

1

The fundamentals of survival and event history analysis

Objectives of this chapter

After reading this chapter, the researcher should be able to:

- Define, recognize and describe the **fundamental concepts and terminology** used in survival and event history analysis.
- Recognize and describe the **reasons** why we use these methods and the **types of problems that can be solved**.
- Define and understand different types of **censored and truncated data** and different types of censoring.
- Define and recognize a **density, survivor and hazard function**.
- Describe the **relationship between a density, survivor and hazard function**.
- Be able to argue **why it is necessary to use survival and event history models and their added value**.
- Recognize the **different types of survival and event history models and classes**.

Introduction: what is survival and event history analysis?

1.1 Survival and event history analysis is an umbrella term for a collection of statistical methods that focus on questions related to timing and duration until the occurrence of an event. The models examine the hazard rate, which is the conditional probability that an event occurs at a particular time interval (t). In other words, we examine how long it takes until the event of interest occurs. It is useful to note that survival models are actually just regression models with somewhat different likelihood estimators than OLS (ordinary least-squares regression). Students who know something about regression should therefore have little difficulty understanding survival models.

An event may take many forms, such as an organ transplant, marriage, birth, death, political revolution or bank merger. Due to fact that many research questions concern timing and duration, this method appeals to multiple scientific disciplines. The techniques described in this book are often referred to

as survival analysis in biostatistics, medical science and epidemiology, reliability analysis in engineering, duration models within economics, and event history analysis within sociology, demography, psychology and political science.

The goal of this book is to introduce these methods in an accessible, practical and engaging manner, starting from basic terminology and ranging to the most cutting-edge techniques used in the field today. Written for accessibility, this book will appeal to students and researchers who want to understand the basics and apply these methods without getting entangled in the mathematical and theoretical technicalities. Readers are offered a blueprint for their entire research project from research question, study design and data preparation to model selection and diagnostics, allowing them to independently master these advanced methods.

This book is written from the perspective of an applied researcher, with numerous examples and hands-on exercises, making it suitable as both a self-learning text or textbook within a course. Readers are provided with guidelines and suggestions on how to prepare data, run various types of models and enhance the expression of results with impressive graphics. Exercises within the body of the text are shown using the powerful and free computer program R, with Appendices and on-line material replicating some of the analyses using Stata (Appendix 2) and reference to SPSS and SAS on the companion web site to this book <http://www.gmw.rug.nl/~millssurvivaleha>.

This chapter begins by describing the fundamental concepts and terminology of these techniques, which is useful for beginners and serves as the foundation for the rest of the book. There is a sizeable amount of terminology related to these models that might be new for some researchers. Readers can therefore also refer to the **glossary** at the end of this book for a quick reference to the definitions of key terms. The mathematical expressions and relation of statistical functions are then presented in a manner that requires only a basic background in mathematics and statistics. We then turn to the logic of why it is necessary to use these types of models with certain data and research problems. The final section provides a brief overview of the different types of survival and event history models, which simultaneously serves as an overview of this book.

Key concepts and terminology

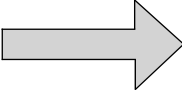
1.2 In survival and event history analysis, the **dependent variable** (also sometimes referred to as the response or outcome) is the hazard rate, which is the conditional probability that an event occurs at a particular time interval (t). In order to obtain statistical estimates of effects on this time to an event, most hazard rate results need to be transformed. Therefore, the dependent variable is a rate. The goal of these models is not only to examine the effects on the time until an event occurs, but also to assess the relationship of survival time to explanatory variables. **Explanatory variables** (also often referred

to interchangeably as covariates or independent variables) assess the impact of certain characteristics (e.g. receiving treatment, size of tumour, level of education) on the dependent variable. As we will explore in the chapters that follow, these variables may be fixed or time-varying. **Fixed** variables do not change across time and include static variables such as sex or place of birth. **Time-varying** variables have values that change over time such as age, labour force experience or size of a tumour.

The classic example of an **event** is death. The occurrence of an event is often referred to as a **failure**, usually attributed to fact that the event is death or disease. As Table 1.1 illustrates, an event could also be an infection, marriage, ‘death’ of a bank or the end of a United Nations peacekeeping mission. (The topic of censoring shown in this table is described shortly.) Another way to understand survival is in terms of **risk**. Take the following question for example, given that an individual has remained in remission from cancer for three years, what is the risk that he or she will experience a relapse within each unit of time?

Survival and event history analysis is highly interdisciplinary and used extensively in epidemiology and health sciences. This includes the study of leukaemia, heart transplant survival and infections of kidney dialysis patients, primary biliary cirrhosis, the effects of air pollution on mortality or AIDS (e.g. Geskus, 2000; Pope et al., 1995). In the social sciences, there are also numerous applications such as the study of organizational change (Hannan and Carroll, 1981), changes in hate crime law (Grattet et al., 1998), social insurance legislation (Usui, 1994), many applications in political science (see Box-Steffensmeier and Jones, 1997), social movements and the evolution of right-wing movements (Olzak, 1989), job mobility (Mills et al., 2006), marriage (Blossfeld and Mills, 2001) and union decline (Western, 1995).

Table 1.1 Examples of survival analysis showing starting time and event status

| <i>Start</i> | <i>Survival time</i> | <i>Event</i> |
|--|---|--|
| Patient with acute myelogenous leukaemia enters remission (Miller, 1997) | | Death or ‘censored’ (i.e. still alive at last observation) |
| Patient joins waiting list for heart transplant (Crowley and Hu, 1997) | | Death or ‘censored’ (i.e. still alive at last observation) |
| Insertion of catheter in kidney dialysis patient (McGilchrist and Aisbett, 1991) |  | Infection or ‘censored’ (i.e. no infection at last observation) |
| Woman in (non-marital) cohabiting relationship becomes pregnant (Blossfeld and Mills, 2001) | | Marriage or ‘censored’ (i.e. still cohabiting at last observation) |
| Commercial bank opens (Bergström et al., 1997) | | Closure of bank or ‘censored’ (i.e. bank still functioning at last observation) |
| Start of U.N. peacekeeping mission (Green et al., 1998) | | End of U.N. peacekeeping mission or ‘censored’ (i.e. mission still ongoing at last observation) |

Since we are concerned with analysing the time to the occurrence of an event, **time** is an essential aspect of these models and can be measured in diverse **units**, such as seconds, days, weeks, months or years. The duration or time that it takes before an event occurs is referred to as **survival time**. It is the time that a person or other unit of analysis (e.g. bolt in a machine, bank, political regime) 'survives' the specified duration. It is also often interchangeably referred to as a spell, episode, interval, waiting time, exposure time, risk period or duration.

The **time axis** may be **continuous or discrete**. If the time of the event is known precisely, it can be measured on a continuous scale (e.g. seconds, days, months). If the time units are unknown within larger units of years or decades, discrete-time methods are often used (Allison, 1982, 1984; Singer and Willet, 2003b). Discrete-time methods are therefore used when we have imprecise measurements and only know that the event occurred within a particular interval (e.g. within a year), but not the exact time; this is discussed in more detail in Chapter 9. In both continuous- and discrete-time models, the risk of the event occurring at time t is being modelled. Whereas the dependent variable in a continuous-time model is a hazard rate, in a discrete-time model it is the odds (if modelled using standard logit/probit models). The necessary precision of the timing of the event is highly dependent on the research question and often related to data restrictions. Although most processes occur in continuous time, they are often measured in discrete time, resulting in many event history models applying the discrete-time approach. Continuous-time models include the exponential survival model and the Cox semi-parametric model (Cox, 1972, 1975; Cox and Oakes, 1984). Results from discrete-time and continuous-time methods will be virtually the same in most models; furthermore, discrete-time models can be used to approximate continuous-time models (Allison, 1982; Yamaguchi, 1991).

These terms provide the basic foundations for simple models, but later chapters will also examine more complex topics such as frailty and recurrent events, competing risks and multistate models and modelling entire trajectories, defined in more detail at the end of this chapter. Models are generally divided into non-parametric, semi-parametric and parametric models, also described within the last section.

Censoring and truncation

1.3 A distinguishing factor of survival and event history models is that they take censoring into account. A simple definition of **censoring** is that we have information about an individual's survival time, but do not know the exact survival time (Kleinbaum and Klein, 2005). Various types of censoring can occur, with the most common type being right-censoring, which will

also be the primary focus in this introductory textbook. In most cases, **truncation** refers to the complete lack of information about the occurrence of the event.

There is often some confusion as to whether observations are censored or truncated. Strictly speaking, truncation refers to the cases where subjects do not appear in the data because they are not observed. Censoring refers to cases when subjects are known to fail within a particular episode, but the exact failure time is unknown. There are additional types of censoring (e.g. **partially left-censoring**), which are described in the glossary or in the detailed discussions of censoring and truncation in sources such as Kalbfleisch and Prentice (1980: 39–41), Allison (1984), Tuma and Hannan (1984: 118), Yamaguchi (1991: 3–9) and Vermunt (1997: 117–130).

Right-censoring

1.3.1 As Figure 1.1 illustrates, **uncensored cases** represent the information where we know both the starting and ending time of episodes. Most discussions of censoring and research projects involve **right-censoring**, which occurs when the event under study is not experienced by the last observation. This commonly occurs in the social sciences when survey data is used. Individuals are often questioned about their retrospective life histories, such as the birth dates of their children or start and end dates of jobs or education. If we were modelling the transition to second childbirth, for instance, using this type of retrospective data, all individuals who had a first child but no second child at the time of observation would be ‘right-censored’ by the survey date. In certain medical studies and panel or longitudinal designs, the individual may not be present for follow-up (e.g. they have moved), dropped out of the study (e.g. due to bad side effects or refusal to participate) or the study simply ends. The first two reasons are often considered as **random**.

A special condition occurs when an episode is right-censored as a result of a **non-random** process, often referred to as Type I censoring. This type of

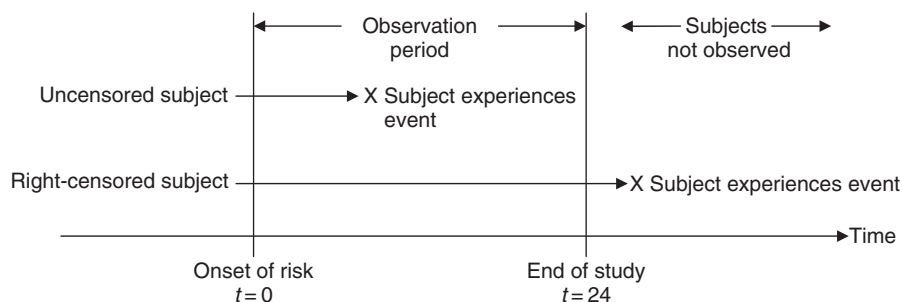


Figure 1.1 Uncensored versus right-censored subjects

censoring comes about more in disciplines such as engineering or those that use experimental designs. In engineering applications, the focus is on the strength of materials, fatigue, rupture in solids and bearings. In these studies, tubes, chips or bearings for instance, are all started for testing at a particular time, and the time until event is recorded until their failure (i.e. breakdown) However, since some items may take a very long time to fail or break down, the experiment is sometimes prematurely terminated at a predetermined non-random time.

An essential point is that although there is no information about the *occurrence* of the event for right-censored cases, there is information about the survival or exposure time until the last point of observation. Right-censored data constitutes missing data insofar as we have information on the event history of the unit of analysis and the time at risk up to the last observation point. Therefore, we require the assumption that censoring is random and that the processes governing censoring and occurrence of events are independent of one another (Tsiatis, 1975). There are straightforward ways to deal with right-censoring during data analysis, which we explore in the chapters that follow.

Interval censoring

1.3.2 Interval censoring refers to the case where we only have information that the event occurred between two known time points, but not the exact timing of the event. This type of censoring comes about when subjects are questioned or tested at fixed time-points during a specified follow-up period. In a longitudinal panel study, for example, data might be collected from subjects once every 2 years. Consider if we were studying employment status (e.g. employed, unemployed, out of labour force), and only the employment status categories were asked every two years and not the timing of changes. If someone were unemployed at the first data collection wave but employed at the second wave we would know that there was a change in employment status, but not exactly *when* the event occurred during this period. In clinical research, patients may be required to visit a clinic once every month for a period of several years. It might be that a patient tested negative at month 9 and positive at month 10, leaving us with the knowledge that the event occurred between the 9th and 10th clinic visit, but we would have no information about exactly *when* the event occurred.

Truncation

1.3.3 Truncation is a condition other than the event of interest that is, for example, used to screen respondents or patients (Klein and Moeschberger, 1997). The most common type of truncation is **left-truncation**, such as when subjects enter the study at a random age. In the case of left-truncation, we do not have information from before the onset of risk to some

time after the onset of risk. In other words, the subject was not observed for some time at the beginning of the process but then came under observation.

Another type of truncation is **interval or gap truncation**, which is similar to left-truncation. This could occur in a clinical study if, for example, a patient is under observation for the first 3 months of the study, drops out for 2 months and then rejoins the study again for the last 7 months. Dropping out of the study for 2 months creates an interval or gap in the period of observation. Both left and interval truncations are dealt with during analyses by omitting the subject from all individual binary-outcome analyses during the truncation period (or gap) due to the fact that they could not have experienced the event during those periods.

Right-truncation can also occur, but is less frequent. Klein and Moeschberger (1997) provide the example of the examination of an episode from HIV infection until the development of AIDS. If the sample only includes those who have developed AIDS prior to the end of the study, those HIV-infected individuals who have not yet progressed to AIDS are excluded from the sample.

Mathematical expression and relation of basic statistical functions

1.4 This section provides a brief outline of the key statistical concepts of survival analysis. The text is written at the level of a non-mathematician and primarily for applied researchers, and for this reason does not go into extensive mathematical or theoretical depth. That being said, understanding these expressions and how they are calculated are key for your general understanding and interpretations of these methods. Non-mathematicians can refer to Box 1.1 for a review of some of the basic notation used in the equations. The mathematical expression may at first glance appear to be somewhat abstract, but will become clearer in the upcoming chapters when we estimate models and explicitly show calculations using real data.

Box 1.1 A REVIEW OF NOTATION FOR NON-MATHEMATICIANS

| | |
|------------|---|
| T | Random variable of survival time ($T \geq 0$) |
| $T \geq 0$ | Means that T can be any number equal to or greater than zero (i.e. cannot have negative values) |
| t | Specific value for T |
| δ | (0,1) Random variable = 1 if failure, = 0 if censored (Greek lower-case letter delta) |

(Continued)

(Continued)

| | |
|---------------------------------|---|
| $\hat{S}(t)$ | Survivor function |
| $h(t)$ | Hazard function |
| $f(t)$ | Density function |
| ∞ | Infinity |
| P | Probability |
| | Given |
| Δt | Small time interval (Greek upper-case letter delta) |
| lim | Limit |
| lim $\Delta t \rightarrow 0$ | Instantaneous potential |

The starting point is to define T as a positive **random variable** that represents the survival times. It is assumed to be continuous (except where we examine discrete-time models). The actual survival time is a value of T , which is denoted as t . The values of T have a particular probability distribution, denoted by a **probability density function** represented by $f(t)$ and a **cumulative density function**, $F(t)$. The distribution function of random variable T is given by:

$$F(t) = \int_0^t f(u) d(u) = \Pr(T \leq t) \quad (1.1)$$

where $\Pr(T \leq t)$ is the probability that a survival time T is less than or equal to some value t . For all points at $F(t)$, the **probability density function** $f(t)$ is defined as:

$$f(t) = \frac{dF(t)}{d(t)} = F'(t) \quad (1.2)$$

This implies:

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \quad (1.3)$$

The density function $f(t)$ expresses the unconditional instantaneous probability that an event occurs in the time interval $(t, \Delta t)$ and is formally specified as:

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr(t \leq T \leq t + \Delta t)}{\Delta t} \quad (1.4)$$

We can see from both Equations 1.3 and 1.4 that the density function is an unconditional failure rate. In other words, it describes the unconditional

(i.e. not conditioned on covariates) instantaneous (at any given instant t) probability of the event (i.e. failure rate).

The **survivor function** is another core concept in survival and event history models and is specified as:

$$\hat{S}(t) = 1 - F(t) = \Pr(T \geq t) \quad (1.5)$$

which expresses the probability that a survival time T is equal to or greater than some time t . $\hat{S}(t)$ denotes the proportion of subjects surviving beyond t . At origin time $t = 0$, $S(0) = 1$, which simply means that all subjects in the study are surviving at $t = 0$. As we will see in the upcoming chapters, $\hat{S}(t)$ is a function that strictly decreases over time as the surviving subjects fail over time.

The occurrence of an event (e.g. failure) and survival are related to each other, which is encapsulated by the **hazard rate**, which also goes by the name of the instantaneous transition or hazard function:

$$h(t) = \frac{f(t)}{\hat{S}(t)} \quad (1.6)$$

The hazard rate indicates the rate at which subjects fail by t given that the subject has survived until t . It is therefore a conditional failure rate, which can be seen by:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \quad (1.7)$$

where the transition rate $h(t)$ represents the instantaneous risk that the event occurs in the time interval $[t, t + \Delta t]$, given survival at or beyond time t . The hazard therefore focuses on failing (i.e. experiencing the event) whereas the survivor function focuses on surviving (i.e. not experiencing the event).

Why use survival and event history analysis? _____

1.5 A core question you need to ask with this (and arguably with any type of data analysis) is: Why is it necessary to engage in this type of analysis? What does it offer that ordinary regression models such as ordinary least-squares regression or logistic regression do not?

Potential problems that might arise if censored data is ignored
1.5.1 Anyone working with censored data should seriously consider using these types of methods. To illustrate the serious problems that may occur when you ignore censored data, we can refer to the study

discussed in Miller (2007) previously shown in Table 1.1 and which is often used in this book as an example dataset, since it is compact and easy to decipher and is also part of the ‘survival’ library package in R. The dataset is described in more detail in the Appendix. The study contains the results from a clinical trial that examined the efficacy of maintenance chemotherapy for individuals with acute myelogenous leukaemia (AML). The goal was to see whether maintenance chemotherapy extended the time until relapse. Once patients entered into remission after treatment by chemotherapy, they were randomly assigned into two groups: ‘maintained’ (continued to receive maintenance chemotherapy) and ‘nonmaintained’ (control group who received no chemotherapy).

Figure 1.2 illustrates how the simple mean duration of the time until the event (death or censoring) would be calculated if we ignored censoring (i.e. removed censored observations) versus accounting for censoring. Using the first approach, which does not take censoring into account, the difference between the two groups appears to be negligible (and the medians that are identical – 23.0 – for both groups). We might conclude from the analysis that the survival time for the group that received maintenance chemotherapy was only slightly more skewed to the right, or in other words had only a marginally higher survival time than the control group. When we calculate the mean properly by also accounting for censored data, we see a markedly larger gap between mean in the maintained versus the non-maintained group. Here the medians have strikingly different values, 31.0 and 23.0 for the maintained and non-maintained groups respectively. In fact, the actual distribution of the maintained group is far more right-skewed and the survival difference between the two groups is very large.

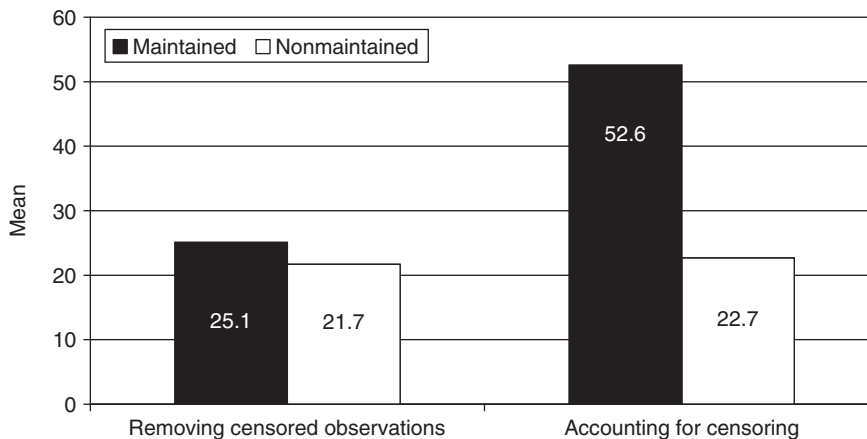


Figure 1.2 Difference in mean in AML study by removing (ignoring) censored observations versus accounting for censoring

What does survival analysis offer that ordinary regression models do not?

1.5.2 These techniques have several unique features that distinguish them from other types of methods. What is the added value of these models? First and foremost, survival analysis *adds information about timing*. As previously described, this in turn makes it possible to account for ‘censoring’. Unlike other techniques, it also takes a different approach by not only focusing on the outcome but also analysing the time to an event. This enables us to compare the survival between two or more groups and to assess the relationship between explanatory variables and survival time. Another unique feature is the ability to include time-varying covariates (i.e. explanatory variables that change their values over time such as age or education), which is not possible in OLS or logistic regression. For this reason, these models are often referred to as dynamic or process models (Aalen et al., 2008; Willekens, 1991).

Overview of survival and event history models and this book

1.6 This book will introduce you to various types of survival and event history models, which are summarized in Table 1.2 in terms of the class and type of model, a brief description and the advantages and disadvantages, and in which chapter they are covered. Do not feel disheartened if you do not understand some of the terminology or subtle differences between the models at this point. As you continue to read each chapter, the key aspects of each class of models should become clear. You can also refer to the glossary at the end of this book for terms that might still be unfamiliar at this point. The table is not intended to be exhaustive, but rather to provide you with a brief synopsis or ‘helicopter view’ of the possibilities.

Non-, semi- and parametric models

1.6.1 A common distinction often made in survival and event history modelling, shown and also in Table 1.2, is between non-, semi- and parametric models. In **non-parametric models**, which include life table and **Kaplan-Meier estimates**, there is no assumption about the shape of the hazard function or about how covariates may affect that shape. As described in Chapter 4, this is often an excellent preliminary descriptive technique to use at the beginning of your data analysis. Due to the fact that non-parametric methods cannot handle the inclusion of multiple covariates, researchers then often turn to regression techniques of semi-parametric and parametric models. **Semi-parametric models** such as the Cox and the piecewise constant exponential model discussed are particularly flexible since they make no

Table 1.2 Summary of survival and event history models

| <i>Class and type of model</i> | <i>Description</i> | <i>Advantages</i> | <i>Disadvantages</i> |
|---|---|--|--|
| Non-parametric (Chapter 4) | | | |
| – Life table estimates | – Makes no assumption about: <ul style="list-style-type: none"> • shape of hazard function • how covariates affect shape of hazard function | – Good method to understand basics and produce descriptive results | – Can only compare limited number of groups |
| – Kaplan-Meier (Product-Limit) estimator | – Effects of covariates shown by stratifying data into groups | – Life table: good for large data and crude measurement of event times – KM: good for smaller data and precisely measured event times | – Does not allow inclusion of multiple covariates and multivariate controls |
| Semi-parametric (Chapter 5) | | | |
| – Cox model (most prominent) | – Makes no assumption about shape of hazard function | – Flexible model, often initial exploratory choice in analyses | – Less appropriate for testing hypotheses about time-dependence (i.e. how hazard varies over time) |
| – Piecewise constant exponential model | – Makes strong assumption about how covariates affect shape of hazard function by assuming proportional hazard between groups over time – Partial-likelihood estimation | – Allows inclusion of multiple covariates, multivariate analysis – Results often similar to parametric models, but without (often) restrictive assumptions | – Less precise than parametric models – Sometimes called 'overfitted' |
| Parametric (Chapter 6) | | | |
| – Exponential, Weibull, logistic, Gamma, Gaussian, complementary log-log, log-logistic, log-normal, Gompertz, Makeham, extreme value, Rayleigh and others | – Researcher needs to decide in advance shape of the hazard function and how covariates impact the hazard function – Maximum likelihood estimation – Preferred when researcher wants to study the nature of time dependence and when time is meaningful in an independent variable – Continuous and discrete-time models | – More precise parameter estimates (if correct model assumptions) – Allows multivariate analysis – Allows analysis of discrete and continuous explanatory variables – Specifies the shape of the hazard function, allowing for predictive modelling | – If the hazard-function shape is incorrectly specified, parameter estimates can be seriously biased – Needs preliminary work to first define shape of hazard function and understand how covariates affect the hazard function – Very sensitive to included or omitted covariates |
| Discrete time and count (Chapter 9) | | | |
| – Logit, probit, logistic, negative binomial regression (NBR models) | – Analyzes the number of events since a defined starting time | – Useful for analysis of rare events (Poisson, NBR) when number of zeros (i.e. number of trials without any events) is large | – Does not examine duration, but rather event-counts |
| – Poisson | – Often for analysis of rare events (Poisson) | | |

Multilevel, frailty or recurrent event models (Chapter 8)

- Recurrent event or multiple episode models
- Frailty models, conditional frailty models (sometimes also referred to as multilevel models, random effect models)
- Some subjects more likely to experience repeated event due to unmeasured cause (unobserved heterogeneity)
- Understanding how covariate effects change across episodes
- Frailty: model as random effect
- Conditional frailty: modifies frailty model to adjust for event dependence, stratifies cases by event number
- Goes beyond single-episode models that only compare effects between covariates to examine how covariate effects change across episodes
- By estimating frailty as cause of unobserved heterogeneity as a random effect, coefficients for measured variables are less biased
- Frailty models may be badly biased if frailty is correlated with the covariates or the wrong distribution is assumed

Competing risk models (Chapter 10)

- Competing risk and multiple destination models: use one of the models described above (e.g. Cox) and make adjustments to risk group depending on whether risks are independent of one another
- Episode can end in two or more different outcomes
- Central assumption is often conditional independence of the risks under analysis
- Considers more complex destination states
- Treats different reasons as different events, allowing comparison of hazard functions across competing risks
- Problem if competing risks are not properly identified
- Hard to cope with assumption of conditional independence of the risks under analysis

Multistate models (Chapter 10)

- Multistate models (also overlaps with competing risk, recurrent event and alternating state models)
- Model for a stochastic process, which at any time point occupies one set of discrete states
- Specify state structure and form of hazard function for each transition
- Appropriate for event-related dependence
- Considers states, not events (problem for recurrent events)
- All data considered longitudinal; less useful for repeated measurements

Sequence analysis (Chapter 11)

- Discrete Markov models and optimal-matching-based clustering
- Obtain a matrix of proximities between sequences via optimal matching (or other metric) and cluster sequences via multidimensional scaling methods
- Provides a holistic view of entire event history
- Derives prominent characteristics of complete trajectories
- Remains highly descriptive if clusters not used as predictors in regression model

assumption about the shape of the hazard. Contrary to non-parametric methods, you are able to include multiple covariates. They are, however, part of the proportional hazards family of models, which (as described in detail in Chapter 5) means that they make a strong assumption about how the covariates affect the shape of the hazard function between groups over time.

The **parametric models** discussed in Chapter 6 include models such as the exponential, Weibull, Gamma, Gompertz and others (see Table 1.2). Leaving theory and mathematical details aside, the main difference between these models and a semi-parametric specification is that the researcher is required to decide in advance about the shape of the hazard function and how covariates might impact the function. The decision process for choosing between a semi-parametric and a parametric model is discussed in more detail in Chapter 7 (see also Figure 7.1).

Outline of this book

1.6.2 Since the goal of this book is to provide a highly practical approach, after the theoretical aspects of the models are defined in each chapter, researchers can continue to actively estimate and interpret the results. When describing the content of the chapters, I realize that I sometimes ‘jump ahead’ and use terminology that may be unfamiliar to some (e.g. ‘lagged covariates’, ‘episode-splitting’). This type of terminology is covered in detail in the chapters that follow, with key terms described briefly within the glossary.

To allow you to estimate the models, it is essential to first understand the basics of R, which is the primary statistical program accompanying this book. Chapter 2 is written for researchers with little or no experience using R, with an Appendix for users of other statistical programs. The first part of the chapter provides an **accessible introduction to R**, followed by basic **descriptive statistics and graphics**. Those already familiar with R, or desiring to estimate models in other programs, can skip this chapter.

Chapter 3 addresses different **types of data** for survival and event history analysis and tackles the often-daunting task of **data restructuring**. Such a practical chapter is rare in textbooks on this topic and will be of interest to those who need to restructure their own data independently. Those with experience in event history data and data restructuring can proceed directly to Chapter 4.

Survival and event history analysis often begins with **non-parametric models**, explored in Chapter 4, which include life-table and Kaplan-Meier (KM) estimates. After describing the fundamentals of the KM approach, this chapter describes how to produce and interpret estimates, plot survival curves, test differences between two groups and stratify the analysis by a covariate.

Semi-parametric models in the form of the Cox regression model are covered in Chapter 5. The Cox proportional model with fixed and time-varying covariates is first described and then models are estimated for each. In this chapter, you will learn how to estimate and interpret a Cox regression model, understand significance and plot the survival function. This is followed by a description of how to create a subject-period data file to accommodate time-varying covariates, adding lagged time-varying covariates to reduce problems of causal ordering. A final section describes how to model interactions with time by using episode-splitting at time intervals.

We then turn to **parametric models** in Chapter 6, where it is necessary to specify the shape of the hazard function in advance. After a description of the mathematical underpinnings of these models, the discussion turns to the distinction between models with a proportional hazards (PH) parameterization versus models that have an accelerated failure time (AFT) approach. The chapter then describes how to estimate and interpret the PH and AFT parameterizations of the exponential, piecewise constant exponential and Weibull models and the AFT estimation of log-logistic and lognormal models.

There are a variety of models to choose from and various decisions that you need to make when choosing an appropriate model. For this reason, Chapter 7 provides a detailed discussion of **model-building and model diagnostics**. This chapter focuses on model-building and the selection of covariates, assessing the overall goodness of fit of your model, testing overall model adequacy via Cox-Snell residuals, testing the proportional hazards assumption via Schoenfeld residuals, checking for influential observations with score residuals and assessing nonlinearity via Martingale residuals and component-plus-residual plots.

In recent years, there has also been growing attention to **recurrent event and frailty models**, sometimes also referred to as multi-level or random effect models or unobserved heterogeneity, which is the focus of Chapter 8. This chapter demonstrates how to examine **recurrent events** such as multiple relapses from remission, repeated heart attacks or infections, multiple marriages, births, cabinet durations or unemployment episodes. This relates to the broader topic of **correlated survival data**, which is often discussed in relation to frailty and unobserved heterogeneity. Correlation of event times can occur in the case of recurrent events or if subjects that experience a single event belong to a particular group or cluster (e.g. family, clinic). **Frailty** entails that some subjects may be more 'frail' and exhibit a higher likelihood of experiencing an event than others do. The assumption therefore is not that all subjects are homogeneous (i.e. similar), but that they are heterogeneous (i.e. different). Frailty models thus account for the **unobserved heterogeneity** that arises due to the potentially unobserved 'frailty' (Aalen, 1988, 1992; Hougaard, 1984, 2000; Vaupel et al., 1979). After description of modelling recurrent events and clustering in groups with shared frailty models, the discussion turns

to the identification of additional frailty models, including unshared, nested, joint and additive. Frailty models are then estimated and interpreted, assuming both Gamma and Gaussian distributions.

Another category of models that is common in contemporary research comprises **discrete-time models**, addressed in Chapter 9. In discrete-time models, the dependent variable is dichotomous (i.e. 0, 1) and data is arranged in the form of a subject-period file of discrete time units. After formal specification and description of the model, the chapter discusses data restructuring to prepare discrete-time files. We then move to the estimation of logit, probit and complementary log-log (cloglog) models, concluding with a reflection on the advantages and disadvantages of these types of models.

Chapter 10 introduces competing risks, but also the more advanced method of multistate models, the later of which is rarely integrated in introductory event history textbooks. This chapter acknowledges that different types of events can occur because there is more than one underlying process; in other words, there may be several competing causes, often referred to as **competing risks** (Crowder, 2001; Hachen, 1988; Larson, 1984). After discussing the three central techniques used to model competing risks, we discuss, and then estimate and compare, the more common latent or cause-specific approach against cumulative incidence curve (CIC) estimates. This is followed by an example of regression analysis with competing risks.

It is rare to find an accessible and practical introduction to **multistate models**, which are models that recognize that the subject may not only be at risk for more than one kind of competing event, but that this event can also occur more than once. These models focus on the evolution of the process and sequence of events and are often assumed to take the form of Markov models. After a brief introduction, the preparation of data to estimate these models is discussed, followed by estimation of Markov multistate models with stratified and proportional hazards and an extended Markov proportional hazards model.

Finally, Chapter 11 covers the more advanced method of **sequence analysis**, which models entire event histories. Just as multistate models, it is intuitively related to event history models, yet this technique has not been introduced in previous introductory textbooks. It is the analysis of categorical sequences of events and concerned with the order in which events occur. These techniques allow researchers to identify typical sequential patterns, establish why certain sequential patterns exist and establish the effects of a given sequential pattern on other outcomes (Abbott, 1995). After a brief review, this chapter describes how to prepare data and describe and visualize sequence data sets, measure similarities and distances between sequences using optimal matching (OM) distances, produce typologies of clusters using cluster analysis and engage in event sequence analysis. The chapter concludes with a discussion of critiques of OM and new advances in the field.

The book concludes with an **Appendix that describes the datasets** used throughout this book (Appendix 1). Although the focus of all applications in this book use the statistical package R, **Appendix 2 replicates virtually all of the commands in Stata** (up to and including Chapter 9). Readers can refer to the companion website to this book for examples of estimation in other packages such as SPSS and SAS. The book also contains a **glossary** of the terminology most commonly used in survival and event history models, described in a non-technical and introductory manner.

Exercises

- 1 Focusing on your own research topic, draw a basic survival model, including an uncensored and right-censored case as shown in Figure 1.1. If you are already working with existing data and a research question, focus on that topic. If you are not yet working with specific data, think of an ideal analysis that you would like to focus on. What is the event? What are the starting and the ending times of the process? Do you have right-censoring or any other types of censoring? Will you need to deal with truncation issues?
- 2 Do a brief literature review and find at least two studies in your area of research that use survival and event history models. What is the dependent or outcome variable? How is censoring defined? Which type of statistical model do they use and why? What are the key independent variables? Are these variables fixed or time-varying? Do they discuss any advantages or limitations of their approach?
- 3 What is the difference and relationship between the survival, hazard and density function?
- 4 Why would you arrive at different mean and median durations of the time until an event if you (a) did or (b) did not include censored observations?
- 5 Name at least two differences between Cox and parametric regression models.
- 6 What is the difference between a fixed and a time-varying covariate? Provide at least one example of each type of covariate.
- 7 What is the added value of survival analysis over other regression models such as OLS or logistic regression?