

Introduction to Discrete-Time Models

Fiona Steele
Centre for Multilevel Modelling
University of Bristol, UK

Discrete-time durations are the norm in social science research where event history data are most commonly collected prospectively and at irregular intervals in panel studies, or retrospectively in cross-sectional studies. In this context, it is rare that data on event timings are collected in a form that can be treated as continuous. Theoretically, events always occur at *points* in time, but in practice imprecise measurement means we know only the time *interval* in which it occurred (e.g. a day, a calendar year). Thus continuous-time processes are usually observed in discrete time. In some cases, the precise timing of an event within a time interval is unimportant and we can happily treat the process as truly discrete; for example, if one intended to analyse the length of time spent by pupils undergoing formal schooling, it is not crucial that time is measured to the nearest second because months or even years would suffice. In other cases, one would ideally like to work in continuous time but only discrete-time data are available. For such situations, a discrete-time model can be used as an approximation to a continuous-time model.

In this chapter we introduce the discrete-time approach to event history analysis, including descriptions of data preparation and alternative discrete-time models. Discrete-time models are applied in an analysis of the first job durations from the German Life History Survey, and results are compared with those from fitting a piecewise constant exponential model. Finally we discuss the analysis of aggregated data, arising from grouping of observed intervals, and the potential impact of aggregation on the estimated coefficients of time-varying covariates.

The aim of this chapter is to give an overview of the discrete-time approach, as a basis for the more complex models. More detailed introductions are given by Allison (1982) and Singer and Willett (2003).

1. Discrete-Time Measurement and Modelling

Although events in the process under study can theoretically occur at any point in time, durations are commonly measured in discrete time units. For example, employment histories are typically collected retrospectively and, because respondents are unlikely to recall the day they started or left a job, they would usually be asked for the month and year of any change in employment status. Such imprecisely measured durations are sometimes referred to as interval censored because, from the available data, we know only that an event occurred within an interval of time. We do not know *exactly* when the event occurred within that interval. We are therefore unable to estimate the continuous transition rate; all we can estimate is the probability that an event occurs within a given interval. This probability, which we shall call the discrete-time transition rate (or hazard), can be viewed as an approximation to the continuous-time transition rate.

One potential problem with using continuous-time models for interval censored data is that such models are formulated for truly continuous event processes from which at most one event can occur at any given point in time. Interval censored data can lead to tied event times, which necessitates the adaptation of the usual estimation procedures to account for ties (see Hosmer and Lemeshow (1999: Chapter 3) for a discussion). An alternative way to handle ties is to use a discrete-time model that explicitly recognises the grouped nature of the data. With a discrete-time approach there is no restriction on the number of events that occur within any given time interval. Furthermore, as for the piecewise constant exponential model, it is straightforward to adapt a discrete-time model to allow for non-proportional transition rates and time-varying covariates can be incorporated easily. For practical purposes, an attractive feature of discrete-time models is that they can be fitted as regression models for binary responses models following some initial data restructuring. This is a particular advantage for more advanced methods because it means we can make use of techniques and software developed for categorical response data.

The data structure required for a discrete-time analysis involves generating a sequence of binary responses for each event time, and is described in the next section. The increased size of the dataset can lead to long computational times for complex models with random effects, although a strategy for minimising the number of records is described in Section 5.

2. Data Structure for Discrete-Time Models

We assume that each individual can experience the event of interest at most once so that there is a single episode for each individual.

Suppose event times are realisations of a random variable T measured in intervals of time indexed by $t = 1, 2, \dots, K$ where K is the number of intervals in which events can occur. These intervals represent a series of consecutive periods in continuous time with boundaries or cut-points given by $\tau_1, \tau_2, \dots, \tau_K$, where $\tau_1 = 0$, interval t ($t = 1, 2, \dots, K - 1$) is given by $[\tau_t, \tau_{t+1})$, and interval K by $[\tau_K, \infty)$. In the case of retrospectively collected histories with event occurrence recorded to the nearest year, for example, intervals might correspond to years, i.e. $[0,1), [1,2)$, etc. More generally, and especially when data are collected prospectively in irregularly-spaced waves of a panel study, intervals may be of different widths.

Denote by t_i the number of intervals for which individual i is observed. In the person-period file required for a discrete-time analysis the duration for individual i is expanded to t_i records. For each record, $t = 1, \dots, t_i$, we define a binary indicator y_{it} such that

$$y_{it} = \begin{cases} 1 & \text{if individual } i \text{ has an event in interval } t \\ 0 & \text{otherwise.} \end{cases}$$

All individuals, regardless of whether or not their duration is censored, will have $y_{it} = 0$ for intervals $t = 1, \dots, t_i - 1$. The response for the last observed interval, y_{it_i} , will be 1 for individuals who experience an event during that interval and 0 for those who are censored. For example, the first individual in Table 3.2 with a censored first job duration of 35 years would contribute 36 records to the person-period file, corresponding to yearly intervals $[0,1), [1,2), \dots, [35,36)$, with $y_{it} = 0$ for each interval t . The second individual, with an uncensored duration of 3 years, would have four records with $(y_{1i}, y_{2i}, y_{3i}, y_{4i}) = (0, 0, 0, 1)$. This data structure is in fact the same as that described for the piecewise constant exponential

model with ungrouped intervals. Intervals can also be grouped in a discrete-time analysis (see Section 5). Although their input data have the same form, however, the piecewise constant exponential model is in terms of the continuous-time transition rate while the discrete-time models considered in this chapter are models for the probability of event occurrence.

3. Discrete-Time Approximation to the Continuous-time Proportional Transition Rate Model

3.1 Relationship between the Continuous-time Transition Rate and the Discrete-time Probability of Event Occurrence

Suppose that all time intervals are of the same width. So we can say that any time interval t is of width s and ranges from a lower limit of t_l to an upper limit (but not including) $t_l + s$. As usual we denote the interval by $[t_l, t_l + s)$.

The probability that an event occurs during interval t , given survival to the start of interval t , is $\Pr(y_t = 1 | y_{t-1} = y_{t-2} = \dots = y_1 = 0)$. We can express this more succinctly as $\Pr(y_t = 1)$ because y_t is only observed if the individual survives to the end of the preceding interval $t-1$ ¹.

Now assume that the transition rate (or hazard of an event occurring) is λ , namely, constant within each interval and the same for all t , i.e. $\lambda_t = \lambda$. A constant rate implies that event times T follow an exponential distribution. If p_t is the probability of an event occurring within (or during) interval t , then it can be calculated as 1 minus the probability of survival beyond time $t_l + s$ (given survival to t_l); that is,

$$p_t = 1 - \Pr(T > t_l + s | T > t_l) = 1 - \frac{\Pr(T > t_l + s)}{\Pr(T > t_l)} \quad (1).$$

¹ To simplify the discussion, the subscript i indexing the individual is omitted.

If event times T follow an exponential distribution with rate λ it can be shown (e.g. Blossfeld et al., 2007, Chapter 4) that $\Pr(T > a) = \exp(-\lambda a)$. Substituting this into (1) gives an expression that enables us to link p_t and the continuous-time transition rate λ :

$$p_t = 1 - \frac{\exp[-\lambda(t_t + s)]}{\exp(-\lambda t_t)} = 1 - \exp(-\lambda s). \quad (2)$$

Thus the probability of event occurrence is also constant and depends on the length of the interval, which is fixed and so we can write $p_t = p$.

3.2. A Discrete-time Approximation to a Continuous-time Proportional Transition Rate Model with Constant Rate

In this section we derive a model for the probability of event occurrence that provides a discrete-time approximation to the proportional transition rate model, using the relationship between the probability and rate given in (2).

A continuous-time proportional transition rate model with a constant rate and a single explanatory variable x can be written

$$\log \lambda_i = \beta_0 + \beta_1 x_i. \quad (3)$$

Taking exponents of both sides of (3) gives $\lambda_i = \exp(\beta_0 + \beta_1 x_i)$ and substitution in (2) gives

$$p_i = 1 - \exp(-s \exp(\beta_0 + \beta_1 x_i)). \quad (4)$$

If we now subtract each side of (4) from 1 and take natural logarithms we get

$$\log(1 - p_i) = -s \exp(\beta_0 + \beta_1 x_i). \quad (5).$$

Finally, multiplying each side by -1 and taking logarithms again gives

$$\log(-\log(1 - p_i)) = \log(s) + \beta_0 + \beta_1 x_i. \quad (6).$$

The transformation of p on the left hand side of (6) is called the *complementary log-log* function, and the model specified by (6) is a complementary log-log model. The response variable is the binary indicator of event occurrence y_i . In the case that all intervals are of the same width, $\log(s)$ is a constant which can be omitted from the model.

Because (6) is derived directly from a proportional transition rate model, the exponentiated coefficient e^β is an approximation to the relative risk in a model for the rate. Therefore the parameter estimates obtained from a complementary log-log model are directly comparable to the coefficients from the corresponding Cox model or PCE model.

The complementary log-log model for p is a discrete-time approximation to the continuous-time piecewise constant exponential (PCE) model for the transition rate λ , and p is an approximation to λ . As λ becomes smaller, the better this approximation will be. The approximation of λ by p will also improve as the width of the time interval, s , becomes smaller. Throughout this book, we shall adopt the convention of much of the discrete-time event history analysis literature and refer to the probability of event occurrence as the discrete-time transition rate or discrete-time hazard, but the relationship with the conventional continuous-time transition rate is now clear.

3.3. Generalisations of the Complementary Log-log Model

The complementary log-log model can be extended in the same way as the piecewise constant exponential model to allow for duration dependency in the transition rate, unequal time intervals, time-varying covariates and non-proportional rates.

Duration dependency

The model in (6) can be extended to allow the transition rate to vary across time intervals:

$$\log(-\log(1 - p_{it})) = \log(s) + \alpha_t + \beta_0 + \beta_1 x_i \quad (7)$$

where α_t is some function of t that is added to the linear predictor, just as for the piecewise constant exponential model. The most flexible model would treat t as a categorical variable, with K dummy variables, leading to a step function. Other possible specifications for duration dependency include polynomial functions.

Unequal time intervals

We have thus far assumed that time intervals are of constant width s . More generally the width of interval t may be s_t , in which case $\log(s_t)$ is included as an offset term. The offset can be fitted by treating $\log(s_t)$ as an explanatory variable with coefficient constrained to equal 1; most software packages allow offsets to be defined in models for categorical and count data.

Time-varying covariates

Because the analysis file for a complementary log-log analysis is in the form of one record per time interval, it is straightforward to allow the values of covariates to vary across time intervals. We can therefore simply replace x_i by x_{it} .

Non-proportional covariate effects

Model (7) is a discrete-time approximation to a continuous-time proportional transition rate model. We can allow the effects of the covariates on the probability of event occurrence to depend on duration by replacing β by β_t . As for the piecewise constant exponential model, we can allow for non-proportional effects of a covariate x_i by including interactions between x_i and each of the duration variables that constitute α_t .

3.4. Example

Table 1 shows results from fitting piecewise constant exponential (PCE) and complementary log-log models to first job durations among GLHS respondents. The models include quadratic duration effects and the following covariates:

- A dummy for sex (*female*);
- Highest *educational attainment* before entry into the labour market (in school year equivalents);
- A measure of the *prestige of the current job* (ranging from 18 to 70 with a mean of 37 and standard deviation of 10);
- *Birth cohort* (grouped into 3 categories: 1929-31, 1939-41, and 1949-51).

For each model, time intervals are of equal width but results are compared for monthly and yearly intervals.

Table 1. Piecewise constant exponential and complementary log-log models fitted to first job durations in monthly and yearly intervals

| | Duration in monthly intervals | | | | Duration in yearly intervals | | | |
|-----------------------|-------------------------------|-------|----------|-------|------------------------------|-------|----------|-------|
| | PCE | | clog-log | | PCE | | clog-log | |
| | β | SE | β | SE | β | SE | β | SE |
| Constant | -4.870 | 0.462 | -4.869 | 0.461 | -2.323 | 0.464 | -2.306 | 0.465 |
| t | -0.002 | 0.003 | -0.002 | 0.003 | -0.031 | 0.039 | -0.035 | 0.039 |
| t ² | -0.000 | 0.000 | -0.000 | 0.000 | -0.001 | 0.002 | -0.001 | 0.002 |
| Female | 0.302 | 0.156 | 0.304 | 0.156 | 0.280 | 0.155 | 0.303 | 0.156 |
| Education (years) | 0.073 | 0.033 | 0.073 | 0.033 | 0.065 | 0.033 | 0.075 | 0.033 |
| Prestige score of job | -0.010 | 0.009 | -0.010 | 0.009 | -0.010 | 0.009 | -0.011 | 0.009 |
| Birth cohort | | | | | | | | |
| 1929-31 (ref) | 0 | - | 0 | - | 0 | - | 0 | - |
| 1939-41 | 0.474 | 0.192 | 0.478 | 0.192 | 0.438 | 0.192 | 0.488 | 0.192 |
| 1949-51 | 0.355 | 0.184 | 0.357 | 0.184 | 0.331 | 0.184 | 0.362 | 0.184 |

When data in the form of monthly records are analysed, the estimated coefficients and standard errors are extremely close. This is a direct result of the fact that the rate of leaving the first job in any given month is very small. There are two important implications of this small transition rate. First, the rate λ_t is closely approximated by p_t . Second, it can be shown that $-\log(1 - p_t) \approx p_t$ for small p_t . Taken together, these results imply that $\log(\lambda_t)$ will be close to $\log(-\log(1 - p_t))$ when the rate (and probability of event occurrence) is small. Therefore, the estimated coefficients of the PCE model for $\log(\lambda_t)$ will be close to the estimated coefficients of the clog-log model for $\log(-\log(1 - p_t))$.

As expected, the estimates from the PCE and clog-log models move further apart when the width of each time interval is increased to one year. Nevertheless, the substantive conclusions about the direction and statistical significance of the effects of each variable are the same for each model.

4. Logit and Probit Discrete-time Models

4.1. A general discrete-time model

We have referred to (7) as a discrete-time model because it is a model for the probability of event occurrence p_t rather than the transition rate λ_t . Strictly speaking, however, the complementary log-log model is a discrete-time *approximation* to a continuous-time proportional transition rate model. The event times being modelled are still assumed to be realisations of a continuous-time process: the observed data are treated as imprecise measurements of events that occur in continuous time. In theory, for example, a job could start on 1 January and end on 26 January in the same year but, if respondents are asked for dates in months and years, we would record a duration of 0 months or an event in the first monthly interval. Thus the data we observe are in discrete time units, but the underlying process is continuous.

While in theory all events in social processes occur in continuous time, there are situations where the exact timing of event occurrence is unimportant. For example, in studies of the

length of time spent in full-time education, students will vary in the precise time at which they leave school or graduate from university. Even if we were able to collect information on the day of leaving education, however, individual differences within a calendar year would usually be of little interest. In such cases, it is more natural to use a model that assumes an underlying discrete-time process. As before we create a binary response y_{it} for each time interval t during which an individual i is at risk of an event, and define the probability of event occurrence as $p_{it} = \Pr(y_{it} = 1)$. Suppose we begin with the assumption that all time intervals are of equal width. We can then regard the observation of an individual over interval t as a ‘trial’ where the probability of ‘success’ (an event occurring) is p_{it} and the probability of ‘failure’ (no event) is $1 - p_{it}$. Viewing the event process in this way, y_{it} is said to follow a Bernoulli distribution, a discrete distribution with a single parameter p_{it} .

A general model for the relationship between p_{it} and an explanatory variable x_i is

$$g(p_{it}) = \alpha_i + \beta_0 + \beta_1 x_i \quad (8)$$

where $g(\cdot)$, known as the *link function*, is a transformation of p_{it} which ensures that predicted values of p_{it} lie between 0 and 1.

4.2. Choice of link function

The complementary log-log function $g(p_{it}) = \log(-\log(1 - p_{it}))$ is one possible choice for the link function which, as we have seen, leads to an approximation of a continuous-time proportional transition rate model. More commonly applied alternatives, however, are the logit and probit functions:

$$\text{logit}(p_{it}) = \log\left(\frac{p_{it}}{1 - p_{it}}\right) \text{ and } \text{probit}(p_{it}) = \Phi(p_{it}),$$

respectively, where $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution (i.e. with a mean of 0 and variance of 1).

Figure 1 shows the logit, probit and complementary log-log transformations of a probability on the range 0 to 1. The following properties can be seen:

- The logit and probit transformations are symmetrical, i.e. $g(p) = -g(1 - p)$. A direct, convenient result of this symmetry is that switching the coding of y so that 0 indicates an event (and 1 no event) will lead to a reversal in the signs of the regression coefficients, but no change in their absolute values (with the exception of the intercept).
- $\text{Logit}(0.5) = \text{probit}(0.5)$, but the two functions move further apart as p gets closer to zero and, by symmetry, closer to 1. In contrast, $\text{logit}(p)$ and $\text{cloglog}(p)$ are almost indistinguishable for small p . Because the probability of event occurrence in an interval of time is usually small, we would therefore expect the predicted probabilities from logit and complementary log-log versions of the same model to be very similar, and to become closer as the width of the time intervals become narrower. Predictions from a probit event history model may be rather different to those from the equivalent logit and clog-log models.

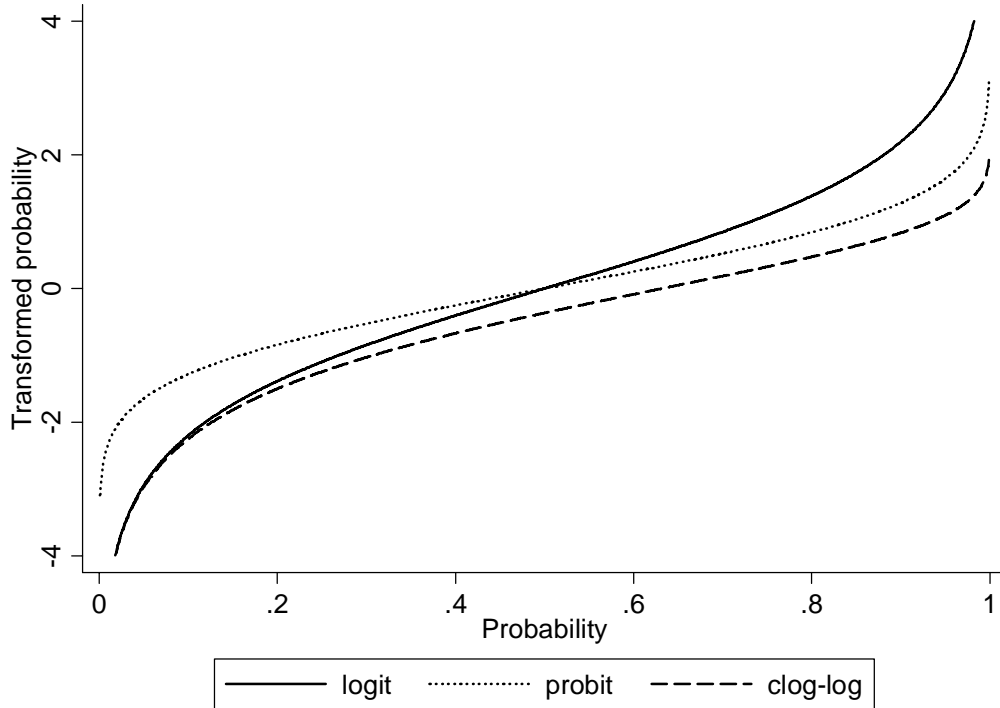


Figure 1. The logistic, probit and complementary log-log transformations of a probability

Table 2 shows predicted probabilities of leaving the first job before the end of years 1, 3 and 5 for the three link functions. Each model includes a quadratic duration effect and the set of covariates shown in Table 1. Separate models were fitted to datasets with monthly and yearly records. The estimates in Table 2 were obtained as follows:

- (i) For each model a predicted probability of leaving the first job, \hat{p}_{ti} , was calculated for each observation in the dataset, i.e. for each time interval (a month or year) of each individual.
- (ii) The average probability of leaving in interval t was then calculated for each gender, denoted by \hat{p}_{tg} .
- (iii) From \hat{p}_{tg} an estimate of leaving during or after the end of interval t (the survival probability) was calculated as $\hat{S}_{tg} = (1 - \hat{p}_{t-1,g}) \hat{S}_{t-1,g}$ for $t > 1$, $\hat{S}_{1g} = 1$.

(iv) Finally, an estimate of leaving before the start of interval t was computed as $1 - \hat{S}_{tg}$.

Table 2. Predicted probability of leaving first job within t years by gender from clog-log, logit and probit models, duration in yearly and monthly intervals

| t (years) | Gender | Duration in monthly intervals | | | Duration in yearly intervals | | |
|-------------|--------|-------------------------------|-------|--------|------------------------------|-------|--------|
| | | Clog-log | Logit | Probit | Clog-log | Logit | Probit |
| 1 | Male | 0.173 | 0.173 | 0.175 | 0.180 | 0.181 | 0.182 |
| 1 | Female | 0.224 | 0.224 | 0.228 | 0.232 | 0.234 | 0.237 |
| 3 | Male | 0.419 | 0.419 | 0.419 | 0.430 | 0.431 | 0.431 |
| 3 | Female | 0.516 | 0.516 | 0.521 | 0.526 | 0.529 | 0.534 |
| 5 | Male | 0.575 | 0.575 | 0.575 | 0.584 | 0.584 | 0.583 |
| 5 | Female | 0.688 | 0.689 | 0.693 | 0.694 | 0.698 | 0.702 |

From Table 2, we see that the predicted probabilities of leaving the first job are almost identical for the complementary log-log and logit models, whether the data are in monthly or yearly intervals. The probabilities for females are slightly larger for the probit, leading to a larger gender effect, but the differences are small.

In many other practical applications the link function will have little impact on predictions, and the choice of link is often therefore based on factors such as interpretational convenience or disciplinary preferences. For example, probit models are widely used in economics while logit models are the most popular choice in epidemiology. In the remainder of this book, the focus will be on logit models for several reasons. First and foremost, most readers will be already be familiar with logistic regression for binary response data and the logit form is the most commonly used. Second, as will be illustrated in the next section, exponentiated logit coefficients have an odds ratio interpretation. Another reason for favouring the logit model is that its extension to multilevel data is well established, and it extends naturally to handle competing risks (a multinomial logit model).

4.3. Application of the logit model to first job durations

Results from a logit model fitted to the first job data with one record per month of exposure are given in Table 3. The ‘logit’ predicted probabilities in Table 2 were calculated from the estimates for β . The estimated coefficients are very close to those estimated from the PCE and complementary log-log models fitted to the same monthly data (Table 1). As noted in the previous section, this is because the probability of event occurrence p_t is small which means that p_t is a good approximation to the transition rate λ_t modelled in the PCE model, and that the $\text{logit}(p_t)$, $\text{log}(-1 - p_t)$ and $\text{log}(\lambda_t)$ transformations will yield similar values.

Table 3. Logit model fitted to first job durations, monthly intervals

| | β | SE | $\exp(\beta)$ | 95% CI for $\exp(\beta)$ |
|-----------------------|---------|-------|---------------|--------------------------|
| Constant | -4.865 | 0.466 | - | - |
| t | -0.002 | 0.003 | 0.998 | (0.992, 1.004) |
| t ² | -0.000 | 0.000 | 1.000 | (1.000, 1.000) |
| Female | 0.307 | 0.158 | 1.360 | (0.998, 1.852) |
| Education (years) | 0.074 | 0.034 | 1.077 | (1.008, 1.150) |
| Prestige score of job | -0.010 | 0.009 | 0.990 | (0.973, 1.008) |
| Birth cohort | | | | |
| 1929-31 (ref) | 0 | - | 1.000 | - |
| 1939-41 | 0.482 | 0.194 | 1.619 | (1.108, 2.366) |
| 1949-51 | 0.360 | 0.186 | 1.433 | (0.997, 2.062) |

The exponentiated coefficients $\exp(\beta)$ are interpreted as odds ratios for logit models and relative risks for PCE and complementary log-log models, although here the small p_t means that the two quantities are almost the same. For a continuous x , e.g. education or job prestige, the odds ratio compares the odds of an event for values of x spaced one unit apart (holding all other explanatory variables constant). We find, for example, that a one year increase in education increases the odds of leaving the first job by a factor of 1.077 or by 7.7%. In the case of a binary x , the odds ratio compares the odds of an event for $x = 1$

relative to $x = 0$. For example, the odds of leaving are 1.36 times higher for women than for men.

Confidence intervals for $\exp(\beta)$ are obtained by first calculating lower and upper confidence limits for β , and then taking the exponential of these limits. For instance, a 95% confidence interval for the population gender effect on the *log*-odds is

$$0.307 \pm (1.96 \times 0.158) = (-0.0027, 0.6167)$$

and a 95% confidence interval for the population female-to-male *odds ratio* is

$(e^{-0.0027}, e^{0.6167}) = (0.997, 1.853)$. (The slight difference between these values and those given in Table 3 are due to rounding.)

If a variable x has no effect on the odds of an event, then its coefficient β will equal zero and $\exp(\beta)$ will equal 1. Thus, a 95% confidence interval for β that includes the value 1 implies that, at the 5% significance level, there is no relationship between the odds of an event and x . From Table 3, for example, we would conclude that gender and job prestige do not have significant effects on first job duration. There is also no significant difference in first job durations of the 1929-31 and 1949-51 birth cohorts but, in any given month, the odds of leaving the first job are higher for the 1939-41 cohort than for those born in 1929-31.

Confidence intervals for β (or approximate t-ratios based on $\hat{\beta} / SE(\hat{\beta})$) can be used to test hypotheses about single parameters, while a likelihood ratio test can be used to test whether a set of coefficients are simultaneously equal to zero. Suppose, for example, that we wish to test for duration effects. We would compare the model in Table 3 (M_1) with the same model after excluding both t and t^2 (M_2). The second model is a constrained version of the first with the coefficients of t and t^2 set to zero. The -2 log-likelihood values for these models are respectively 1908.8 and 1923.7. The test statistic is the difference between these values, 14.9, which on comparison with a chi-squared distribution on 2 degrees of freedom yields a p-value of 0.0006. We therefore conclude that, on average, the probability of leaving one's first job depends on the length of time in the job. Furthermore, the fact that the estimated

coefficients of t and t^2 are both negative implies that the probability of leaving decreases with duration.

5. Aggregating Time Intervals

The expansion of each duration into a person-period format can lead to an extremely large dataset, particularly when time intervals are short and the observation period long. While the rapid increase in computing power means that this is not an important concern for the simple models considered here, it becomes more of an issue for random effects models. Fortunately, it is often possible to aggregate intervals with minimal loss of information.

So far in this chapter, analyses of first job durations have been carried out on monthly and yearly time intervals. The duration in years was calculated simply by dividing the duration in months by 12. However, this is a crude method of aggregation that ignores the fact that individuals will vary in the time at which they experience events during a year. In the continuous-time piecewise constant exponential model, different exposure times are handled by including as an offset term n_{it} , the length of exposure of individual i in time interval t . The same approach can be adopted for complementary log-log models. We have already seen in Section 3.3 how unequal time intervals can be accommodated by treating the width of time interval t , s_t , as an offset. More generally, s_t could be replaced by n_{it} , the exposure time of individual i in interval t . We then include $\log(n_{it})$ as an offset term.

Alternatively, for a discrete-time process, we can view n_{it} as a set of Bernoulli trials taking place for individual i in interval t . For example, a person who has an event or is censored during the sixth month of a yearly interval is observed for six trials. At each trial (month) we record whether or not an event occurs to the individual, where the total number of events during interval t is denoted by r_{it} . In fact r_{it} can only equal 0 or 1 because the events considered thus far are not repeatable, so that $r_{it} = y_{it}$. The rate of event occurrence in interval t is therefore r_{it} / n_{it} which follows a *binomial* distribution. Compared to a Bernoulli distribution for binary data, a binomial distribution has n_{it} as an additional parameter, commonly referred to as the *denominator*. Most statistics software packages with routines

for logit, probit and clog-log models can also handle binomial data. (In Stata, for example, the `blogit` command can be used to fit a binomial logit model.)

If the probability of event occurrence is constant within each interval t and the values of any time-varying covariates are constant across t , aggregation will not lead to any loss of information provided n_{it} is incorporated in the model. In practice, the first of these assumptions will usually be reasonable and small departures from a piecewise constant transition rate will have little impact on the estimated coefficients of the covariates. The second assumption may be more difficult to justify and can lead to more substantial differences between results for the same model fitted to aggregated and raw data. Many event history analyses are focused on the relationship between a change in the value of a time-varying covariate and the timing of an event. One example where it is important to have precise information on the relative timing of changes in covariate values and events is an investigation of how the marriage rate among cohabiting couples changes during pregnancy, following a birth, and with the number and age of children. Another example is a study of the nature of the relationship between the time taken to achieve promotion and the presence and age of children. When intervals are aggregated, the researcher must choose an appropriate point in the grouped interval at which time-varying variables are defined. Whatever choice is made, some loss of information is inevitable. For example, if pregnancy status is measured at the start of a group interval (e.g. a year), pregnancies that start and finish within the interval will be missed.

We examine the impact of aggregating time intervals in an analysis of the timing of first marriage among a sample of 3876 British cohabiting women. The data are from the British Cohort Study which has as its subjects all individuals born in Britain during a single week of 1970 (Bynner et al. 1997). In an interview when the respondents were aged 30, they were asked to recall the start and end dates of cohabitations and marriages and the births of children. All dates were recorded to the nearest month. In the illustrative analysis presented here we consider four covariates: the woman's age at the start of the cohabitation, the number of years of post-compulsory education, an indicator of whether or not she had experienced the breakdown of her parents' relationship before age 16, and an indicator of her current pregnancy status. Education and pregnancy status are treated as time-varying covariates. A quadratic function in t , the current duration of the cohabitation, is also included to allow for

the possibility that the probability of marriage may depend on the length of time that a woman has lived with her partner. Further analyses of the formation and outcomes of *all* cohabiting and marital partnerships between ages 16 and 30, and with a richer set of covariates, can be found in Steele et al. (2006).

The same logit model is fitted to data with one record per month, six-month and year of cohabitation until marriage or, for censored cases, age 30. In each case the response variable is a binary indicator of event occurrence (marriage) but, for data aggregated to six- and 12-month intervals, we account for the number of months of exposure within each interval. Thus, for example, a woman who marries during her 14th month of cohabitation would contribute 14 one-month intervals, three six-month intervals, and two yearly intervals (see Table 4).

Table 4. Data structures for a woman who marries in the 14th month of cohabitation

| t | Indicator of marriage (y_t) | Exposure in months (n_t) |
|----------------------------|---------------------------------|------------------------------|
| Data in 6-month intervals | | |
| 1 | 0 | 6 |
| 2 | 0 | 6 |
| 3 | 1 | 2 |
| Data in 12-month intervals | | |
| 1 | 0 | 12 |
| 2 | 1 | 2 |

Table 5 shows the estimated coefficients and standard errors from analyses of data at different levels of aggregation. The coefficient of t when durations are in six-month intervals is approximately six times the value for one-month durations ($6 \times 0.014 = 0.084$). Similarly, the coefficient of t when the data are in yearly intervals is roughly 12 times the one-month estimate. As expected there is little effect of aggregation on the effects of the time-invariant variables, age and family disruption. There is also little impact on the coefficients for education; although there is some within-individual variation in this variable, its values are fixed over time for most women in the sample because they typically complete education before cohabiting. In contrast, there are substantial differences in the estimated coefficients

and statistical significance of current pregnancy status. The expected positive association between pregnancy and the odds of marriage is found only for the monthly data; when the data are aggregated to yearly intervals, the relationship becomes negative and non-significant.

Table 5. Logit models fitted to first cohabitation to marriage transitions, data in 1, 6 and 12 month intervals

| Variable | 1-month intervals | | 6-month intervals | | 12-month intervals | |
|-------------------------------|-------------------|-------|-------------------|-------|--------------------|-------|
| | β | SE | β | SE | β | SE |
| Constant | -5.564 | 0.210 | -5.531 | 0.213 | -5.488 | 0.217 |
| Cohabitation duration (t) | 0.014 | 0.003 | 0.088 | 0.018 | 0.167 | 0.037 |
| t^2 | -0.000 | 0.000 | -0.005 | 0.001 | -0.019 | 0.004 |
| Age at cohabitation (years) | 0.047 | 0.009 | 0.045 | 0.009 | 0.044 | 0.009 |
| Post-16 education (years) | | | | | | |
| 0 (reference) | 0 | - | 0 | - | 0 | - |
| 1 | 0.089 | 0.067 | 0.081 | 0.067 | 0.069 | 0.067 |
| 2 | 0.105 | 0.068 | 0.102 | 0.068 | 0.096 | 0.068 |
| 3-5 | -0.038 | 0.072 | -0.064 | 0.073 | -0.079 | 0.073 |
| ≥ 6 | 0.062 | 0.089 | 0.050 | 0.089 | 0.030 | 0.089 |
| Family disruption before 16 | -0.269 | 0.062 | -0.267 | 0.062 | -0.267 | 0.062 |
| Currently pregnant | 0.476 | 0.077 | 0.129 | 0.073 | -0.077 | 0.069 |

As noted earlier, the analyst must decide to which point in the grouped interval a time-varying covariate should refer. The pregnancy status variable in Table 5 is an indicator of whether a woman is pregnant in *any* month during the grouped interval. This measure has the advantage of including information on all pregnancies. An alternative approach is to consider pregnancy status in the *first* month of each six-month and yearly interval, although clearly this indicator would miss pregnancies starting during an interval. The coefficients for these definitions are compared in Table 6. Perhaps surprisingly, we find that the variables based on pregnancy status in the first month have coefficients that are closet to the one-month estimate of 0.476. However, both measures lead to a loss of information on the timing of conception relative to the timing of marriage. Suppose, for example, that two women become pregnant in the second month of a yearly interval: the first marries during the pregnancy in the sixth month of the same interval, while the second marries following the

birth in the twelfth month. The indicator based on status at the first month will ignore both pregnancies, leading to an underestimate of the relationship between pregnancy and marriage. The alternative indicator will pick up both pregnancies, but will not recognise the fact that one of the women married before the birth while the other married after. It is particularly important to account for the timing of marriage relative to the birth because previous studies (e.g. Blossfeld et al., 1999; Steele et al., 2006) have found that the odds of cohabitation increase sharply during pregnancy, but then drop to below pre-conception levels after the birth. This pattern in the relationship between fertility and marriage is most often attributed to selection whereby couples with a favourable attitude towards marriage will tend to marry before the birth, leaving behind couples who are less inclined to marry. By using an indicator based on pregnancy status at any month during an interval, and therefore not distinguishing pre- and post-marital cohabiting births, we are grouping together women with quite different odds of marriage.

Table 6. Estimated pregnancy status effects on the first cohabitation to marriage transition

| Interval width | Definition of pregnancy status | β | SE |
|----------------|---------------------------------------|---------|-------|
| 1 month | Current status | 0.476 | 0.077 |
| 6 month | 1 st month in interval | 0.245 | 0.084 |
| 6 month | Pregnant in any month during interval | 0.129 | 0.073 |
| 12 month | 1 st month in interval | 0.101 | 0.090 |
| 12 month | Pregnant in any month during interval | -0.077 | 0.069 |

To summarise, aggregation can be a useful way of reducing the size of a discrete-time dataset, especially when fitting more complex random effects models. Furthermore, in applications where all covariates of interest are time-invariant and the transition rate can be assumed constant within grouped intervals, there will be little impact on the regression coefficients. Where there are time-varying covariates, however, aggregation should be used with caution; it is important to assess the robustness of the results to the width of time intervals.

References

- Allison P.D. (1982) Discrete-time methods for the analysis of event histories. In: *Sociological Methodology* (ed. Leinhardt S), pp. 61-98. Jossey-Bass, San Francisco
- Blossfeld H.-P., Golsch K. and Rohwer G. (2007) *Event History Analysis with Stata*. Lawrence Erlbaum Associates, Mahwah, New Jersey.
- Blossfeld H.-P., Klijzing E., Pohl K. and Rohwer G. (1999) Why do cohabiting couples marry? An example of a causal event history approach to interdependent systems. *Quality and Quantity*, **33**, 229-242.
- Hosmer D.W. and Lemeshow S. (1999) *Applied Survival Analysis: Regression Modeling of Time to Event Data*. Wiley, New York.
- Singer J.D. and Willett J.B. (2003) *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press, New York.
- Steele F., Kallis C. and Joshi H. (2006) The formation and outcomes of cohabiting and marital partnerships in early adulthood: the role of previous partnership experience. *Journal of the Royal Statistical Society, Series A*, **169**, 757-780.