often used interchangeably with survival analysis, the term *event history analysis* is used primarily in social science applications where events may be repeatable and an individual's *history* of events is of interest.

1.2 The collection and management of event history data

Event histories are collected in a number of social surveys. In Britain, for example, partnership, employment and birth histories are collected in the 1958 and 1970 birth cohort studies (NCDS and BCS70) and the British Household Panel Study (BHPS). Event history data are almost always collected retrospectively, with respondents asked to recall the dates of events that have occurred since a certain age or during a fixed window of time before the interview. Respondents may be asked to recall events in the order that they occurred or in reverse chronological order, depending on the significance of the start of the observation period with reference to the process under study. For instance, if information is required on partnerships formed since age 16 it is most natural to start with the first partnership, followed by the second, moving forwards in time towards the interview. This strategy was adopted in the NCDS when, at the age 33 interview, respondents were asked for details of all partnerships formed from age 16. At age 42 respondents were reinterviewed and asked to recall the dates of partnerships since age 33. Because the dates of partnerships formed around age 33 are likely to be more difficult to recall than the first partnership, respondents were first asked about their current partnership, followed by their most recent ex-partner and so on, moving backwards in time towards age 33.

The potential for recall error will depend on the salience of the events to individuals and the length of the recall period. For this reason, surveys carried out in less-developed countries under the Demographic and Health Surveys (DHS) programme collect full birth histories while contraceptive use histories are collected for only a six-year window before the survey. Among the techniques adopted to minimise recall error is the use of milestone events as reference points. For example, female respondents in a DHS survey are asked about periods of contraceptive use with reference to the dates of birth and names of children born during the observation period. Another strategy for improving data quality is to collect event times in the form of a calendar. In DHS contraceptive history calendars, births and their preceding pregnancies are entered first, followed by the most recent episode of contraceptive use. Recording births and pregnancies first allows these events to

be used as anchors and, in addition, allows inconsistencies such as the use of contraceptives during pregnancy to be detected and reconciled during the interview.

Event histories are usually stored in the form in which they were collected or in as compact a format as possible; this rarely corresponds to a data structure that is convenient for analysis. For example, partnerships histories collected in the cohort studies are in the form of a set of variables including the start and end dates of marriages and cohabitations, and indicators of whether each marriage was preceded by a period of cohabitation. In DHS datasets, a woman's contraceptive history is stored as a single variable occupying 72 columns, one for each month of the calendar. Prior to an event history analysis the data must be restructured so that there is a record for each episode, where an episode is a continuous period during which an individual was at risk of experiencing an event. If there was a maximum of one event per individual, there will be a single record for each individual. Manipulating event history data into a form suitable for analysis requires some programming skill and is often extremely time consuming.

1.3 Issues in the analysis of event history data

There are two particular features of event history data which must be considered in their analysis. First, the event of interest may not yet have occurred to a proportion of the sample. The duration to event occurrence, the response variable in an event history analysis, is said to be *right censored* for these individuals. Second, we are usually interested in the effects of explanatory variables on the timing of events, and some of these variables may change value over the course of the observation period; such variables are called *time-varying covariates*.

1.3.1 Incompletely observed durations

Denote by t_i the event time for individual i ($i = 1, \dots, n$). There are usually individuals who have not experienced the event by the end of the observation period. Suppose, for example, that we are interested in the age at first marriage which we collect retrospectively by asking respondents to recall the date of their marriage. Some of these respondents will still be single at the time of interview and therefore their event times (age at marriage) are incompletely observed; all we can infer is that their age at marriage is greater than their age at interview. This form of incomplete observation is called *right censoring*. Denote by c_i the censoring time for individual i . For any individual we observe the minimum of t_i and c_i which we denote by y_i , and a censoring indicator which is defined as

$$\delta_i = \begin{cases} 1 & \text{if censored, i.e. } t_i > c_i \\ 0 & \text{if uncensored, i.e. } t_i \leq c_i \end{cases}$$

Although right censoring is the most common form of incomplete observation, durations may also be left truncated or left censored. *Left truncation* arises when an individual is exposed to the risk of experiencing the event before the start of the observation period, and is a common problem when event histories are collected retrospectively for a fixed window of time. For example, in DHS surveys the episodes of women who were using contraceptives at the start of the six-year calendar period are all left truncated because women were not asked to recall the start date of these episodes. *Left censoring* arises when a history is only partially observed so that some events occurred before the start of the observation period. Again, left censored durations are common when a restricted observation period is used, such as in the DHS calendar. These different forms of incomplete observation are illustrated in Figure 1. The event histories for four individuals are shown. The duration of the first individual is completely observed. The second individual has experienced two events; the duration to the first event is left censored because the event occurred before the start of

the recall period, while the second duration is right censored. The durations of the third and fourth individuals are left truncated because they both became at risk before the start of the observation window. The fourth individual's event time is also right censored.

Another form of incomplete observation is *interval censoring*. A duration is said to be interval censored if the event time is not recorded precisely but is known to fall within a given range. In fact *all* durations in retrospectively collected event histories should, strictly speaking, be viewed as interval censored. To reduce respondent burden and minimise recall error, event times are typically recorded to the nearest month or year. Depending on the width of these intervals, it may be more appropriate to use an event history model which assumes measurement in discrete rather than continuous time (see Section 4).

Left truncation and left censoring are difficult to handle. In the absence of additional information about these observations, the most common approach taken is to exclude such observations from the analysis. Fortunately, left truncation and left censoring occur much less frequently in practice than right censoring which is much easier to accommodate (under certain assumptions which are discussed below). See Guo (1993) and Hosmer and Lemeshow (1999: Chapter 7) for further discussion of event history analysis of left-censored and left-truncated data.

One way of dealing with right-censored observations would be to simply exclude them from the analysis. This approach is not recommended because it can lead to a drastic reduction in sample size and, more importantly, substantial bias. To illustrate the problem of bias, suppose that we are interested in studying the age at first marriage among a cross-section of women, based on retrospective reports of their marriage dates. Any woman who is still single at the time of interview is right censored. Suppose we drop them from the analysis sample. The excluded cases are unlikely to be a random subset of the original sample; they are likely to be predominantly younger women who have not yet had a chance to marry and older women who have delayed or rejected marriage. Omitting these women will bias the sample towards older women and young women who married early.

All methods discussed in this review retain right-censored observations under the assumption that censoring times c_i are independent of event times t_i , i.e. censoring is non-informative. Essentially we assume that individuals are not selectively withdrawn from the sample because they are more or less likely to experience the event. In experimental research this assumption may be questionable. To take an example, suppose a clinical trial is conducted to test the effect of a new drug on patient survival. Informative (or non-random) censoring would arise if weaker patients taking the drug are more likely to suffer negative side effects and, as a result, more likely to withdraw from the trial. When event history data are collected retrospectively, as is the case in most social research, it is usually reasonable to assume non-informative censoring because c_i is determined by the interview date which is random with respect to the event time t_i .

1.3.2 Time-varying covariates

Together with longitudinal information on the timing of events, we may also have data on changes in individual characteristics or circumstances over time. For example, from employment histories collected in the British cohort studies and the BHPS it is possible to determine whether a person is in full-time education at a given point in time. In an analysis of age at first partnership, we might be interested in the relationship between an individual's probability of partnering at time t and their educational status at that time. Educational enrolment is an example of a *time-varying covariate*. While one approach would be to take the value of such variables at one point in time, such as the start of the observation period, this is wasteful and does not allow us to explore how the timing of

an event relates to a *change* in the value of a covariate. In Sections 3 and 4 we discuss how time-varying covariates can be included in event history models.

2. Descriptive Event History Analysis

2.1 Definitions: hazard and survivor functions

The first step of an event history analysis is to examine the distribution of event times using the hazard function and the survivor function. We assume that event times are realisations of a random variable T .

The *hazard function* is defined as

$$h(t) = \lim_{dt \rightarrow 0} \frac{\Pr(t \leq T < t + dt \mid T \geq t)}{dt}.$$

$\Pr(t \leq T < t + dt \mid T \geq t)$ is the probability that an event occurs during the interval $(t, t + dt)$ given that no event has occurred before time t . The notation $\lim_{dt \rightarrow 0}$ is shorthand for ‘the limit as the width of the interval dt gets infinitesimally small’, so the hazard expresses the probability that an event occurs within a very small interval of time (given that an event has not already occurred). For this reason, the hazard function is also commonly known as the instantaneous risk. Other terms for $h(t)$ includes the hazard rate and failure rate.

The *survivor function* is the probability that no event has occurred before time t :

$$S(t) = \Pr(T \geq t).$$

Individuals who have not yet experienced the event are said to have ‘survived’. The term ‘survivor’ comes from biostatistics where the event of interest is often death. Note that $S(t)$ is a decreasing function of t , with $S(0) = 1$ and $S(t) \rightarrow 0$ as $t \rightarrow \infty$.

The complement of the survivor function, known as the *cumulative distribution function*, is the probability that an event occurs before t :

$$F(t) = 1 - S(t) = \Pr(T < t)$$

2.2 Estimation of the hazard and survivor functions

The usual way of estimating the hazard function and other summary functions of the duration distribution is to construct a *life table*. In a life table durations are grouped into intervals of time, so that t now refers to a time interval rather than a point in time. These intervals are usually, but not necessarily, of equal width. Denote the hazard in interval t by h_t , where $t = 1, \dots, g$. For each interval we calculate the following:

- r_t the number of individuals at risk of an event at the start of interval t ;
- d_t the number of events during interval t ;
- w_t the number of individuals who are censored (or who withdraw) during interval t .

Estimates of h_t are based on the ratio of the number of events (d_t) to the number at risk (r_t) with some adjustment made for censoring. For example, the *actuarial estimator* assumes that the censored observations w_t are distributed uniformly across the interval t , so that censored cases are on average at risk for half the interval, i.e.

$$\hat{h}_t = \frac{d_t}{r_t - w_t/2}. \quad (2.1)$$

We denote by S_t the survivor function for interval t . An estimate of S_t can be derived directly from \hat{h}_t as follows. The probability of survival up to the start of interval $t = 1$ is by definition equal to 1. S_2 is the probability of survival to the start of the second interval, which equals the probability that no event occurs in $t = 1$, i.e. $S_2 = 1 - h_1$. Similarly S_3 , the probability of survival to the start of the third interval, equals the probability that no event occurs in either $t = 1$ or $t = 2$ (given survival to $t = 2$), i.e. $S_3 = (1 - h_1) \times (1 - h_2)$. Thus an estimate of S_t is:

$$\begin{aligned} \hat{S}_1 &= 1; \\ \hat{S}_t &= (1 - \hat{h}_1)(1 - \hat{h}_2) \dots (1 - \hat{h}_{t-1}) \text{ for } t > 1 \\ &= \prod_{j=1}^{t-1} (1 - \hat{h}_j). \end{aligned} \quad (2.2)$$

Note that sometimes the survivor function is defined as $S_t = \Pr(T > t)$ in which case the above expression for \hat{S}_t is multiplied by $(1 - \hat{h}_t)$.

Finally, the cumulative density function is estimated as $\hat{F}_t = 1 - \hat{S}_t$. Most software packages will compute standard errors and confidence intervals for the sample estimates \hat{h}_t , \hat{S}_t and \hat{F}_t . (See Hosmer and Lemeshow (1999: Chapter 2) for details of variance estimation, including Greenwood's formula for the variance of the survivor function.)

2.3 Example: age at first partnership

We will illustrate estimation of the hazard and survivor functions in an analysis of age at first partnership, which includes marriage and cohabitation, among a subsample of 500 respondents from the National Child Development Study, a study of all men and women born in Britain during a single week of 1958. We analyse data collected when the respondents were age 33. An analysis of the full sample can be found in Berrington (2003).

The duration of interest is the number of years between the sixteenth birthday and the start of the first partnership. The sample contains 35 individuals who had not partnered by age 33 and whose durations are therefore right censored. Because of the cohort design, censoring cannot occur before age 33. Table 1 shows an extract of the life table based on yearly intervals. From the estimates of the hazard at each age we can see, for example, that 4.2% of those who did not form their first partnership at age 16 did so at age 17. Turning to the survivor function, we estimate that 94.1% were unpartnered at their 18th birthday (the start of the third year of the observation period) and 7.8% were still unpartnered at age 32.

The hazard and survivor functions are usually displayed in graphical form (see Figures 2 and 3). The hazard of first partnership increases rapidly until age 22, then remains fairly constant before dropping sharply at age 31 (Figure 2). The survivor function is less informative because it always decreases with time (age). Nevertheless, we can see that the greatest decrease in the proportion remaining unpartnered is during the early twenties when the hazard is consistently high (Figure 3).

3. Continuous-time Event History Models

3.1 Introduction

The aim of most event history analyses is to identify factors that are associated with the timing of the event of interest. In the following sections we describe how covariates can be incorporated in a regression model. The values of these covariates may be fixed over time or time varying. There is a wide range of event history models from which to choose. One distinction between models is based on whether event times are assumed to be measured in continuous or discrete time. In this section we consider continuous-time models; the discrete-time approach is described in Section 4. Models can also be classified as either proportional hazards or accelerated life models, according to the way in which covariates are assumed to affect the timing of events (see Section 3.2). The most important consideration when choosing an appropriate model, however, is the nature of the distributional assumption for event times. The assumption of exponentially distributed event times, for example, implies a constant hazard, while distributions such as the Weibull and gamma imply monotonically increasing or decreasing hazards. The most flexible continuous-time model, and therefore the most frequently applied, is the Cox proportional hazards model. The Cox model is discussed in detail in Sections 3.3-3.5, with an example of its application and interpretation in Section 3.6.

3.2 Proportional hazards and accelerated life models

3.2.1 Proportional hazards model

For each individual i we observe a vector of covariates \mathbf{x}_i with values fixed across time. The hazard at time t is now a function of t and \mathbf{x}_i , which we denote by $h(t; \mathbf{x}_i)$. Denote by $h_0(t)$ the hazard at $\mathbf{x}_i = \mathbf{0}$, i.e. $h(t; \mathbf{x}_i = \mathbf{0})$. If all covariates are categorical, $h_0(t)$ is the hazard for individuals in the reference (baseline) category of each variable. For this reason $h_0(t)$ is often referred to as the *baseline hazard*. A proportional hazards (PH) model is written as

$$h(t; \mathbf{x}_i) = h_0(t) g(\mathbf{x}_i), \quad (3.1)$$

where $g(\mathbf{x})$ is some function of the covariates. If the values of the covariates are changed from their reference categories (or, more generally, from zero) to a value \mathbf{x}^* , then the hazard is multiplied by $g(\mathbf{x}^*)$. Therefore, the covariates are assumed to have a multiplicative effect on the hazard.

The PH assumption implies that the effect of a change in \mathbf{x} on the hazard is the same for all values of t . To see this, consider the hazard functions for two different sets of covariate values, \mathbf{x}_1 and \mathbf{x}_2 . From (3.1) the ratio of the hazards at these two values of \mathbf{x} is

$$\frac{h(t; \mathbf{x}_1)}{h(t; \mathbf{x}_2)} = \frac{g(\mathbf{x}_1)}{g(\mathbf{x}_2)},$$

which does not depend on t .

3.2.2 Accelerated life model

An accelerated life (AL) model is based on the idea that individuals experience time in different units. For example, suppose we wish to compare mortality risks of humans ($x = 0$) and dogs ($x = 1$). Dogs have a shorter lifespan than humans, so dogs are said to age faster than humans. If a year of human life is approximately equal to seven dog years, the relationship between the survivor functions for humans and dogs can be expressed as

$$S(t; x = 1) = S(7t; x = 0).$$

More generally an AL model can be written

$$S(t; \mathbf{x}_i) = S_0(g(\mathbf{x}_i) t),$$

where $S_0(t)$ is the survivor function at $\mathbf{x} = \mathbf{0}$.

The AL model assumes that a change in covariate values from $\mathbf{0}$ to \mathbf{x}^* accelerates time by a factor $g(\mathbf{x}^*)$ or, equivalently, reduces the median survival time by a factor $g(\mathbf{x}^*)$. While the PH model assumes that the covariates have a multiplicative effect on the *hazard*, the AL model assumes that covariates have a multiplicative effect on the *timescale*.

In practice PH models are used far more frequently than AL models, to the extent that Hosmer and Lemeshow (1999; p.273) state that “It [the PH model] is now accepted as the standard method for regression analysis of survival times in many applied settings.” However, it is not always necessary to make the distinction between these two types of model. Both the Weibull model and the exponential model (a special case of the Weibull) can be viewed as a PH or AL model, and their parameters interpreted as covariate effects on the hazard or the timescale. See Hosmer and Lemeshow (1999: Chapter 8) for further details.

3.3 Cox proportional hazards model

The most commonly applied event history model is the Cox proportional hazards model. In the Cox model, $g(\mathbf{x}) = \exp(\boldsymbol{\beta}' \mathbf{x})$ so that (3.1) becomes

$$h(t; \mathbf{x}_i) = h_0(t) \exp(\boldsymbol{\beta}' \mathbf{x}_i), \quad (3.2)$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients.

The model is sometimes written in linear form by taking natural logarithms of both sides of (3.2) to give

$$\log h(t; \mathbf{x}_i) = \log h_0(t) + \boldsymbol{\beta}' \mathbf{x}_i, \quad (3.3)$$

where \log denotes the natural logarithm. One reason for the popularity of the Cox model is its flexibility; the baseline hazard function $h_0(t)$ is left completely unspecified. Another attractive feature of the model is that the exponents of the regression coefficients $\boldsymbol{\beta}$ can be interpreted as

relative risks. To illustrate the interpretation of the parameters of the Cox model, suppose we have a single covariate x which is coded 0 or 1. From (3.2) we obtain

$$h(t; x = 0) = h_0(t)$$

$$h(t; x = 1) = e^\beta h_0(t),$$

so e^β is the ratio of the hazard for $x = 1$ to the hazard for $x = 0$. e^β is referred to as the *relative risk* or *hazard ratio*. If $\beta = 0$ (or $e^\beta = 1$) then x has no effect on the hazard. A positive β (or $e^\beta > 1$) implies that the group with $x = 1$ has a higher hazard, or shorter event time, than the group with $x = 0$. A negative β (or $e^\beta < 1$) implies that the hazard is higher among the group with $x = 0$. For continuous x , the hazard changes by a factor of e^β for each 1-unit change in x .

We can test the null hypothesis that $\beta = 0$ using an approximate t-test, based on the ratio $\hat{\beta} / SE(\hat{\beta})$, or a likelihood ratio test. Confidence intervals for the relative risks e^β are calculated by taking the exponents of the lower and upper limits of confidence intervals for β . Examples of hypothesis testing and confidence intervals are given in Section 3.6.

3.4 The proportional hazards assumption

As noted in Section 3.2.1, a key assumption of the PH model is that the effects of covariates are the same for all values of t . This assumption should be checked before interpreting the results from fitting a Cox model. The PH assumption is usually checked graphically. For illustration suppose that we have a single covariate x coded 0 or 1. We can estimate the hazard function separately for each group, using the method described in Section 2.2, and plot the hazards on the same graph. If the PH assumption holds, the vertical distance between the points for the two groups should be the same for each value of t (see Figure 4a). Figure 4b shows an example of non-proportional hazards where the groups have similar shaped hazards but different peaks, leading to a cross-over in the curves. A similar plot could be produced for a continuous x after grouping values into a small number of categories. The PH assumption may be assessed more formally using a hypothesis test such as the Grambsch and Therneau test described by Hosmer and Lemeshow (1999: Chapter 6).

If the PH assumption does not hold there are several ways that we can proceed. Suppose we find that the effect of x varies with t . One option is to split the time axis into sections within which the PH assumption is reasonable. This approach is often used in studies of child mortality where the effects of covariates such as gender vary according to the period of childhood. The first five years may be partitioned into three periods: neonatal (the first month), postneonatal (1-11 months), and child (1-5 years). A separate analysis would then be carried out for each period. A disadvantage of this approach, however, is that it will be inefficient if the effects of some covariates are the same for each period. Moreover, if separate models are fitted it is not possible to test for equality of covariate effects across periods. Another option to handle non-proportional hazards is to fit a stratified Cox model, where the strata are defined by the covariate(s) with non-proportional effects. The stratified model allows the form of the baseline hazard to vary across strata, while the effects of other covariates are assumed time-invariant. An alternative approach is to fit a discrete-time model. Using a discrete-time approach it is straightforward to test and allow for non-proportional hazards (see Section 4).

3.5 Time-varying covariates

Thus far we have assumed that the values of covariates do not change over the observation period. In many applications there will be a mixture of time-constant and time-varying covariates which we

collectively denote by $\mathbf{x}(t)$. Kalbfleisch and Prentice (1980: Chapter 5) classify time-varying covariates as either *internal* or *external*. Internal variables are individual-specific and therefore require longitudinal data at the individual level, or retrospectively collected information from which a time-varying covariate may be derived. The value of an external time-varying covariate at a particular point in time applies to the whole sample, or to subgroups of individuals in the sample. For example, in a study of marital dissolution an internal time-varying covariate might be household income, while an external variable could be regional divorce rates or property prices. Although event history models that allow for time-varying covariates make no distinction between internal and external variables, it is important to consider the nature of any time-varying covariate before including it in a model and interpreting the results.

As in any other regression model, covariates in an event history model are assumed *exogenous*. A variable is said to be exogenous if its values are determined outside the system under study. In contrast, an *endogenous* variable is jointly determined with the outcome variable; that is, the two variables are influenced by a common, or correlated, set of unobserved variables. Time-varying covariates, particularly internal variables, are highly likely to be endogenous. In a study of marital dissolution, for example, there may be unobserved factors influencing both household income at t and the risk of dissolution at the same point in time, such as family illness. The effects of potentially endogenous variables must be interpreted with caution, and certainly not in a causal way. We return briefly to the issue of endogeneity in Section 8.3.

Particular care should be taken when interpreting the effects of time-varying covariates on survival probabilities. A common way to demonstrate the magnitude of a covariate effect is to show estimated survival curves for each value of x , holding constant the values of other covariates, but this may be misleading for time-varying covariates. To illustrate the problem, suppose we have a time-varying covariate $x(t)$ which takes the values 0 and 1. If we compare estimates of $S(t)$ corresponding to $x(t)=0$ and $x(t)=1$ we are in fact comparing two extreme groups of individuals: one whose values on $x(t)$ are fixed at 0 *for the whole observation period*, and another whose values are fixed at 1. At least one of these extremes may be rare or unobserved in the sample or even theoretically impossible. A more meaningful comparison would involve calculating estimates of the survivor function for typical trajectories in $x(t)$.

All major statistical software packages can fit the Cox PH model with time-varying covariates. However, it can be awkward to specify time-varying covariates and the way in which this is done is software dependent. It is important to think carefully about the information needed to construct the time-varying covariate when preparing event history data for analysis, and to check the requirements of the software to be used. To give an example, suppose we wish to include a time-varying indicator of enrolment in full-time education, $\text{FULLTIME}(t)$, in our study of age at first partnership. This indicator is derived by comparing an individual's current age (t) with the age they left school (AGELEFT):

$$\text{FULLTIME}(t) = \begin{cases} 0 & t > \text{AGELEFT} \\ 1 & t \leq \text{AGELEFT}. \end{cases}$$

Note that in most software, $\text{FULLTIME}(t)$ will be stored internally, i.e. it does not appear in the data file. This is because the data file will consist of one record per individual while $\text{FULLTIME}(t)$ varies within a woman.

3.6 Example: age at first partnership

In our analysis of age at first partnership in the NCDS, we consider the effects of educational enrolment and the following time-constant categorical covariates: gender, region of residence, and father's social class. The results from fitting a Cox model are given in Table 2.

The 95% confidence interval for each relative risk, $\exp(\beta)$, provides a test of the null hypothesis of no effect. The null is rejected at the 5% level if the confidence interval does not contain the value 1. Based on this criterion, we find that the effects of gender and educational enrolment are both significant at the 5% level. In general, a likelihood ratio test is preferred. A likelihood ratio test is used to compare a pair of nested models. For example, to test the significance of father's social class, we compare the following two models: the model shown in Table 2 and the model without social class. The test statistic is the difference between the -2 log-likelihood values for the two models, which is compared to a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between models (here, the number of social class categories minus 1). The test statistic for comparing models with and without social class is 3.3 on 2 degrees of freedom ($p=0.192$). We therefore conclude that social class has no effect on age at first partnership. (See Yamaguchi (1991: Chapter 2) for further discussion and examples of likelihood ratio testing.)

The hazard of first partnership is almost 1.5 times higher for women than for men, which implies that women partner at an earlier age. Enrolment in full-time education is associated with delayed partnership formation: being in education reduces the hazard of partnering by a factor of $(1-0.337)\times 100\%=66.3\%$. However, we should hesitate to interpret the effect of educational enrolment as causal because the decisions about when to leave education and when to partner are likely to be jointly determined, i.e. enrolment is potentially endogenous.

4. Discrete-time Event History Models

4.1 The discrete-time approach

Although events in the process under study can theoretically occur at any point in time, durations are usually measured in discrete time units such as months or years, particularly when collected retrospectively. For this reason it may be considered more natural to use a model for discrete event times. The discrete-time approach has several advantages over continuous-time methods. One potential problem with continuous-time models, particularly when durations are measured in reasonably broad time intervals, is the assumption that only one event can occur at any given point in time t . Estimation procedures for continuous-time models need to be adapted if there are tied event times (see Hosmer and Lemeshow (1999: Chapter 3) for a discussion). It is also straightforward to allow for non-proportional hazards and time-varying covariates in a discrete-time model. Finally, discrete-time models are essentially logistic regression models which are already familiar to most social scientists.

A disadvantage of the discrete-time approach is the need for further data manipulation prior to analysis. The data structure required for a discrete-time analysis is described in Section 4.2 but, briefly, a sequence of binary responses is generated from each event time. The increased size of the dataset can lead to long computational times for complex models with random effects (Sections 5, 6 and 8), although a strategy for minimising the number of records is described in Section 4.3. However, an advantage of the user carrying out this extra stage of data manipulation is that discrete-time methods are more transparent than continuous-time methods. In fact a similar restructuring of the data is carried out when a Cox model is specified, but in most statistical software this is done internally and concealed from the user. While this reduces the data management burden, it can also make the mechanics of fitting a Cox model more difficult to understand.

4.2 Data structure for a discrete-time analysis

The first step in a discrete-time analysis is to expand the data file as follows. From the observed data, the event/censoring time y_i and the censoring indicator δ_i , we create, for each time interval t up to y_i , a binary response y_{it} which is coded as follows:

$$y_{it} = \begin{cases} 0 & t < y_i \\ 0 & t = y_i, \delta_i = 1 \\ 1 & t = y_i, \delta_i = 0. \end{cases}$$

For example, if an individual has an event during the third time interval of observation their discrete responses will be $(y_{1i}, y_{2i}, y_{3i}) = (0, 0, 1)$, while someone who is censored at $t=3$ will have response vector $(0, 0, 0)$. The restructured dataset is often called a *person-period file*. Figure 5 shows how the records for two individuals in the person-based file for age at first partnership are converted to person-period format. The original file contained a variable for age at leaving school (AGELEFT) which is used to create the time-varying covariate FULLTIME, as described in Section 3.5.

4.3 Discrete-time models

The discrete-time hazard for interval t is the probability of an event during interval t , given that no event has occurred in a previous interval, i.e.

$$h_{it} = \Pr(y_{it} = 1 \mid y_{si} = 0, s < t),$$

which is the usual response probability for a binary variable.

Therefore, after restructuring the data into person-period format, the event indicator y_{it} can be analysed using any model appropriate for binary responses. A commonly used model is the logit model:

$$\text{logit}(h_{it}) = \log\left(\frac{h_{it}}{1 - h_{it}}\right) = \alpha(t) + \boldsymbol{\beta}' \mathbf{x}_{it}, \quad (4.1)$$

where the covariates \mathbf{x}_{it} may be fixed or time-varying and $\alpha(t)$ is some function of t , which we refer to as the baseline logit-hazard. In a discrete-time model the form of $\alpha(t)$ needs to be specified by the analyst, usually after inspection of a plot of the hazard function (see Section 2.2). If the hazard is approximately linear, then a linear function $\alpha(t) = \alpha_0 + \alpha_1 t$ may be fitted by including t as an explanatory variable in the model. A quadratic function is fitted by including both t and t^2 in the model. The most flexible form for $\alpha(t)$ is a step function which is specified by treating t as a categorical variable. Suppose that we group t into g categories, and define a dummy variable D_j for each category j then the baseline logit-hazard takes the form $\alpha(t) = \alpha_1 D_1 + \alpha_2 D_2 + \dots + \alpha_g D_g$. Alternatively, but equivalently, we can define an overall intercept term and $g - 1$ dummies. Either parameterisation leads to a piecewise-constant hazards model where the hazard is assumed constant within each of the g categories.

In (4.1) a coefficient β is interpreted as the effect of a 1-unit change in a covariate x on the log-odds of an event in interval t . The exponent of β is interpreted as the multiplicative effect of x on the odds or an *odds ratio*. As for the Cox PH model, we can test the null hypothesis that $\beta = 0$ using a t-test, likelihood ratio test, or by examining confidence intervals for β or $\exp(\beta)$.

Model (4.1) is commonly known as the *proportional odds* model because, as in the proportional hazards model, the effects of covariates are assumed constant across the observation period. It is straightforward to relax the assumption of proportionality. To allow the effect of a covariate x to depend on t , we simply include as an additional explanatory variable the interaction between x and t , or whatever functions of t are contained in $\alpha(t)$. Thus we can allow the shape of the baseline hazard to depend on x .

As noted in Section 4.1, a potential drawback of the discrete-time approach is the size of the person-period dataset, especially if the width of the intervals t is short relative to the length of the observation window. One way to reduce the number of records is to use broader intervals, and to weight by the exposure time within each interval. To take an example, suppose that an individual has an event in the 10th month of observation. Rather than creating 10 records, one for each month, we could group time into six-month intervals. The individual would then contribute two records with response vector $y_{it}=(0,1)$ and exposure time vector $n_{it}=(6,4)$. The responses y_{it} now follow a binomial distribution with denominator n_{it} , and (4.1) becomes a binomial logit model. The denominators can be thought of as weights. The grouped intervals do not need to be of equal width. For example, we might have longer intervals at the start and end of the observation period if few events occur at those times. Aggregating intervals leads to minimal loss of information provided that, within each grouped interval, the hazard is fairly constant and it is reasonable to fix the values of time-varying covariates, for example to their values at the start of the interval.

4.4 Example: reanalysis of age at first partnership

Table 3 shows the results of fitting a discrete-time logit model to age at first partnership in the NCDS. Prior to the analysis, the event time for each individual has been converted to a sequence of yearly observations as shown in Figure 5. The first decision in a discrete-time analysis is how to specify the baseline logit-hazard. After examining a plot of the hazard (Figure 2), it was decided to fit $\alpha(t)$ as a quadratic function in age (t) by including t and t^2 as explanatory variables in the model.

For comparison, the results from the Cox model are also shown in Table 3. Although there are some differences between models in terms of statistical significance (using p-values) the conclusions from the two models are broadly similar. The values of $\exp(\hat{\beta})$ now have a different interpretation, however; in a logit model they are interpreted as odds ratios rather than relative risks. Thus, for example, the odds of partnering at age t are 1.61 times higher for women than for men.

When the hazard is very small, as is often the case when an event is rare or time intervals are narrow, the estimated coefficients from the logit and Cox models will be much closer than in this example. This is because as $h_t \rightarrow 0$ the odds ratio $h_t/(1-h_t) \rightarrow h_t$, and therefore the left-hand sides of (3.3) and (4.1) become closer. For this reason, the discrete-time logit model can be viewed as an approximation to the Cox model.

5. Unobserved Heterogeneity

5.1 The problem and implications of unobserved heterogeneity

In an attempt to capture variation in the hazard rate across individuals in the population we include covariates in an event history model. It is unlikely, however, that we will include all important variables. There may be variables we would like to include but which are not available in our dataset, and there will be others whose importance is unknown. Variability between individuals that is due to unmeasured characteristics is known as *unobserved heterogeneity*. As in any regression analysis, omitted variables are a source of model misspecification. If unmeasured variables are correlated with the covariates included in the model, failure to account for unobserved heterogeneity will lead to biased parameter estimates. Furthermore, allowing for unobservables often leads to a change in the shape of the baseline hazard rate. To illustrate how the baseline hazard might change, suppose that the study population consists of two subpopulations whose hazards are both constant over time, but with one group having a higher hazard than the other. If we can distinguish these groups using a dummy variable which we include as a covariate, the baseline hazard rate estimated from the model will be constant. Now suppose that the variable defining these groups is unobserved. In that case, the estimated hazard will be based on a mixture of two latent groups with different hazard rates. This aggregate baseline hazard will be a decreasing function of time because individuals with a high hazard have the event early so that, over time, the risk population increasingly consists of individuals from the lower risk group. Blossfeld and Rohwer (2002: Chapter 10) provide several numerical examples which illustrate this point further.

5.2 Incorporating unobserved heterogeneity

The standard approach to allow for unobserved heterogeneity is to include in the model a *random effect*, also known as *frailty*, which represents unobserved risk factors that are specific to an individual and fixed over time. We usually assume some distributional form for these random effects. Here, we focus on the Cox model with gamma distributed frailty, and the discrete-time logit model with normal frailty. It is also possible to fit non-parametric models, in which the random effects follow a completely flexible discrete distribution, but these methods tend to be implemented via specialist programs written by researchers for their own use rather than mainstream statistical software.

One important consideration when fitting random effects frailty models is that the regression coefficients have a different interpretation to those in the models described thus far. This difference in interpretation is illustrated in Section 6.4. Another note of caution in the application of frailty models is that results can vary according to the type of model fitted, the estimation algorithm used, and the distribution assumed for frailty. See Blossfeld and Rohwer (2002: Chapter 10) for further discussion. Unobserved heterogeneity is usually better identified, and sometimes *only* identified, when there are groups which share the same frailty (i.e. random effect value). These groups might be geographical areas or institutions which share unobserved characteristics, or individuals with repeated events where, for each event, the hazard depends on individual-specific unobservables (see Section 6).

5.2.1 The Cox model with frailty

The model of (3.2) may be generalised to

$$h(t; \mathbf{x}_i, u_i) = h_0(t) u_i \exp(\boldsymbol{\beta}' \mathbf{x}_i) \quad (5.1)$$

where u_i is the random effect or frailty for individual i . In (5.1) frailty acts multiplicatively on the hazard. Therefore an individual with $u_i > 1$ has an above-average hazard (i.e. they are more “frail”) while someone with $u_i < 1$ has a below-average hazard. Because the hazard function must be greater than zero, u_i must be a positive quantity. For this reason, and for mathematical convenience, u_i is most often assumed to follow a gamma distribution with mean 1 and variance σ_u^2 . See Hosmer and Lemeshow (1999: Chapter 9) for further discussion of proportional hazards models with gamma frailty. As noted by Hosmer and Lemeshow, one practical obstacle to the application of (5.1) has been the lack of available software. However, the Cox and parametric survival models with gamma frailty can now be fitted in Stata.

5.2.2 The discrete-time logit model with frailty

In a discrete-time model frailty is usually incorporated by including a normally distributed random effect in the linear predictor. An extension of the logit model of (4.1) which allows for unobserved heterogeneity is given by

$$\text{logit}(h_{ii}) = \alpha(t) + \boldsymbol{\beta}' \mathbf{x}_{ii} + u_i \quad (5.2)$$

where u_i is a random effect for individual i . We assume that the random effects follow a normal distribution with zero mean and variance σ_u^2 . The random effect variance is interpreted as the variance between individuals that is due to unobserved time-invariant characteristics, i.e. residual variance. Model (5.2) can be fitted using any software for random effects logistic regression, including Stata, MLwiN and SAS.

5.3 Example: age at first partnership

Age at first partnership was reanalysed using frailty models. However, we do not show results from this analysis because the estimates were found to vary substantially according to the type of model fitted and the estimation algorithm used to fit the discrete-time logit model. In the case of the Cox model, a likelihood ratio test failed to reject the null that $\sigma_u^2=0$ ($p=0.497$) and the estimates obtained from the frailty model were identical to those from the standard model (shown in Table 2). Random effects logit models were fitted in Stata, which uses numerical quadrature, and MLwiN, using quasiliikelihood and Markov chain Monte Carlo (MCMC) methods¹. The results from using these three alternative estimation procedures were quite different, suggesting that the model is poorly identified when applied to these data. In contrast, when random effects models are fitted to repeated events data (see Section 6), the results are broadly consistent across estimation procedures. Therefore, for an example of the interpretation of random effects models, see Section 6.4.

6. Repeated Events

6.1 Examples

Many events that we study in social research may occur more than once to an individual over the observation period. For example, individuals may move in and out of co-residential relationships multiple times, they may have more than one child, and they may change job repeatedly. If repeated events are observed we can model the duration of each *episode*, where an episode is

¹ See Rasbash et al. (2004: Chapter 9) for a brief discussion of the different estimation algorithms for fitting random effects models to discrete response data.

defined as a continuous period during which an individual is at risk of experiencing a particular event. When an event occurs, a new episode begins and the duration ‘clock’ is reset to zero.

6.2 Issues in the analysis of repeated events

The simplest way to handle repeated events is to model each event separately. For example, in a study of marital dissolution one might fit separate models for first and second marriages. Such an approach is inefficient because it is likely that some of the determinants of dissolution may operate in the same way for first and subsequent marriages. Furthermore, if some of the characteristics that affect each event occurrence are unobserved, the results from modelling repeated events independently may be misleading. Kravdal (2001) compared results from separate and joint random effects models in a study of the duration of second and third birth intervals in Norway. Based on separate models for each birth order, both controlling for age, a counterintuitive positive effect of education was found, i.e. more educated women apparently had shorter birth intervals. When birth intervals were analysed using a joint random effects model, the effect of education became negative. Kravdal argued that the implausible result obtained from the separate models was due to selection: educated and uneducated women having their second child at the same age differ on unobservables which affect a woman’s chance of conception whatever the birth order.

A preferred approach is to model repeated events jointly. However, in doing so, it is important to recognise the potential for correlation between the duration of episodes from the same individual. Event times will be correlated if there are unobserved individual characteristics, fixed over time, which affect an individual’s hazard in all episodes. For example, some individuals may have an increased risk of separation from any partner because of unmeasured personality traits. The most common way of allowing for correlation between repeated episodes is to include in the model a random effect representing unobserved risk factors shared by each of an individual’s episodes, sometimes referred to as *shared frailty*.

One advantage of pooling repeated episodes is that it allows us to test explicitly whether variables that predict the occurrence of the first event also impact on subsequent events. This is achieved by including dummy variables for event order in the model and testing whether they interact with covariates. A joint modelling approach using a shared frailty model also permits investigation of the effects of previous events on the likelihood of later events of the same type. For example, does the experience of marital separation place an individual at a higher or lower risk of dissolution should they remarry?

6.3 A discrete-time model for repeated events

When events are repeatable, event history data have a two-level hierarchical structure with episodes (level 1) nested within individuals (level 2). Thus repeated events may be analysed using multilevel models. The two-level random effects logit model for repeated events, or shared frailty model, takes the same form as the frailty model in (5.2). If we denote by h_{ij} the hazard of an event in interval t of episode j of individual i , the model may be written as

$$\text{logit}(h_{ij}) = \alpha(t) + \beta' \mathbf{x}_{ij} + u_i \quad (6.1)$$

where covariates \mathbf{x}_{ij} may vary across time intervals (i.e. time-varying covariates), episodes, or individuals. After controlling for the individual-specific unobservables represented by the random effect, we assume that the durations of episodes for the same individual are independent.

In multilevel modelling terminology, model (6.1) is a two-level *random intercept model*. In such a model, the log-odds of an event in interval t is shifted up or down by an amount u_i for a given individual but the effects of duration and covariates are assumed to be constant across individuals. In a more general random coefficient model, the effects of the predictor variables may vary randomly across individuals. Random coefficient models can be fitted using any multilevel modelling software. It is also straightforward to extend to further hierarchical levels, for example to allow for area effects.

6.4 Example: birth intervals of Hutterite women

The application of model (6.1) is illustrated in an analysis of birth intervals among Hutterite women. The Hutterites are a North American population who do not use contraceptives and therefore have a very high fertility rate (see McDonald and Rosina (2001) for further analyses of these data). The analysis sample contains 159 women who contribute 944 birth intervals. The analysis is restricted to closed birth intervals (i.e. those that had ended in a conception before the survey) to avoid difficulties in analysing long intervals from sterile women. (We return to this issue briefly in Section 8.2.) We will carry out a discrete-time analysis with intervals of one-month width. After expanding the data to person-period format, there are 8361 monthly observations.

An episode corresponds to the number of months between a birth and the conception of the next child. The response variable y_{ij} indicates whether a conception occurs in month t of birth interval j of woman i . The baseline hazard $\alpha(t)$ is specified as a step function, which is fitted by including as explanatory variables dummies for grouped birth interval duration (with categories <6, 6-11, 12-23, 24-35 and 36+ months). The following covariates, all defined at the episode level, are also included: maternal age and marital duration at the start of the birth interval, and an indicator of whether the last born child died within one year.

Table 4 shows the results of fitting a two-level random effects logit model to the Hutterite data. These results were obtained using quasilielihood methods (2nd order PQL) in MLwiN but Stata, which uses numerical quadrature, produces very similar results. The results from a model which ignores unobserved heterogeneity, and therefore treats the durations of birth intervals for the same woman as independent (conditional on covariates), are shown for comparison. The between-woman residual variance is estimated as 0.310 and is large relative to its standard error of 0.061. When the same model is fitted via maximum likelihood (e.g. in Stata) we obtain a likelihood ratio test statistic of 63.8 (1 d.f.; $p < 0.001$). We therefore conclude that there is strong evidence of residual heterogeneity between women.

There are some large differences between the two sets of parameter estimates in Table 4. Although failure to account for unobserved heterogeneity may have led to biased estimates in the standard logit model, it is important to note that the parameters in the two models have a different interpretation. Regression coefficients in the standard model (i.e. without frailty) have a *population average* or *marginal* interpretation, while coefficients in a random effects model have a *cluster-specific* or *conditional* interpretation. Suppose we have a covariate x with population average coefficient β_{PA} and cluster-specific coefficient β_{CS} . Here, a ‘cluster’ corresponds to a woman since we have birth intervals nested within women. β_{PA} is the effect of a 1-unit increase in x in the population, after adjusting for any other covariates included in the model. β_{CS} is also the effect of a 1-unit increase in x after adjusting for covariates, but it is the effect among women with the same value of u_i , i.e. sharing the same unobserved characteristics. If x varies within women, i.e. it is time-varying or defined at the episode level, a more meaningful interpretation of β_{CS} is as the effect of a within-woman change in x .

We will illustrate the interpretation of population average and cluster-specific coefficients using the results in Table 4. Consider the effect of the death of the last born child on the succeeding birth interval (adjusting for the effects of maternal age and marital duration). From both models we obtain a positive estimate which might be due to a ‘replacement’ effect whereby couples attempt to compensate for a child death by having another child quickly. The odds ratio estimated from the standard model is $\exp(0.272) = 1.31$. Thus, *on average*, the odds of a conception among women who have recently experienced the death of their youngest child are 1.31 times higher than for women whose youngest has survived. The corresponding odds ratio from the random effects model is $\exp(0.460) = 1.58$ which is interpreted as the effect of a child death on the odds of a conception for a *given woman*; a woman’s odds of a conception increase by 58% if her youngest dies during infancy. For covariates that are fixed across the observation period, i.e. defined at the woman level, cluster-specific effects can be difficult to interpret. What does it mean to talk of an effect of x among women with the *same random effect value*? An alternative way to interpret the results of a logit model is to compute a predicted response probability (hazard) for each value of x , holding the values of other covariates constant. Rasbash et al. (2004: Chapter 9) describe how predicted probabilities can be calculated from a random effects model using a simulation approach, in which probabilities are averaged over repeated draws from the estimated random effects distribution. The probabilities obtained using this method are population averaged, and therefore comparing probabilities for different values of x (or transforming them to odds ratios) is equivalent to interpreting population average coefficients.

7. Competing Risks

7.1 Introduction

So far we have assumed that there is only one possible transition available to an individual. In many situations, there may be more than one transition or type of event which leads to the termination of an episode. Multiple types of event are commonly referred to as *competing risks*. The different event types are said to be ‘competing’ because they are mutually exclusive; only one type of transition can occur in a given episode, after which the episode ends. Examples of competing risks are causes of death and, in studies of the duration of employment episodes, different reasons for leaving a job (e.g. redundancy, dismissal, a new job opportunity, retirement or maternity leave).

In a competing risks analysis, we examine the relative frequency of different types of event and the dependency of the event-specific hazards on covariates. Different event types may have different determinants or the effects of the same covariates may be event specific. For example, having specialist qualifications may decrease the risk of redundancy and dismissal and increase the chance of moving to a new job; other covariates may affect only a subset of the possible events. We can use competing risks analysis to explore such hypotheses.

7.2 Definitions

As before we assume that event times are realisations of a random variable T . In addition we observe, for each individual, an indicator of the type of event which we assume is a realisation of another random variable R . The hazard function for event type r ($r = 1, \dots, k$), sometimes called the cause-specific hazard, is defined as

$$h^{(r)}(t) = \lim_{dt \rightarrow 0} \frac{\Pr(t \leq T < t + dt, R = r | T \geq t)}{dt}.$$

The unconditional probability that an event of type r occurs in the interval $(t, t + dt)$ is the *event-specific density function*

$$f^{(r)}(t) = \lim_{dt \rightarrow 0} \frac{\Pr(t \leq T < t + dt, R = r)}{dt} = h^{(r)}(t) S(t),$$

where $S(t) = \Pr(T \geq t)$ as before.

Another useful quantity in competing risks analysis is the probability that a event of type r occurs before t , which is the *event-specific cumulative density function*

$$\begin{aligned} F^{(r)}(t) &= \Pr(T < t, R = r) \\ &= \int_0^t f^{(r)}(s) ds. \end{aligned}$$

7.3 Models for competing risks

For each individual i we observe an event or censoring time y_i and a categorical variable r_i which indicates if they were censored ($r_i = 0$) and, if uncensored, the type of event they experience ($r_i = 1, 2, \dots, k$).

A common way of formulating competing risks is in terms of a set of potential (unobserved) event times. We denote by $t_i^{(r)}$ ($r = 1, \dots, k$) the time at which event type r occurs to individual i . For a given individual, the occurrence of one type of event removes them from the risk set for other events so, for uncensored individuals, we observe $t_i = \min(t_i^{(1)}, t_i^{(2)}, \dots, t_i^{(k)})$. Most models, including those described below, assume that the latent event times are independent. While we may suspect that this assumption is questionable in some situations, e.g. for causes of death with similar risk factors, models that allow for dependency between $t_i^{(r)}$ are unidentified except under certain conditions (see Crowder (2001) for further discussion).

7.3.1 Separate models for each type of transition

Under the assumption of independent $t_i^{(r)}$, it can be shown that a competing risks model may be fitted as a set of k independent models, one for each type of event. In the model for event type r all events other than r are treated as censored, so we work with a censoring indicator $\delta_i^{(r)}$ which is defined as

$$\delta_i^{(r)} = \begin{cases} 1 & \text{if } r_i \neq r \\ 0 & \text{if } r_i = r. \end{cases}$$

We can then analyse $(y_i, \delta_i^{(r)})$ using any model that can be applied to single event data. For example, the Cox proportional hazards model of (3.3) generalises to

$$\log h^{(r)}(t; \mathbf{x}_i^{(r)}(t)) = \log h_0^{(r)}(t) + \boldsymbol{\beta}^{(r)'} \mathbf{x}_i^{(r)}(t), \quad r = 1, \dots, k. \quad (7.1)$$

Eq. (7.1) specifies a very general competing risks model in which the baseline log-hazard $\log h_0^{(r)}(t)$, covariates $\mathbf{x}^{(r)}(t)$ and covariate effects $\boldsymbol{\beta}^{(r)}$ may all vary across event types. In practice the covariate vectors $\mathbf{x}^{(r)}(t)$ will usually partially or completely overlap.

In discrete time we define, for each time interval t , a vector of binary responses $\mathbf{y}_{ii} = (y_{ii}^{(1)}, y_{ii}^{(2)}, \dots, y_{ii}^{(k)})$ where $y_{ii}^{(r)}$ is coded as follows:

$$y_{ii}^{(r)} = \begin{cases} 0 & t < y_i \\ 0 & t = y_i, r_i = r \\ 1 & t = y_i, r_i \neq r. \end{cases}$$

The discrete-time event-specific hazard is defined as

$$h_{ii}^{(r)} = \Pr(y_{ii}^{(r)} = 1 \mid \mathbf{y}_{si} = 0, s < t)$$

and the logit model (4.1) for single events becomes

$$\text{logit}(h_{ii}^{(r)}) = \alpha^{(r)}(t) + \boldsymbol{\beta}^{(r)'} \mathbf{x}_{ii}^{(r)}, \quad r = 1, \dots, k. \quad (7.2)$$

The hazard rates estimated under models (7.1) and (7.2) are called *independent* or *gross rates*. The independent rate associated with event type r is often interpreted as the theoretical or underlying rate that would apply if all types of event other than r were eliminated. Independent rates are useful for making comparisons across populations or time periods where the composition of event types differs. For example, in a comparison of cancer mortality rates for the UK and India we should make some adjustment for differences in life expectancy between the two countries because cancer deaths tend to occur at older ages. The use of independent cancer mortality rates allows comparison of the hypothetical risk of a cancer death in the absence of other causes of death, under the assumption of independent cause-specific rates.

7.3.2 Modelling event types simultaneously: the multinomial logit model

In discrete time, an alternative competing risks model is the multinomial logit model. Rather than defining a vector of k binary responses for each in time interval t , we define a single categorical response y_{ii} :

$$y_{ii} = \begin{cases} 0 & t < y_i \\ r_i & t = y_i. \end{cases}$$

The event-specific hazard is

$$h_{ii}^{(r)} = \Pr(y_{ii} = r \mid \mathbf{y}_{si} = 0, s < t), \quad r = 1, \dots, k$$

and the hazard that no event of any event occurs during interval t is given by

$$h_{ii}^{(0)} = \Pr(y_{ii} = 0 \mid \mathbf{y}_{si} = 0, s < t) = 1 - \sum_{r=1}^k h_{ii}^{(r)}. \quad (7.3)$$

The discrete-time multinomial logit model is given by

$$\log\left(\frac{h_{ii}^{(r)}}{h_{ii}^{(0)}}\right) = \alpha^{(r)}(t) + \boldsymbol{\beta}^{(r)'} \mathbf{x}_{ii}^{(r)} \quad r = 1, \dots, k. \quad (7.4)$$

The multinomial logit model consists of k equations, each contrasting the risk of one type of event with the risk that no event occurs.

The exponent of a regression coefficient $\exp(\beta^{(r)})$, sometimes called a *relative risk ratio*, is interpreted as the multiplicative effect of a 1-unit increase in x on the risk of event type r versus the risk that no event occurs. As an alternative to relative risk ratios, which can be awkward to interpret, it is often helpful to calculate predicted event-specific hazards or cumulative density functions. Rearrangement of (7.4) leads to the following expression for the event-specific hazard functions:

$$h_{ii}^{(r)} = \frac{\exp[\alpha^{(r)}(t) + \boldsymbol{\beta}^{(r)'} \mathbf{x}_{ii}^{(r)}]}{1 + \sum_{l=1}^k \exp[\alpha^{(l)}(t) + \boldsymbol{\beta}^{(l)'} \mathbf{x}_{ii}^{(l)}]}, \quad r = 1, \dots, k, \quad (7.5)$$

with $h_{ii}^{(0)}$ calculated by subtraction as in (7.3).

The survivor function is

$$\hat{S}_{1i} = 1; \quad \hat{S}_{ii} = \prod_{j=1}^{t-1} (1 - \hat{h}_{ji}), \quad t > 1, \quad (7.6)$$

where $\hat{h}_{ji} = \sum_{r=1}^k \hat{h}_{ji}^{(r)}$.

Estimates of the event-specific density and cumulative density functions can be calculated as follows:

$$\hat{f}_{ii}^{(r)} = \hat{h}_{ii}^{(r)} \hat{S}_{ii}, \quad (7.7)$$

$$\hat{F}_{ii}^{(r)} = \sum_{j=1}^{t-1} \hat{f}_{ji}^{(r)}. \quad (7.8)$$

The hazard rates estimated under model (7.4) are often called *dependent* or *net rates*. The estimated hazard $\hat{h}_{ii}^{(r)}$ represents the risk of an event of type r in the presence of all other types of event, i.e. the observed rates. Comparisons of dependent rates across populations should be interpreted with caution. It would be misleading, for example, to compare dependent cancer mortality rates for the UK and India. The dependent rate would be higher for the UK, but this is at least partly explained by the lower life expectancy in India which means that a large proportion of the population die from other causes before they are at risk of cancer. If age-specific cause-of-death data are unavailable, independent rates may be of greater interest (see also Section 7.3.1).

7.4 Application of the multinomial logit model to contraceptive use dynamics

The multinomial logit model (7.4) is applied in a study of the duration of episodes of contraceptive use among Indonesian women. An episode is defined as a continuous period of using the same contraceptive method. We distinguish between two types of event which lead to the end of an episode: a transition to non-use (referred to as discontinuation), and a transition to another method (a method switch). The analysis sample consists of 17,843 episodes from 12,595 women. (We ignore any correlation between the duration of episodes from the same woman.) The data are in discrete-time format with one record per six-month time interval of contraceptive use, leading to a total of 68,525 observations. We consider the effects of five covariates: method used during the episode (categorised as pill/injectable, Norplant®/intrauterine device (IUD), other modern reversible, traditional), age at the start of the episode (<25, 25-34, 35-49), level of education (none, primary, secondary or higher), type of region of residence (rural, urban) and household socio-economic status (low, medium, high). A more detailed analysis which allows for multiple episodes per woman and incorporates transitions from non-use to use is given in Steele et al. (2004).

The results of the competing risks analysis are shown in Table 5. The model consists of two simultaneous equations, where each equation represents a contrast between a type of event (discontinuation or method switch) and continuing use of the same method (the 'no event' category). The baseline hazard takes the form of a step function, which is fitted by including dummy variables for (grouped) duration. We see that the risk of discontinuation is highest in the first year of use, and fairly constant thereafter. The hazard of switching methods is high in the first six months, then declines sharply. Users of long-term hormonal methods (Norplant® and IUD) are less likely than users of any other method to abandon contraceptive use or to change to a different method. Users of traditional methods (rhythm and withdrawal) are also relatively unlikely to switch methods. Turning to the effects of demographic and socio-economic characteristics, we find that age has a negative effect on both discontinuation and switching, with older women being more likely to continue use of the same method. Urban and educated women are more likely to discontinue or switch methods than are rural and uneducated women. Socio-economic status has different effects on discontinuation and switching: a higher standard of living is associated with a low discontinuation rate, but higher switching rate, possibly reflecting access to a wider choice of methods for better-off women.

8. Further Topics

8.1 Multiple states

Over the course of an event history, one can think of individuals as experiencing different types of event or, equivalently, as passing through different 'states'. For example, one can view the experience of an unemployment 'event' as a transition between two states, employment and unemployment. The models considered thus far consider transitions from a single state although, in a competing risks framework, there may be multiple destination states. After experiencing an event which results in a transition to another state, an individual is no longer observed unless they later re-enter the state of interest. Thus, in an analysis of employment episodes, periods of unemployment are ignored. In a multiple state model, we are interested in the duration spent in more than one state, which may be a subset of all possible states.

A simple way of handling multiple states is to fit a separate model for transitions from each state, using the methods described in the previous sections. For example, a model for the dissolution of cohabiting unions (a transition from the cohabitation state) would be fitted independently of a model for marital breakdown (a transition from the marriage state). There are two major drawbacks to this approach. Suppose we are interested in testing whether the presence of children has the same effect on the risk of separation for married and cohabiting couples. If separate models are fitted for

transitions from marriage and cohabitation, we cannot test whether covariate effects are state-dependent. Another potential problem with the independent modelling approach is that it does not allow for correlation between the unmeasured determinants of transitions from different states. For example, Steele et al. (2005a) found that women with a high risk of separating from a marital partner were also at high risk of separating from a cohabiting partner. This correlation was not fully captured by covariates, and therefore led to a positive residual correlation between the risks of separation for married and cohabiting women. It is particularly important to account for such correlation if one is interested in assessing the effect of a transition from one state on the risk of a subsequent transition from another state (see Steele et al. (2005b) for further discussion.)

A preferred way of handling multiple states, which avoids the problems described above, is to model transitions from each state simultaneously. Random effects models can be useful for representing simultaneous models and are straightforward to implement if a discrete-time model is used. For each state, we fit an equation of the form (5.2), or (6.1) if there are repeated events. The equations are linked by allowing for correlation between the random effects. The first step is to construct a 'pooled' person-period file with one record for each time interval spent in *any* state. Dummy variables are then defined to indicate which state is occupied during interval t and these are included as explanatory variables in the model. The state dummies are interacted with explanatory variables to allow for state-specific covariate effects. State-specific random effects are fitted by allowing the coefficient of each state dummy to vary randomly across individuals. Finally, we estimate pairwise correlations between the random effects to allow for residual correlation across states. Further details are given in Steele et al. (2004).

8.2 Long-term survivors

The methods described in this review assume that the entire population is at risk of experiencing the event of interest throughout the observation period. For many types of event, however, there will be individuals who have a zero hazard and will therefore never experience the event. For example, there are married couples who will never divorce and infertile men and women who cannot have children. Such individuals are commonly referred to as *long-term survivors*. The problem is that the distinction between long-term survivors and the rest of the population is unobservable: while some right-censored individuals may be long-term survivors, others may have the event after the end of the study period. Empirical evidence of a surviving fraction is heavy censoring at long durations.

Mixture models, also called mover-stayer, split population, or cure models, have been developed to allow for long-term survivors. These models assume a zero hazard rate among long-term survivors throughout the observation period. Typically, a mixture model consists of a logistic regression model for event occurrence combined with an event history model for event timing, conditional on event occurrence. Because membership of the long-term survivor group is unobserved for sample members not experiencing the event in the observation period, the dependent variable in the logistic regression model is only partially observed. For this reason, special estimation procedures, such as the EM algorithm, are required. At the time of writing, mixture models have not been implemented in the major statistical packages; most authors have written their own software. However, a simple mixture model, in which the probability of long-term survivorship is assumed constant rather than depending on covariates, can be estimated using the SABRE program.

For further discussion of mixture models for long-term survivors, see Li and Choe (1997) and McDonald and Rosina (2001). Li and Choe (1997) use a mixture model to study second birth intervals in China where, as a result of the one-child policy, a substantial proportion of couples will never have a second child. McDonald and Rosina (2001) apply a random effects mixture model in an analysis of Hutterite birth intervals, where long-term survivors are sterile women.

8.3 Correlated event histories

So far we have considered methods for studying the timing of events in a single history or ‘process’. Examples of events and processes are births in the fertility process, changes of job in the employment process, and the formation and dissolution of marriages and cohabitations in the partnership process. Typically outcomes of one process will influence the occurrence of events in another process. For example the presence of children, who constitute prior outcomes of the fertility process, is often found to be negatively associated with the risk of partnership dissolution. Previous researchers have explored the relationship between childbearing and dissolution by including the number of children as a covariate in a model for dissolution. This approach, however, ignores the possibility that decisions about childbearing and partnerships are subject to shared influences, some of which will be unobserved. In other words, fertility outcomes may be endogenous with respect to partnership transitions. Lillard and Waite (1993) developed a simultaneous equations model to allow for the joint determination of marital stability and fertility in the US. Using this approach they found that some women were more likely than others to have unstable marriages (due to unmeasured time-invariant characteristics), and that such women were less likely to have children during marriage. Failure to take into account the simultaneity between processes led to biased estimates of the effects of having children on the risk of marital dissolution.

Steele et al. (2005a) describe a discrete-time model for multiple, correlated processes which can be fitted using MLwiN. The model can be framed as a multivariate response model. For each time interval, we have a response corresponding to each event history. For example, to study the link between marital dissolution and marital fertility we would have a bivariate response for each time interval of marriage: the first response indicates whether the marriage has dissolved, and the second indicates whether a conception (or birth) has occurred. The bivariate response model consists of two equations, one for each response. Each equation includes individual-specific random effects which are correlated across equations to allow for residual correlation between the hazards of dissolution and conception.

9. Software

All of the methods described in this review can be implemented in several software packages. Some of these software options are listed here. Note, however, that this section does not provide an exhaustive review of the event history analysis capabilities of any package. Readers are advised to consult the manuals of their package of choice for full details.

Standard models (without random effects). The Cox model (with or without time-varying covariates), and discrete-time logit and multinomial logit models can be fitted in any major statistical package, e.g. SPSS, SAS and Stata. See the NCRM website for SPSS and Stata syntax, together with the datasets used in this review. Syntax for converting an episode-based file to person-period format prior to a discrete-time analysis is also provided.

Random effects models. The Cox frailty model can be fitted in Stata (using `stcox` with the `frailty` option). Random effects logit models can be fitted in Stata (using `xtlogit`) and SAS (using `proc nlmixed`) as well as specialist software for multilevel modelling (e.g. MLwiN). See the NCRM website for Stata syntax and MLwiN instructions. Readers may also find useful a series of software reviews available from the Centre for Multilevel Modelling website (www.mlwin.com/softrev/index.html), which include syntax for fitting random effects models in a range of packages.

10. Further reading

There are numerous books on event history analysis, more commonly referred to as survival analysis outside the social sciences. This section provides a brief commentary on a small selection of those considered most appropriate for social researchers because they have an applied rather than mathematical focus.

- Blossfeld, H.-P. and Rohwer, G. (2002) *Techniques of Event History Modelling: New Approaches to Causal Analysis*. 2nd Edition. New Jersey: Lawrence Erlbaum Associates.

The overarching theme of this book is the use of event history analysis for understanding social processes. The opening chapter outlines the limitations of cross-sectional and panel designs for causal modelling, and discusses the strengths and weaknesses of event history data. The focus is on continuous-time models, with detailed discussion of time-varying covariates. There is a chapter on unobserved heterogeneity. A nice feature of the book is that a single dataset is analysed throughout, with comparisons of results across different models and model specifications. References are given to research papers with more in-depth analysis and substantive interpretation of the same data. Syntax and output from the TDA program are included.

- Hosmer, D.W. and Lemeshow, S. (1999) *Applied Survival Analysis: Regression Modeling of Time to Event Data*. New York: Wiley.

This text gives a more technical coverage of event history analysis than the others reviewed here, and the examples are not from social science. Nevertheless, it is very clearly written and much of the material should be accessible to readers who do not have a strong background in statistics. The focus is on continuous-time models, particularly the proportional hazards model, with a section on random effects (frailty) models for repeated events. There is a whole chapter devoted to methods for assessing model fit. The book is not tied to a particular software package but there is an accompanying Solutions Manual, providing Stata code for all examples, and datasets are available from the publishers' website.

- Yamaguchi, K. (1991) *Event History Analysis*. Applied Social Research Methods Series, Vol. 28. Sage: Newbury Park.

A clear and thorough practical guide to both continuous-time and discrete-time methods, with a range of examples from social research. A particular strength of the book is that each analysis is preceded by a statement of the substantive hypotheses to be tested. Therefore readers can see how a model can be modified to test a particular hypothesis. A full interpretation of the results of each analysis is given. Also included are models for multiple states (or 'two-way transitions'), but there is no discussion of random effects models and only a brief description of competing risks. The book is not linked to a particular software package, although there are examples of SAS and BMDP code.

- Steele, F. (2005) Multilevel Discrete-time Event History Analysis. Lecture notes and practical exercises. Download from www.mlwin.com/team/mmmmpceh.html#workshop

These training materials, written for a two-day workshop on discrete-time event history analysis, are available free of charge. Among the topics discussed are models for multiple states, competing risks and correlated histories. Data and practical exercises in MLwiN (and an introductory exercise in SPSS) can also be downloaded.

References

- Berrington, A. (2003) Change and continuity in family formation among young adults in Britain. *S3RI Applications and Policy Working Paper A03/04*, Southampton Statistical Sciences Research Institute, University of Southampton, UK.
- Blossfeld, H.-P. and Rohwer, G. (2002) *Techniques of Event History Modelling: New Approaches to Causal Analysis*. 2nd Edition. New Jersey: Lawrence Erlbaum Associates.
- Crowder, M. (2001) *Classical Competing Risks*. Boca Raton: Chapman and Hall.
- Guo, G. (1993) Event history analysis of left-truncated data. In P. Marsden (Ed.), *Sociological Methodology*, vol. 23, pp. 217-242. San Francisco: Jossey-Bass.
- Hosmer, D.W. and Lemeshow, S. (1999) *Applied Survival Analysis: Regression Modeling of Time to Event Data*. New York: Wiley.
- Kalbfleisch, J.D. and Prentice, R.L. (1980) *The Statistical Analysis of Failure Time Data*. New York: Wiley.
- Kravdal, Ø. (2001) The high fertility of college educated women in Norway: an artefact of the separate modelling of each parity transition. *Demographic Research*, 5, article 6.
- Li, L. and Choe, M.K. (1997) A mixture model for duration data: analysis of second births in China. *Demography*, 34: 189-197.
- Lillard, L. and Waite, L. (1993) "A joint model of marital childbearing and marital disruption", *Demography*, 30: 653-681.
- McDonald, J.W. and Rosina, A. (2001) Mixture modelling of recurrent event times with long-term survivors: analysis of Hutterite birth intervals. *Statistical Methods & Applications*, 10: 257-272.
- Rasbash, J., Steele, F., Browne, W. and Prosser, B. (2004) *A User's Guide to MLwiN, Version 2.0*. London: Institute of Education.
- Steele, F., Goldstein, H. and Browne, W. (2004) A general multistate competing risks model for event history data, with an application to a study of contraceptive use dynamics. *Statistical Modelling* 4: 145-159.
- Steele, F., Kallis, C., Goldstein, H. and Joshi, H. (2005a) The relationship between childbearing and transitions from marriage and cohabitation in Britain", *Demography*, 42 (to appear).
- Steele, F., Kallis, C. and Joshi, H. (2005b) The formation and outcomes of cohabiting and marital partnerships in early adulthood: the role of previous partnership experience. Working paper. Download from <http://www.mlwin.com/team/mmpceh.html>
- Yamaguchi, K. (1991) *Event History Analysis*. Applied Social Research Methods Series, Vol. 28. Sage: Newbury Park.

Table 1. Life table estimates of the age at first partnership

t	r_t	d_t	w_t	\hat{h}_t	\hat{S}_t
1 (age 16)	500	9	0	0.018=9/500	1
2	491=500-9	20	0	0.042=20/491	0.982=1-0.018
3	471=491-20	32	0	0.068	0.941=0.982×(1-0.042)
.
.
17 (age 32)	39	3	0	0.077	0.078

Note: that the hazard and survivor functions are not estimated for the last interval (age 33) because most individuals who have not partnered by the start of this interval are censored.

Table 2. Results from a Cox proportional hazards analysis of age at first partnership

	$\hat{\beta}$	$\exp(\hat{\beta})$	95% CI for $\exp(\beta)$
Female	0.398	1.489	(1.220, 1.817)
Region			
<i>Scotland and the North</i>	0.238	1.268	(0.939, 1.712)
<i>Wales and Midlands</i>	0.155	1.168	(0.850, 1.606)
<i>Southern and Eastern</i>	0.081	1.084	(0.765, 1.536)
<i>South East, including London*</i>	0	1	-
Father's social class			
<i>I or II (professional and managerial)</i>	-0.288	0.749	(0.549, 1.023)
<i>III</i>	-0.148	0.863	(0.674, 1.104)
<i>IV or V (manual)*</i>	0	1	-
Enrolled in full-time education	-1.089	0.337	(0.225, 0.505)
-2 log-likelihood	4253.1		

*Reference category.

Table 3. Results from fitting discrete-time logit model and Cox model to age at first partnership

	Discrete-time logit model			Cox model	
	$\hat{\beta}$	$\exp(\hat{\beta})$	p-value	$\hat{\beta}$	p-value
Constant	-14.228	-	-	†	-
t	1.033	-	<0.001	†	-
t ²	-0.021	-	<0.001	†	-
Female	0.475	1.608	<0.001	0.398	<0.001
Region					
<i>Scotland and the North</i>	0.289	1.335	0.081	0.238	0.121
<i>Wales and Midlands</i>	0.193	1.212	0.273	0.155	0.339
<i>Southern and Eastern</i>	0.106	1.112	0.579	0.081	0.649
<i>South East, including London*</i>	0	1	-	0	-
Father's social class					
<i>I or II (professional and managerial)</i>	-0.350	0.705	0.044	-0.288	0.069
<i>III</i>	-0.181	0.834	0.194	-0.148	0.240
<i>IV or V (manual)*</i>	0	1	-	0	-
Enrolled in full-time education	-1.137	0.321	<0.001	-1.089	<0.001

*Reference category.

†Parameter not estimated.

Table 4. Results from analysis of repeated birth intervals among Hutterite women, before and after adjusting for unobserved heterogeneity (UH)

	Without UH		With UH	
	Estimate	(SE)	Estimate	(SE)
Constant	-1.591	(0.377)	-1.523	(0.599)
Months since last birth (t)				
<6*	0	-	0	-
6-11	0.567	(0.081)	0.760	(0.084)
12-23	0.680	(0.097)	1.098	(0.106)
24-35	0.145	(0.232)	0.759	(0.247)
36+	0.026	(0.269)	1.016	(0.303)
Maternal age at last birth (years)	-0.016	(0.017)	-0.022	(0.027)
Marital duration at last birth (years)	-0.040	(0.016)	-0.040	(0.025)
Last born child died within a year	0.272	(0.152)	0.460	(0.171)
Random effect variance (σ_u^2)	-	-	0.310	(0.061)

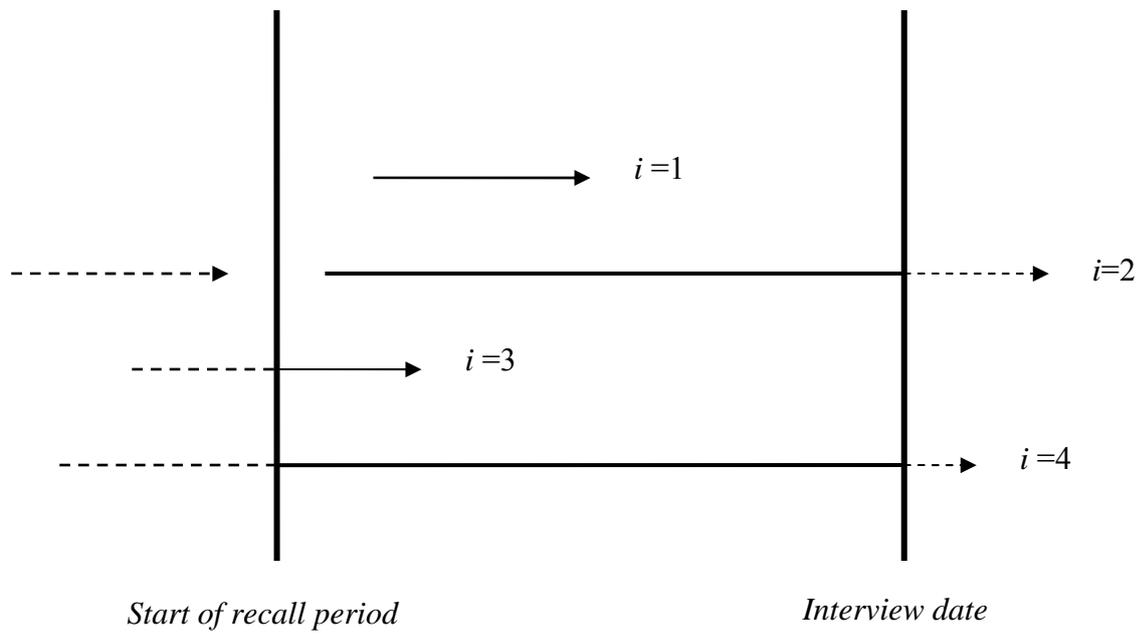
*Reference category.

Table 5. Results from discrete-time multinomial logit analysis of transitions from contraceptive use

	Discontinuation vs. Continue with same method		Method switch vs. Continue with same method	
	Estimate	(SE)	Estimate	(SE)
Constant	-2.205	(0.079)	-3.070	(0.114)
Months of use (<i>t</i>)				
<6*	0	-	0	-
6-11	-0.079	(0.047)	-0.510	(0.050)
12-23	-0.165	(0.044)	-0.830	(0.050)
24-35	-0.105	(0.051)	-0.917	(0.064)
36+	-0.124	(0.056)	-0.917	(0.071)
Method				
Pill/injectables*	0	-	0	-
Norplant®/IUD	-1.190	(0.059)	-1.049	(0.066)
Other modern	0.481	(0.114)	0.380	(0.120)
Traditional	0.027	(0.064)	-0.642	(0.094)
Age at start of episode				
<25*	0	-	0	-
25-34	-0.399	(0.034)	-0.246	(0.039)
35-49	-0.693	(0.059)	-0.330	(0.063)
Education				
None*	0	-	0	-
Primary	0.023	(0.070)	0.457	(0.103)
Secondary+	0.236	(0.073)	0.898	(0.105)
Urban residence (vs. rural*)	0.127	(0.037)	0.094	(0.042)
Socio-economic status				
Low*	0	-	0	-
Medium	-0.110	(0.047)	0.355	(0.064)
High	-0.172	(0.053)	0.282	(0.069)

*Reference category.

Figure 1. Examples of incomplete observation in a restricted recall period



Notes: Arrowheads indicate event times. Dashed lines indicate 'at risk' periods falling outside the observation window which are therefore unobserved.

Figure 2. The hazard of first partnership by age

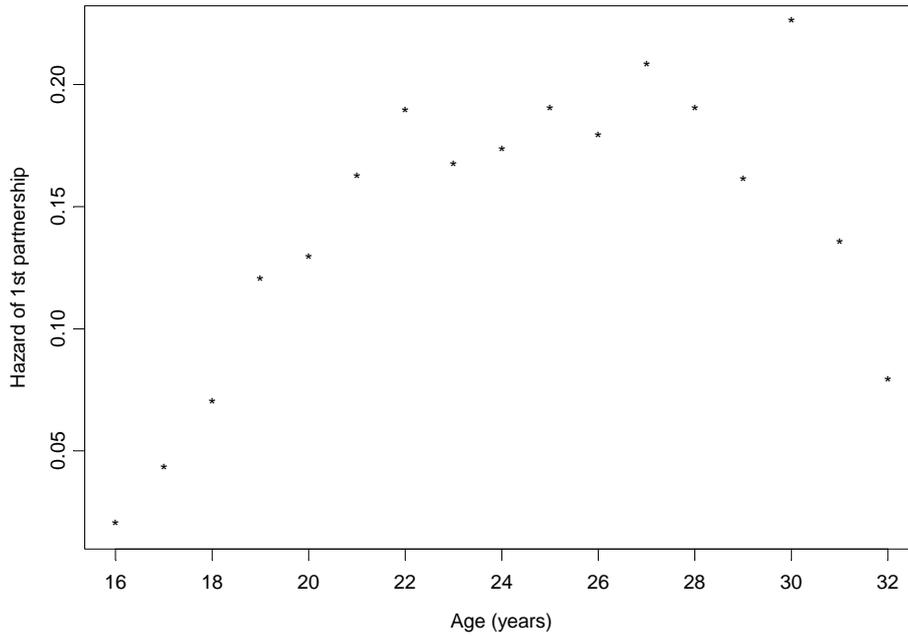


Figure 3. The proportion unpartnered (survival probability) by age

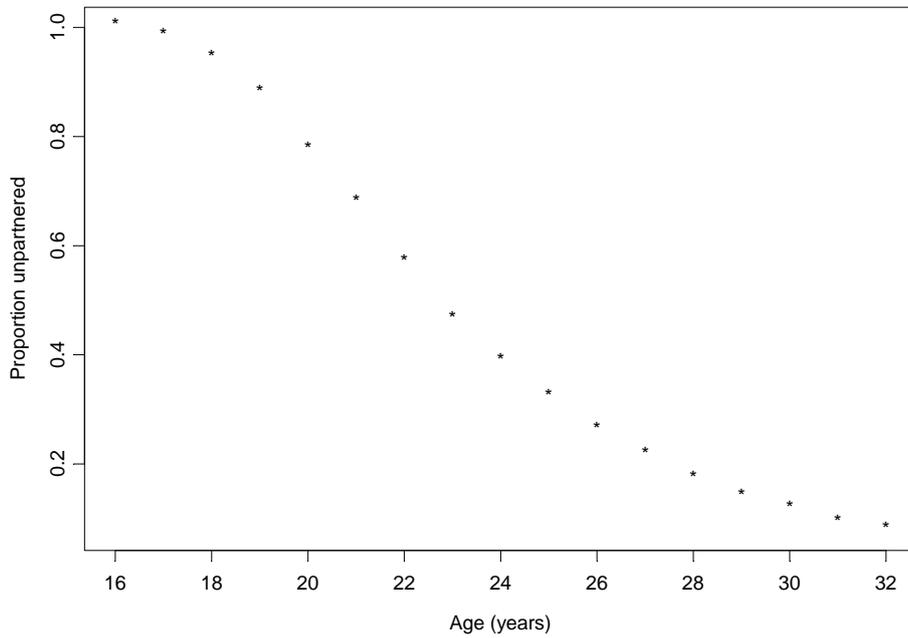


Figure 4. Examples of (a) proportional and (b) non-proportional hazards for two groups

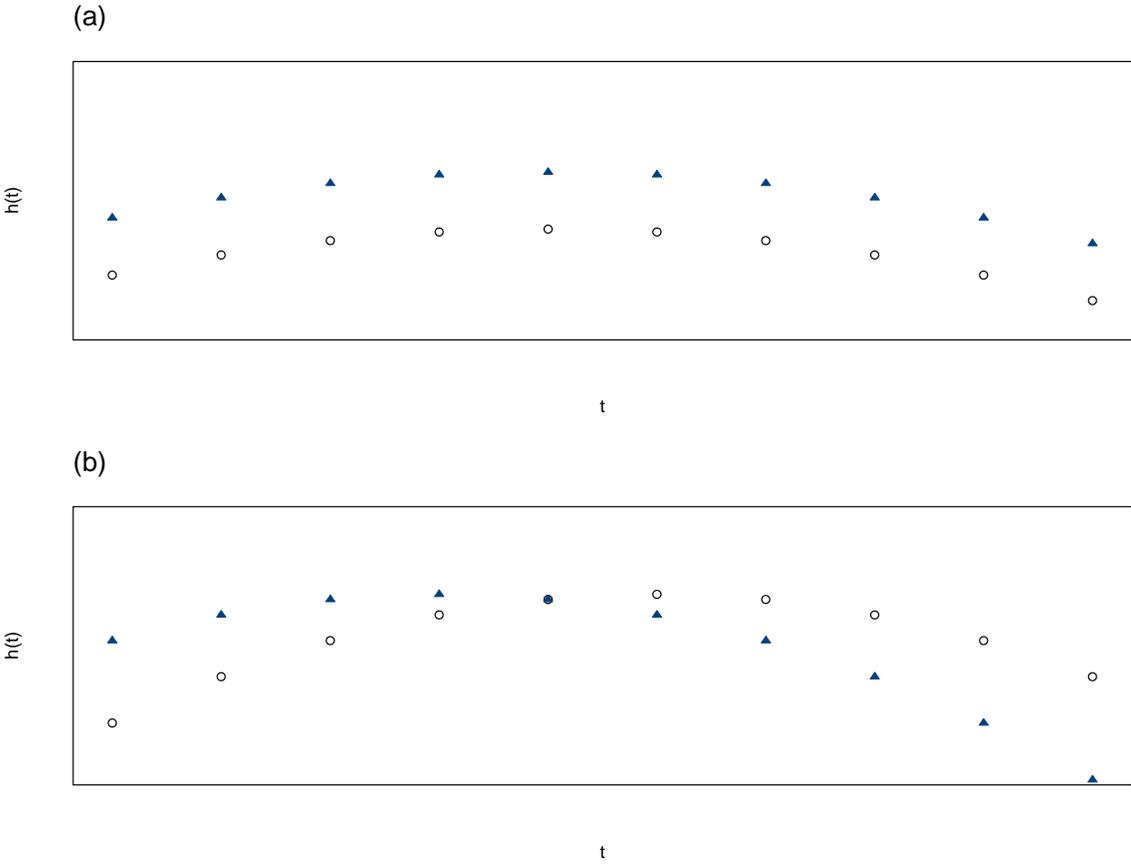


Figure 5. Example of restructuring data for a discrete time analysis

Person-based file

i	y_i	δ_i	FEMALE $_i$	AGELEFT $_i$
1	19	0	1	16
2	25	0	0	21



Person-period file

i	t	y_{it}	FEMALE $_i$	FULLTIME $_i(t)$
1	16	0	1	1
1	17	0	1	0
1	18	0	1	0
1	19	1	1	0
2	16	0	0	1
2	17	0	0	1
2	18	0	0	1
2	19	0	0	1
2	20	0	0	1
2	21	0	0	1
2	22	0	0	0
2	23	0	0	0
2	24	0	0	0
2	25	1	0	0