Gayle, Vernon, and Paul Lambert. "The analysis of duration data." *What is Quantitative Longitudinal Data Analysis?*. London: Bloomsbury Academic, 2018. 39–56. *Bloomsbury Collections*. Web. 26 Apr. 2023. <<http://dx.doi.org/10.5040/9781472545268.0011>>.

Downloaded from Bloomsbury Collections, www.bloomsburycollections.com, 26 April 2023, 11:48 UTC.

Copyright © Vernon Gayle and Paul Lambert 2018. You may share this work for non-commercial purposes only, provided you give attribution to the copyright holder and the publisher, and provide a link to the Creative Commons licence.

4 The analysis of duration data

Introduction

Duration analyses are generally employed to analyse how long something remains in a particular state, for example how long do graduates remain in their first job after leaving university. The central idea in duration analyses is that there is a fundamental change from one state to another at a specific point in time. The change in state is marked by an event. The term 'event' is not usually used to describe a gradual change and the event must consist of a relatively sharp disjunction between what precedes and what follows (Allison, 1984).

Techniques suitable for modelling duration data are routinely known as duration models, survival models, Cox regression, Cox models, failure-time analysis, hazard models and event history analysis. These terms are often used interchangeably. The different names partly reflect the development of methods for the analysis of duration data in different academic disciplines.

The foundational paper on duration models is Cox (1972). Models for duration data allow the data analyst to assess the relative influence of a number of explanatory factors upon how long it takes for an event to occur. The utility of duration models is evident when we consider that the models have been used to study the lifetimes of machine components in engineering, the duration of unemployment in economics, the time taken to complete cognitive tasks in psychology, the lengths of tracks on a photographic plate in particle physics and survival times of patients in clinical trials (see Cox and Oakes, 1984).

In the social sciences Heckman and Borjas (1980) used duration modelling approaches to study unemployment, Blossfeld and Hakim (1997) studied female part-time employment and Mulder and Smits (1999) investigated first-time home ownership. In the area of marriage and cohabitation, Lillard, Brien and Waite (1995) studied premarital cohabitation and

subsequent marital dissolution, Kiernan and Mueller (1998) undertook an analysis of divorce using the BHPS and the NCDS and Boyle et al. (2008) examined union dissolution using the Austrian Family and Fertility Survey (FFS). Using the BHPS Chan and Halpin (2002) examined the influence of gender role attitudes and the domestic division of labour on divorce, and Pevalin and Ermisch (2004) investigated mental health, union dissolution and re-partnering.

Measuring durations

There are three requirements for correctly determining duration. First, a starting time must be unambiguously defined. Second, time must have a defined unit of measurement. Third, the event must be clearly defined (Cox and Oakes, 1984).

We begin with a simple example relating to breastfeeding. The majority of both academic and non-academic writing on breastfeeding is profoundly in favour of the practice (Schmied and Lupton, 2001). The definition of exclusive breastfeeding is that the infant receives only breastmilk, and no other liquids or solids (except medicines or vitamins and mineral supplements). Imagine that a set of data on exclusive breastfeeding were being analysed three years after the babies were born. It is fair to assume that all of the babies would have finished exclusively breastfeeding. Indeed results from the 2010 UK Infant Feeding Survey¹ reported that across the UK the level of exclusive breastfeeding had decreased to one per cent six months after birth.²

In this example birth provides an unambiguous starting point, and the duration of exclusive breastfeeding would probably be measured in days or weeks. One standard measure of the outcome or event in breastfeeding research is the first time that formula, other liquids or solids are given to a baby. The duration (the number of days or weeks that the baby was exclusively breast fed) is a continuous measure and it would be possible to estimate a standard linear regression model. One option might be to

¹ See <http://www.hscic.gov.uk/catalogue/PUB08694/ifs-uk-2010-chap2-inc-prev-dur.pdf> (accessed 7 June 2016).

² Readers who are fans of the hit UK television comedy 'Little Britain' might, however, remember the character Harvey Pincher, who was breast fed for much longer than most children.

model the log of the durations as a continuous outcome. Using the log of the outcome variable is sometimes advised when the outcome variable has a pronounced positive skew, which is common for measures of duration. This model is sometimes referred to as the accelerated life model (Goldstein, 1999).

In analyses where the social process is complete, by which we mean that all of the sample have achieved the outcome, the accelerated life model is appropriate. Mortality analyses of veterans who fought in the First World War is one such example. Similarly we envisage that at the end of this century researchers studying mortality will have access to 'birth to death records' for the UK 1946 birth cohort. In genuine empirical social science studies it is common to have individuals for whom the event does not take place within the window of observation, and to address this methodological issue, more complex techniques of analysis are required.

Figure 4 is a diagram based on data from a hypothetical study of unemployment. The study began at month 0 and concluded 18 months later. During the 18-month observation window Person A exited unemployment (by gaining a full-time job) 3 months into the study, Person B did the same 15 months into the study and Person C finished their unemployment spell 18 months into the study. The measure of the unemployment duration for these three participants in the study is unproblematic. Person D did not exit unemployment until after the study ended, and a researcher would not have access to the precise time at which this happened. Person E never exited unemployment. The obvious question is what should the value of their duration of unemployment be for Person

Figure 4 A diagram of a hypothetical study of unemployment

D and Person E? Both could be ascribed a duration value of 18 months. This would mean that Person C, who genuinely was unemployed for 18 months, would be treated similarly to Person D, who was unemployed for 21 months, and Person E, who in fact never exited unemployment. This method of approximation is evidently imperfect.

The durations for Person D and Person E are examples of censored data. Estimating a standard regression model using data which include censored observations is problematic because they do not have appropriate values for the duration measure. The impact on the results obtained using a standard regression model may sometimes be negligible. Plewis (1997) states that when there is a very small proportion of censored cases, they will have little effect and an accelerated life model might still be suitable. In our experience few supervisors, examiners and referees will be convinced that when there are censored observations this is a suitable approach!

The durations for Person D and Person E in Figure 4 are 'right censored'. Right-censored data are very typical in large-scale social science datasets. Data can also be left censored, which occurs when the observation window interrupts an ongoing process. Consider data from a study of romantic relationships. At the beginning of the study some couples may be just at the start of their relationship, but other couples will be in existing relationships that began long ago. Left-censored data are less common in social science datasets because it is often feasible to ask respondents retrospectively about the starting point of a relevant 'spell'. Studies such as the BHPS collect retrospective information, for example on marital relationships and job tenure.

Duration models

Cox (1972) proposed a modelling approach for analysing durations which include censored observations. The main technical feature of this modelling approach is that we no longer directly model the duration. In duration models the focus is on modelling the probability that an event occurs at time t , conditional on it not having occurred before t .

Here is a more intuitive explanation. Think of a professor walking down a corridor. We could ask the question: What is their probability of dying in the next nanosecond, given that they have not already died? A nanosecond later we could again ask the question: What is their probability of dying in the next nanosecond, given that they did not die in the previous

nanosecond? Now imagine that we continued to make these probabilistic calculations repeatedly using these infinitesimal periods between two time points. We could eventually sum all of these probabilities and we would have a probability distribution. This probability distribution is often called the hazard function. As Allison (1984) states, the fundamental concept in duration models is that the hazard is the outcome that is modelled. This might not at first be obvious and indeed some years ago Sir David Cox commented that a less-appealing aspect of the Cox model is the difficulty of forming a direct physical interpretation of the hazard (Reid, 1994).

We will now move to a short example using some duration data to illustrate the main concepts associated with duration models. We will include some illustrative examples using the software package Stata, which will be discussed further in Chapter 7.

An English city-based college has a computing skills program for women who are unemployed. They collect data in order to evaluate the program. The individual student's enrolment ends when they pass an online test. In particular the college principal is an advocate of massive open online courses (MOOCs) so the college randomizes one group of participants to a course taught by standard methods and another group to the new MOOC. At the beginning of the study 628 women were enrolled on the program. By the end of the study 508 women had passed the test and 120 were right censored (because they had not passed by the end of the study).

A Stata codebook for the data is provided in Figure 5. The dataset includes some additional variables such as the woman's age at enrolment, the number of previous jobs she has had, which of the two college campuses she was inducted at and her educational level. The woman's

Variable	Obs	Unique	Mean	Min	Max	Label
id	628	628	314.5	1	628	student id
time	628	338	234.7038	2	1172	number of days until test passed
test	628	2	.8089172	0	1	test passed (or censored)
age	623	31	32.36918	20	56	age at enrolment
no_jobs	611	28	4.574468	0	40	number of previous jobs
mooc	628	2	.4904459	0	1	taught by massive open online course
campus	628	2	.2929936	0	1	college campus
quals1	628	2	.4601911	0	1	no qualifications
quals2	628	2	.1815287	0	1	lower qualifications (below A'level)
quals3	628	2	.3582803	0	1	higher qualifications (above A'level)

Figure 5 Stata output: Compact codebook for the college skills program dataset

educational level is measured by three dummy variables that take the value of 0 or 1.

The overall mean number of days that the women were enrolled was 235. The mean number of days enrolled for the women who passed the test was 154. The mean number of days enrolled for the women who had not passed the test by the end of the study was 576.

It is important to stress that the duration analysis that is presented below will assess the relative influence that a set of factors have upon how quickly the college students passed the online test. This is a substantively different emphasis to exploring the influences on the 'prevalence' or chance of a woman passing the online test. This is a subtle but important point to consider because it is easy for data analysts to misleadingly express their modelling results in terms of general prevalence, rather than in terms of effects on durations.

The start of a duration analysis in Stata requires that you declare that survival time data are being analysed using the *stset* command

```
stset time, failure(test)
```

The first part of the command declares which variable measures time (which in our example is helpfully called *time*). The second part of the command declares which variable denotes the event. Unfortunately, for historical reasons the 'event' is called a 'failure'. In our example the 'failure' is passing the online test! In other examples where the outcome is death or a light bulb burning out, it is easier to think of the event as a failure.

It is possible to get an overall description of the duration data using Stata's *stdes* command:

```
stdes
```

The output is reported in Figure 6. At the beginning of the study the risk set contains all 628 of the women in the study. When a woman passes the test she leaves the risk set and cannot re-enter it. The minimum time before a woman passed the test was 2 days and the maximum time on the program was 1,172 days. The median time on the program was 166 days. Once again an unhelpful terminology features in the output. In our example the 'time at risk' is the period that the woman is at risk of passing the test.

A descriptive graph called a Kaplan-Meier plot is useful for summarizing overall distributions of the time variable. These curves are sometimes called survival curves or survivor functions, and they illustrate the pattern and timing of individuals 'leaving the risk set' (i.e. passing the test). The survivor functions for both women taught by standard methods and women taught using the new MOOC are plotted in Figure 7. The curves appear to indicate that women taught by standard methods leave the

```

failure _d: test
analysis time _t: time

```

Category	total	per subject			
		mean	min	median	max
no. of subjects	628				
no. of records	628	1	1	1	1
(first) entry time		0	0	0	0
(final) exit time		234.7038	2	166	1172
subjects with gap	0				
time on gap if gap	0				
time at risk	147394	234.7038	2	166	1172
failures	508	.8089172	0	1	1

Figure 6 Stata output: *stdes* command for the college skills program data

Figure 7 Stata output: Kaplan-Meier plot of time to passing the test by teaching methods (college skills program data)

risk set (i.e. pass the online test) sooner than their counterparts who are taught using the new MOOC.

We can formally test the difference in the *survivor functions* of the women taught by standard teaching methods and the women taught by the new MOOC using the following Stata syntax:

```
sts test mooc, logrank
```

The results of the test of the equality of survivor functions are reported in Figure 8. The observed number of passes within the observation window for the women who were taught by standard methods was 265, which was higher than the expected number 236. By contrast the observed number of passes for women taught using the new MOOC was 243, which was lower than the expected number 272. The null hypothesis states that there is no structured or systematic difference between the survivor curves for the women taught by standard methods and the women taught using the new MOOC. By default Stata performs a log-rank test (see Mantel and Haenszel, 1959). We can reject the null hypothesis since $p < 0.01$. There are a number of alternative tests which could be computed using Stata, for example the Wilcoxon (Breslow-Gehan) test (Breslow, 1970; Gehan, 1965). In practice there is usually little difference between the results of these tests (see Lee and Wang, 2003).

We now move on to estimating the Cox regression model, which is a standard approach for modelling durations. We include age at

Figure 8 Stata output: Log-rank test for equality of survivor functions

enrolment, number of previous jobs, type of teaching method, college campus and previous education. The Cox model is estimated using the following Stata syntax:

```
stcox age no_jobs mooc campus quals2 quals3, nohr
```

In this syntax all of the variables are explanatory (i.e. things that might influence differences in durations). The outcome variable is the 'hazard' of passing the test, which is related to the duration, and it is already known to Stata because it has been declared in the *stset* specification and therefore does not need to be explicitly identified in the Stata syntax for estimating this model. Cox regression models can be thought of as another type of statistical model that falls under the generalized linear model (glm) umbrella, where the model assesses how multiple explanatory variables may influence variations in an outcome (see Nelder and Wedderburn, 1972).

The output for the Cox regression model is reported in Figure 9. The first lines of the output report that the failure is measured by the variable *test* and the analysis time (i.e. the duration) is measured by the variable *time*.

```

      failure _d: test
analysis time _t: time

Iteration 0:  log likelihood = -2868.555
Iteration 1:  log likelihood = -2851.6989
Iteration 2:  log likelihood = -2851.0884
Iteration 3:  log likelihood = -2851.0863
Refining estimates:
Iteration 0:  log likelihood = -2851.0863

Cox regression -- Breslow method for ties

No. of subjects =          610                Number of obs   =          610
No. of failures =          495
Time at risk   =        142994

LR chi2(6)      =          34.94
Prob > chi2     =          0.0000

-----+-----
      _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      age |   -.0237543   .0075611    -3.14   0.002   -.0385737   -.0089349
no_jobs |    .034745   .0077538     4.48   0.000    .0195478   .0499422
mooc    |   -.2540169   .091005    -2.79   0.005   -.4323834   -.0756504
campus  |   -.1723881   .1020981    -1.69   0.091   -.3724966   .0277205
quals2  |    .2467753   .1227597     2.01   0.044    .0061706   .4873799
quals3  |    .125668    .1030729     1.22   0.223   -.0763513   .3276873
-----+-----

```

Figure 9 Stata output: Cox regression model time to passing the test (college skills program data)

The next block of output reports the iterative steps that are undertaken to estimate the final model. The Cox proportional-hazards likelihood is a particularly difficult function to estimate numerically so the software places extra effort into ensuring appropriate precision, which is why Stata reports that it is 'refining estimates'.

The output reports that the 'Breslow method for ties' is used. If time could be measured on a true continuous scale then no observations would leave the risk set at the same time. In reality, because of the resolution or scale on which we measure time, there may be observations that simultaneously leave the risk set (i.e. they are tied). This affects the size of the risk set because we do not know who left first. Several methods have been proposed for addressing this problem and the default approach in Stata is the Breslow method (see Breslow, 1974). Stata reports that there are 610 women in the analysis and 495 passed the test (i.e. were failures). The total time at risk was 142,994 days. The log likelihood of the current model can be converted into a deviance (i.e. $-2 * \log \text{likelihood}$). The difference between the deviance of the null model and the deviance of the current model is reported as the likelihood ratio chi-square (LR χ^2). This can be calculated manually in Stata:

$$\text{display } (-2 * e(l_0)) - (-2 * e(l))$$

The LR χ^2 is 34.94 at 6 degrees of freedom (i.e. 1 degree of freedom for each explanatory variable in the model), and it is significant ($p < 0.01$). The LR χ^2 value can be considered to be analogous to the F test in a standard linear regression model.

The next panel of the output can be interpreted in a similar fashion to a *logit* model. Coefficients for each explanatory variable are reported on the logit scale along with their standard errors and a z test (which is the ratio of the estimate and its standard error). A p value for the estimate is reported and a 95 per cent confidence interval is constructed. The coefficients can be considered as the effect on the log of the hazard (Allison, 1984).

The overall message from the results in Figure 9 is that the age at enrolment, the number of previous jobs and the type of teaching method are all significant. The negative coefficients for age and for being taught by an MOOC suggest that on average they are associated with a lower hazard of passing the test (i.e. a longer duration until completion). A higher number of jobs is associated on average with a higher hazard of passing the test

(i.e. quicker completion). The effects of the college campus are not significant. The effects of previous education are less obvious. We can formally test the overall (or global) significance of qualifications in this model using the following Stata syntax:

```
test quals2 quals3
```

The chi-square test of previous education in the Cox model that is reported in Figure 10 is not significant ($p = 0.11$). This result can also be derived by comparing the difference between deviance (i.e. $-2 \times \log$ likelihood) for the model reported in Figure 9 with the deviance of a model with the previous education measures excluded.

```
( 1)  quals2 = 0
( 2)  quals3 = 0

      chi2( 2) =      4.36
      Prob > chi2 =      0.1130
```

Figure 10 Stata output: Test of the effects of previous education in Cox regression model of time to passing the test (college skills program data)

```
      failure _d: test
analysis time _t: time

Iteration 0:  log likelihood = -2868.555
Iteration 1:  log likelihood = -2855.3635
Iteration 2:  log likelihood = -2854.6762
Iteration 3:  log likelihood = -2854.6735
Refining estimates:
Iteration 0:  log likelihood = -2854.6735

Cox regression -- Breslow method for ties

No. of subjects =      610          Number of obs   =      610
No. of failures =      495
Time at risk   =     142994

Log likelihood = -2854.6735          LR chi2(3)      =      27.76
                                      Prob > chi2     =      0.0000
```

	_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
age		.9794475	.0072674	-2.80	0.005	.9653067 .9937955
no_jobs		1.036128	.0078949	4.66	0.000	1.020769 1.051718
mooc		.7940896	.0716076	-2.56	0.011	.6654445 .9476047

Figure 11 Stata output: Hazard ratios Cox regression model time to passing the test (college skills program data)

A more suitable model might therefore only include age at enrolment, number of previous jobs and method of teaching.

Some data analysts prefer to estimate the Cox regression model and display estimates as *hazard ratios* (which can be thought of as odds). The following Stata syntax omits the option *nohr* (i.e. 'no hazard ratio'):

```
stcox age no_jobs mooc
```

Conceptually the hazard of passing the test at time t for the women who were taught by the new MOOC was $(1 - 0.79) \times 100 = 21$ per cent lower (*ceteris paribus*). Or expressed another way the women who were taught by the standard methods passed the test sooner than the women who were taught by the new MOOC. Conceptually the hazard of passing the test at time t increased (*ceteris paribus*) by 3.6 per cent for each one unit increase in a woman's number of previous jobs (i.e. $(1.036 - 1) \times 100 = 3.6$ per cent). In some research areas the effect of an explanatory variable in a Cox model is described as a 'relative risk'.

One way to quantify these effects is to plot the survival functions to illustrate comparisons. Figure 12 shows the survival functions for women who were aged 30 and had 5 previous jobs. The survival function for

Figure 12 Stata output: Time to passing the test – survival functions comparing women aged 30 with 5 previous jobs by teaching methods (college skills program data)

women taught by the MOOC indicates that it generally takes them longer to leave the risk set (i.e. pass the test).

In the current model we have made no explicit consideration about the shape of the hazard function. An assumption of the Cox model is that the hazards are proportional, that is, that the effect of a given covariate does not change over time (Persson, 2002). In the statistical literature this is sometimes referred to as the ‘proportional hazards assumption’. There are several methods for verifying that a model satisfies the assumption of proportionality. One method is to include time-dependent covariates in the model. The options *tvc* and *texp* are available in Stata. An alternative method for assessing the proportionality assumption is to test Schoenfeld residuals (Grambsch and Therneau, 1994). The test can be undertaken in Stata using the following syntax after the *stcox* model has been estimated:

```
estat phtest, detail
```

Parametric duration models

The Cox regression model estimates the baseline survivor function without any reference to a theoretical distribution (Hamilton, 2012). This is why in the statistical literature the model is sometimes referred to as a ‘semi-parametric’ model. In practice, this means that models can be estimated without the researcher knowing (or having to assume) an appropriate distribution, which is an attractive feature of the model for empirical research. There are a number of ‘parametric’ versions of the model, which are based on the idea that the survival times follow a known distribution. These models have the same general form as the Cox model but define the baseline hazard differently. A useful way to conceptualize parametric functions of durations is to consider them as representing the accumulation of something. This accumulation begins at $t = 0$ and therefore we can assume that the measurement is 0 at $t = 0$. For example, primary smoking-related lung damage begins when the person starts smoking cigarettes. The general methodological advice is that how the accumulation process works should inform the choice of parametric function used in the modelling framework.

A range of parametric duration models can be estimated using Stata’s *streg* command. Here we will mention only two distributions. The first is

the exponential distribution which specifies a constant hazard. The underlying assumption is that events are independent and are not influenced by time. Therefore, in an analysis, individuals would be no more likely to experience the event later in the study than they were earlier in the study. This distribution is suitable if there is no 'ageing' effect in the social process that is being studied. For many social science research questions this is an unrealistic assumption. The second distribution is the Weibull distribution, which does not require failure rates to be constant over time and allows them to smoothly increase or decrease.

Zorn (2000) provides an advanced, but accessible, account of the potential challenges associated with duration dependence in the context of social science research. When there are several possible parametric models that could be used for an analysis, thought should be placed into which model is most appropriate. A sensible strategy for evaluating alternative models is to use a measure such as the Akaike Information Criterion (AIC) (see Akaike, 1974).

Competing risk models

The techniques that have been outlined above have all involved the analysis of a simple single event. We can easily envisage social science examples where a process ends with more than one possible outcome. Imagine a study of young people claiming unemployment benefits. The spell is unemployment and it may be ended by one of several substantively interesting alternative outcomes, for example gaining a full-time job, gaining a part-time job, entering a training scheme, returning to education or leaving the labour market. The 'competing-risks model' extends the Cox model to handle multiple outcomes (see Fine and Gray, 1999). We can think of these multiple outcomes as 'risks' of things that compete to end the respondent's spell (i.e. time in the initial state).

For many social science analyses where the phenomenon of interest is characterized by more than two states the competing-risks model may be more suitable than a simpler Cox model. The competing-risks model can be estimated in Stata using the *stcrreg* command. We do not present a competing-risks example as it is sufficiently similar to the Cox model. A limitation of the competing-risks model is that the relative hazards of different risks can be quite complicated to describe appropriately. When there are a number of different competing risks, the identification of

model parameters can be difficult. In practice researchers often choose to strategically limit the number of different competing risks that they explore within one model. The drawback is that this can sometimes lead to an inappropriate simplification of the characteristics of more complex social processes.

Discrete-time models

Discrete-time models are an alternative approach to duration data analysis and are also widely used in the social sciences. In discrete-time duration data the spell is partitioned into a number of different components that are aligned to distinctive or 'discrete' time periods. An essential property of the model used for the analysis of discrete time is the discrete hazard. This is the conditional probability that an individual will experience an event during the current discrete-time period, given that they have not previously experienced it. Singer and Willett (2003) advocate discrete-time approaches as a convenient alternative to continuous-time approaches and they argue that the simplicity of discrete-time approaches makes them an ideal entrée into the analysis of duration data. Stata does not have a suite of in-built commands for discrete-time models. This is because the models can be easily estimated using standard techniques.

The techniques and methods that have been outlined so far in this chapter use data that are recorded in continuous time. Continuous-time data are usually organized in a 'wide format' (this is often known as a 'person-time' dataset). In this format each row of the data represents a respondent or unit in the study. Discrete-time analysis requires what is sometimes referred to as a 'person-period' dataset. This is a 'long format' dataset where each row represents a 'person' at a given discrete-time period. For example, a dataset might contain several rows per person, one for each annual wave of a household panel survey. In Chapter 7 we outline the formats of 'long' and 'wide' datasets.

Like most data enabling tasks there are a number of possible approaches to organizing discrete-time data. The data need to be 'expanded' from a (continuous) 'person-time' format to a (discrete) 'person-period' format (i.e. from 'wide' to 'long' format). There are a number of facilities to do this, but we advocate using the Stata command *prsnperd*, which is a utility for creating discrete-time datasets from continuous-time data, although some researchers prefer Stata's *stsplit* command.

The package *dthaz* provides tools for discrete-time analysis and must first be downloaded using the syntax *net install dthaz* from <http://www.alexisdinno.com/stata>.

Once the data are in the person-period format, it is possible to estimate a discrete-time model using standard cross-sectional modelling techniques, for example a logistic regression model.

Discrete-time duration data analysis has long been recognized as a useful alternative means of exploring durations in the social sciences (Tuma, 1982). Although the format imposes some granularity upon the measure of duration, it does mean that standard models can be used to analyse the data. The techniques have been successfully applied in many studies. For example, Henley (1998) estimated single and competing risk discrete-time duration models of residence duration using the BHPS. Ermisch and Francesconi (2000) used discrete-time models to analyse cohabitation using BHPS. Berrington and Diamond (2000a) used discrete-time competing-risks models to analyse first partnerships in the NCDS.

Conclusions

Duration models are useful tools when the focus of the research question is the time to an event, and the event is a relatively crisp change between states. Because of the general appeal of analysing time to an event, duration models have been applied in areas as diverse as medicine, engineering, economics and sociology. The Cox model provides a robust and flexible starting point and Therneau (1997) states that it has become the workhorse of duration analyses with censored data.

The distinction between analysis in continuous-time frameworks and discrete-time frameworks is sometimes highlighted. Allison (2010) provides the following succinct, and in our view cogent, recommendations. First, if the exact time at which an event occurs is known then it is appropriate to use a continuous-time approach. Second, if only the month or the year of the event is recorded, discrete-time methods are better suited. Third, a very good indication of the need for discrete-time methods is the presence of a large number of ties (e.g. where two or more individuals experience an event at the same recorded time). Fourth, in some circumstances time is genuinely discrete, in the sense that events can occur only at certain discrete points in time.

Many different duration models can be estimated. The software Stata supports many options using its *st* suite (*st* stands for survival times). Blossfeld, Golsch and Rohwer (2007), Cleves (2016) and Royston and Lambert (2011) provide extensive and informative sources on using the *st* suite. The methods outlined in this chapter can be extended and more complex modelling approaches are available in both continuous-time and discrete-time frameworks. For example, multiple failure-time data, or multivariate survival data as it is alternatively known, is a natural extension to the techniques that have been outlined above. These more advanced techniques are suitable for the analysis of data from time-to-occurrence studies when two or more events occur for the same subject. Duration models can also be extended to circumstances where the durations for different units are related to each other, for example family members or classmates. In these studies, failure times are correlated within clusters of cases, violating the assumption of independence required in standard survival analysis (see Rabe-Hesketh and Skrondal, 2012). An outline of multiple-spell discrete-time analysis methods is provided by Willett and Singer (1995), and Steele, Goldstein and Browne (2004) document and apply a general multilevel multistate competing-risks model for event history data.

The central focus upon how long individuals remain in a certain state until an explicit event takes place means that the analysis of duration data is a more specific and narrowly defined approach in comparison to other methods of longitudinal data analysis. In the next chapter we turn our attention to more general strategies for analysing repeated contacts data.