Survival Analysis

Stephen P. Jenkins

18 July 2005

Contents

List of Tables

List of Figures

Preface

These notes were written to accompany my Survival Analysis module in the masters-level University of Essex lecture course EC968, and my Essex University Summer School course on Survival Analysis.[1] (The first draft was completed in January 2002, and has been revised several times since.) The course reading list, and a sequence of lessons on how to do Survival Analysis (based around the Stata software package), are downloadable from

http://www.iser.essex.ac.uk/teaching/degree/stephenj/ec968/index.php.

Please send me comments and suggestions on both these notes and the do-it-yourself lessons:

Email: stephenj@essex.ac.uk

Post: Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom.

Beware: the notes remain work in progress, and will evolve as and when time allows.. Charts and graphs from the classroom presentations are not included (you have to get something for being present in person!). The document was produced using Scientific Workplace version 5.0 (formatted using the 'Standard LaTeX book' style).

My lectures were originally based on a set of overhead transparencies given to me by John Micklewright (University of Southampton) that he had used in a graduate microeconometrics lecture course at the European University Institute. Over the years, I have also learnt much about survival analysis from Mark Stewart (University of Warwick) and John Ermisch (University of Essex). Robert Wright (University of Stirling) patiently answered questions when I first started to use survival analysis. The Stata Reference Manuals written by the StataCorp staff have also been a big influence. They are superb, and useful as a text not only as program manuals. I have also drawn inspiration from other Stata users. In addition to the StataCorp staff, I would specifically like to cite

[1] Information about Essex Summer School courses and how to apply is available from http://www.essex.ac.uk/methods.

the contributions of Jeroen Weesie (Utrecht University) and Nick Cox (Durham University). The writing of Paul Allison (University of Pennsylvania) on survival analysis has also influenced me, providing an exemplary model of how to explain complex issues in a clear non-technical manner.

I wish to thank Janice Webb for word-processing a preliminary draft of the notes. I am grateful to those who have drawn various typographic errors to my attention, and also made several other helpful comments and suggestions. I would like to especially mention Paola De Agostini, José Diaz, Annette Jäckle, Lucinda Platt, Thomas Siedler and the course participants at Essex and elsewhere (including Frigiliana, Milan, and Wellington).

The responsibility for the content of these notes (and the web-based) Lessons is mine alone.

If you wish to cite this document, please refer to:

Jenkins, Stephen P. (2004). Survival Analysis. Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK. Downloadable from http://www.iser.essex.ac.uk/teaching/degree/stephenj/ec968/pdfs/ec968lnotesv6.pd

Chapter 1

Introduction

1.1 What survival analysis is about

This course is about the modelling of *time-to-event data*, otherwise known as *transition data* (or *survival time data* or *duration data*). We consider a particular life-course 'domain', which may be partitioned into a number of mutually-exclusive states at each point in time. With the passage of time, individuals move (or do not move) between these states. For some examples of life-course domains and states, see Table 1.1.

For each given domain, the patterns for each individual are *described* by the time spent within each state, and the dates of each transition made (if any). Figure 1, from Tuma and Hannan (1984, Figure 3.1) shows a hypothetical marital history for an individual. There are three states (married, not married, dead) differentiated on the vertical axis, and the horizontal axis shows the passage of time t. The length of each horizontal line shows the time spent within each state, i.e. *spell lengths*, or *spell durations*, or *survival times*. More generally, we could imagine having this sort of data for a large number of individuals (or firms or other analytical units), together with information that describes the characteristics of these individuals (to be used as explanatory variables in multivariate models).

This course is about the methods used to model transition data, and the relationship between transition patterns and characteristics. Data patterns of the sort shown in Figure 1 are quite complex however; in particular, there are *multi-state transitions* (three states) and *repeat spells* within a given state (two spells in the state 'not-married'). Hence, to simplify matters, we shall focus on models to describe survival times within a *single state*, and assume that we have *single spell* data for each individual. Thus, for the most part, we consider exits from a single state to a single destination.[1]

[1] Nonetheless we shall, later, allow for transitions to multiple destination states under the heading 'independent competing risk' models, and shall note the conditions under which repeated spell data may be modelled using single-spell methods.

Domain	**State**
Marriage	married
	cohabiting
	separated
	divorced
	single
Receipt of cash benefit	receiving benefit x
	receiving benefit y
	receiving x and y
	receiving neither
Housing tenure	owned-outright
	owned with mortgage
	renter – social housing
	renter – private
	other
Paid work	employed
	self-employed
	unemployed
	inactive
	retired

Table 1.1: Examples of life-course domains and states

We also make a number of additional simplifying assumptions:

- the chances of making a transition from the current state do not depend on transition history prior to entry to the current state (there is *no state dependence*);

- entry into the state being modelled is exogenous – there are *no 'initial conditions' problems*. Otherwise the models of survival times in the current state would also have to take account of the differential chances of being found in the current state in the first place;

- the model parameters describing the transition process are fixed, or can be parameterized using explanatory variables – the process is *stationary*.

The models that have been specially developed or adapted to analyze survival times are distinctive largely because they need to take into account some special features of the data, both the 'dependent' variable for analysis (survival time itself), and also the explanatory variables used in our multivariate models. Let us consider these features in turn.

1.2 Survival time data: some notable features

Survival time data may be derived in a number of different ways, and the way the data are generated has important implications for analysis. There are four main types of sampling process providing survival time data:

1. *Stock sample* Data collection is based upon a random sample of the individuals that are currently in the state of interest, who are typically (but not always) interviewed at some time later, and one also determines when they entered the state (the spell start date). For example, when modelling the length of spells of unemployment insurance (UI) receipt, one might sample all the individuals who were in receipt of UI at a given date, and also find out when they first received UI (and other characteristics).

2. *Inflow sample* Data collection is based on a random sample of all persons entering the state of interest, and individuals are followed until some prespecified date (which might be common to all individuals), or until the spell ends. For example, when modelling the length of spells of receipt of unemployment insurance (UI), one might sample all the individuals who began a UI spell.

3. *Outflow sample* Data collection is based on a random sample of those leaving the state of interest, and one also determines when the spell began. For example, to continue our UI example, the sample would consist of individuals leaving UI recept.

4. *Population sample* Data collection is based on a general survey of the population (i.e. where sampling is not related to the process of interest), and respondents are asked about their current and/or previous spells of the type of interest (starting and ending dates).

Data may also be generated from combinations of these sample types. For example, the researcher may build a sample of spells by considering all spells that occurred between two dates, for example between 1 January and 1 June of a given year. Some spells will already be in progress at the beginning of the observation window (as in the stock sample case), whereas some will begin during the window (as in the inflow sample case).

The longitudinal data in these four types of sample may be collected from three main types of survey or database:

1. *Administrative records* For example, information about UI spells may be derived from the database used by the government to administer the benefit system. The administrative records may be the sole source of information about the individuals, or may be combined with a social survey that asks further questions of the persons of interest.

2. *Cross-section sample survey, with retrospective questions* In this case, respondents to a survey are asked to provide information about their spells

in the state of interest using retrospective recall methods. For example, when considering how long marriages last, analysts may use questions asking respondents whether they are currently married, or ever have been, and determining the dates of marriage and of divorce, separation, and widowhood. Similar sorts of methods are commonly used to collect information about personal histories of employment and jobs over the working life.

3. *Panel and cohort surveys, with prospective data collection* In this case, the longitudinal information is built from repeated interviews (or other sorts of observation) on the sample of interest at a number of different points in time. At each interview, respondents are typically asked about their current status, and changes since the previous interview, and associated dates.

Combinations of these survey instruments may be used. For example a panel survey may also include retrospective question modules to ask about respondents' experiences before the survey began. Administrative records containing longitudinal data may be matched into a sample survey, and so on.

The main lesson of this brief introduction to data collection methods is that, although each method provides spell data, the nature of the information about the spells differs, and this has important implications for how one should analyze the data. The rest of this section highlight the nature of the differences in information about spells. The first aspect concerns whether survival times are complete, censored or truncated. The second and related aspect concerns whether the analyst observes the precise dates at which spells are observed (or else survival times are only observed in intervals of time, i.e. grouped or banded) or, equivalently – at least from the analytic point of view – whether survival times are intrinsically discrete.

1.2.1 Censoring and truncation of survival time data

A survival time is *censored* if all that is known is that it began or ended within some particular interval of time, and thus the total spell length (from entry time until transition) is not known exactly. We may distinguish the following types of censoring:

- *Right censoring*: at the time of observation, the relevant event (transition out of the current state) had not yet occurred (the spell end date is unknown), and so the total length of time between entry to and exit from the state is unknown. Given entry at time 0 and observation at time t, we only know that the completed spell is of length $T > t$.

- *Left censoring*: the case when the start date of the spell was not observed, so again the exact length of the spell (whether completed or incomplete) is not known. Note that this is the definition of left censoring most commonly used by social scientists. (Be aware that biostatisticians typically

use a different definition: to them, left-censored data are those for which it is known that exit from the state occurred at some time before the observation date, but it is not known exactly when. See e.g. Klein and Moeschberger, 1997.)

By contrast, *truncated* survival time data are those for which there is a systematic exclusion of survival times from one's sample, and the sample selection effect depends on survival time itself. We may distinguish two types of truncation:

- *Left truncation*: the case when only those who have survived more than some minimum amount of time are included in the observation sample ('small' survival times – those below the threshold – are not observed). Left truncation is also known by other names: *delayed entry* and *stock sampling with follow-up*. The latter term is the most-commonly referred to by economists, reflecting the fact that data they use are often generated in this way. If one samples from the stock of persons in the relevant state at some time s, and interviews them some time later, then persons with short spells are systematically excluded. (Of all those who began a spell at time $r < s$, only those with relatively long spells survived long enough to be found in the stock at time s and thence available to be sampled.) Note that the spell start is assumed known in this case (cf. left censoring), but the subject's survival is only observed from some later date – hence 'delayed entry'.

- *Right truncation*: this is the case when only those persons who have experienced the exit event by some particular date are included in the sample, and so relatively 'long' survival times are systematically excluded. Right truncation occurs, for example, when a sample is drawn from the persons who exit from the state at a particular date (e.g. an *outflow sample* from the unemployment register).

The most commonly available survival time data sets contain a combination of survival times in which either (i) both entry and exit dates are observed (*completed spell data*), or (ii) entry dates are observed and exit dates are not observed exactly (*right censored incomplete spell data*). The ubiquity of such right censored data has meant that the term 'censoring' is often used as a shorthand description to refer to this case. We shall do so as well.

See Figure 2 for some examples of different types of spells. *** insert and add comments ***

We assume that the process that gives rise to censoring of survival times is *independent* of the survival time process. There is some latent failure time for person i given by $T_i^*$and some latent censoring time C_i^*, and what we observe is $T_i = \min\{T_i^*, C_i^*\}$. See the texts for more about the different types of censoring mechanisms that have been distinguished in the literature. If right-censoring is not independent – instead its determinants are correlated with the determinants of the transition process – then we need to model the two processes jointly.

An example is where censoring arises through non-random sample drop-out ('attrition').

1.2.2 Continuous versus discrete (or grouped) survival time data

So far we have implicitly assumed that the transition event of interest may occur at any particular instant in time; the stochastic process occurs in *continuous time*. Time is a continuum and, in principle, the length of an observed spell length can be measured using a non-negative real number (which may be fractional). Often this is derived from observations on spell start dates and either spell exit dates (complete spells) or last observation date (censored spells). Survival time data do not always come in this form, however, and for two reasons.

The first reason is that survival times have been grouped or banded into discrete intervals of time (e.g. numbers of months or years). In this case, spell lengths may be summarised using the set of positive integers (1, 2, 3, 4, and so on), and the *observations* on the transition process are summarized discretely rather than continuously. That is, although the underlying transition process may occur in continuous time, the data are not observed (or not provided) in that form. Biostatisticians typically refer to this situation as one of *interval censoring*, a natural description given the definitions used in the previous subsection. The occurence of tied survival times may be an indicator of interval censoring. Some continuous time models often (implicitly) assume that transitions can only occur at different times (at different instants along the time continuum), and so if there is a number of individuals in one's data set with the same survival time, one might ask whether the ties are genuine, or simply because survival times have been grouped at the observation or reporting stage.

The second reason for discrete time data is when the underlying transition process is an *intrinsically discrete* one. Consider, for example, a machine tool set up to carry out a specific cycle of tasks and this cycle takes a fixed amount of time. When modelling how long it takes for the machine to break down, it would be natural to model failure times in terms of the number of discrete cycles that the machine tool was in operation. Similarly when modelling fertility, and in particular the time from puberty to first birth, it might be more natural to measure time in terms of numbers of menstrual cycles rather than number of calendar months.

Since the same sorts of models can be applied to discrete time data regardless of the reason they were generated (as we shall see below), we shall mostly refer simply to discrete time models, and constrast these with continuous time models.

Thus the more important distinction is between discrete time data and continuous time data. Models for the latter are the most commonly available and most commonly applied, perhaps reflecting their origins in the bio-medical sciences. However discrete time data are relatively common in the social sciences. One of the themes of this lecture course is that one should use models that reflect the nature of the data available. For this reason, more attention is given to discrete time models than is typically common. For the same reason, I give

more explicit attention to how to estimate models using data sets containing left-truncated spells than do most texts.

1.2.3 Types of explanatory variables

There are two main types. Contrast, first, explanatory variables that describe

- the characteristics of the observation unit itself (e.g. a person's age, or a firm's size), versus
- the characteristics of the socio-economic environment of the observation unit (e.g. the unemployment rate of the area in which the person lives).

As far model specification is concerned, this distinction makes no difference. It may make a significant difference in practice, however, as the first type of variables are often directly available in the survey itself, whereas the second type often have to be collected separately and then matched in.

The second contrast is between explanatory variables that are

- *fixed* over time, whether time refers to calendar time or survival time within the current state, e.g. a person's sex; and
- *time-varying*, and distinguish between those that vary with survival time and those vary with calendar time.

The unemployment rate in the area in which a person lives may vary with calendar time (the business cycle), and this can induce a relationship with survival time but does not depend intrinsically on survival time itself. By contrast, social assistance benefit rates in Britain used to vary with the length of time that benefit had been received: Supplementary Benefit was paid at the short-term rate for spells up to 12 months long, and paid at a (higher) long-term rate for the 13th and subsequent months for spells lasting this long. (In addition some calendar time variation in the benefit generosity in real terms was induced by inflation, and by annual uprating of benefit amounts at the beginning of each financial year (April).)

Some books refer to *time-dependent* variables. These are either the same as the time-varying variables described above or, sometimes, variables for which changes over time can be written directly as a function of survival time. For example, given some personal characteristic summarized using variable X, and survival time t, such a time-dependent variable might be $X \log(t)$.

The distinction between fixed and time-varying covariates is relevant for both analytical and practical reasons. Having all explanatory variables fixed means that analytical methods and empirical estimation are more straightforward. With time-varying covariates, some model interpretations no longer hold. And from a practical point of view, one has to re-organise one's data set in order to incorporate them and estimate models. More about this 'episode splitting' later on.

1.3 Why are distinctive statistical methods used?

This section provides some motivation for the distinctive specialist methods that have been developed for survival analysis by considering why some of the methods that are commonly used elsewhere in economics and other quantitative social science disciplines cannot be applied in this context (at least in their standard form). More specifically, what is the problem with using either (1) Ordinary Least Squares (OLS) regressions of survival times, or with using (2) binary dependent variable regression models (e.g. logit, probit) with transition event occurrence as the dependent variable? Let us consider these in turn.

OLS cannot handle three aspects of survival time data very well:

- censoring (and truncation)
- time-varying covariates
- 'structural' modelling

1.3.1 Problems for OLS caused by right censoring

To illustrate the (right) censoring issue, let us suppose that the 'true' model is such that there is a single explanatory variable, X_i for each individual $i = 1, \ldots, n$, who has a true survival time of T_i^*. In addition, in the population, a higher X is associated with a shorter survival time. In the sample, we observe T_i where $T_i = T_i^*$ for observations with completed spells, and $T_i < T_i^*$ for right censored observations.

Suppose too that the incidence of censoring is higher at longer survival times relative to shorter survival times. (This does not necessarily conflict with the assumption of independence of the censoring and survival processes – it simply reflects the passage of time. The longer the observation period, the greater the proportion of spells for which events are observed.)

CHART TO INSERT

Data 'cloud': combinations of 'true' X_i, T_i^*

By OLS, we mean: regress T_i, or better still $\log T_i$ (noting that survival times are all non-negative and distributions of survival times are typically skewed), on X_i, fitting the linear relationship

$$\log(T_i) = a + bX_i + e_i \tag{1.1}$$

The OLS parameter estimates are the solution to $\min_{a,b} \sum_{i=1}^{n} (e_i)^2$. $\hat{a}$ is the vertical intercept; $\widehat{b}$ is the slope of the least squares line.

Case (a) Exclude censored cases altogether

Sample data cloud less dense everywhere but disproportionately at higher t

CHART TO INSERT

Case (b) Treat censored durations as if complete

CHART TO INSERT
Under-recording – especially at higher t
sample OLS line has wrong slope again

1.3.2 Time-varying covariates and OLS

How can one handle time-varying covariates, given that OLS has only a single dependent variable and there are multiple values for the covariate in question?

If one were to choose one value of a time-varying covariate at some particular time as 'representative', which one would one choose of the various possibilities? For example:

- the value at the time just before transition? But completed survival times vary across people, and what would one do about censored observations (no transition observed)?
- the value of the variable when the spell started, as this is the only definition that is consistently defined? But then much information is thrown away.

In sum, time-varying covariates require some special treatment in modelling.

1.3.3 'Structural' modelling and OLS

Our behavioural models of for example job search, marital search, etc., are framed in terms of decisions about whether to do something (and observed transitions reflect that choice). I.e. models are not formulated in terms of completed spell lengths. Perhaps, then, we should model transitions directly.

1.3.4 Why not use binary dependent variable models rather than OLS?

Given the above problems, especially the censoring one, one might ask whether one could use instead a binary dependent regression model (e.g. logit, probit)? I.e. one could get round the censoring issue (and the structural modelling one), by simply modelling whether or not someone made a transition or not. (Observations with a transition would have a '1' for the dependent variable; censored observations would have a '0'.) However, this strategy is also potentially problematic:

- it takes no account of the differences in time in which each person is at risk of experiencing the event. One could get around this by considering whether a transition occurred within some pre-specified interval of time (e.g. 12 months since the spell began), but ...

- one still loses a large amount of information, in particular about when someone left if she or he did so.

Cross-tabulations of (banded) survival times against some categorical/categorised variable cannot be used for inference about the relationship between survival time and that variable, for the same sorts of reasons. (Crosstabulations of a dependent variable against each explanatory variable is often used with other sorts of data to explore relationships.) In particular, the problems include:

- the dependent variable is mis-measured and censoring is not accounted for;
- time-varying explanatory variables cannot be handled easily (current values may be misleading)

1.4 Outline of the book

The preceding sections have argued that, for survival analysis, we need methods that directly account for the sequential nature of the data, and are able to handle censoring and incorporate time-varying covariates. The solution is to model survival times indirectly, via the so-called 'hazard rate', which is a concept related to chances of making a transition out of the current state at each instant (or time period) conditional on survival up to that point. The rest of this book elaborates this strategy, considering both continuous and discrete time models.

The hazard rate is defined more formally in Chapter 2. I also draw attention to the intimate connections between the hazard, survivor, failure, and density functions. Chapter 3 discusses functional forms for the hazard rate. I set out some of the most commonly-used specifications, and explain how models may be classified into two main types: proportional hazards or accelerated failure time models. I also show, for a selection of models, what the hazard function specification implies about the distribution of survival times (including the median and mean spell lengths), and about the relationship between differences in survival times and differences in characteristics (summarised by differences in values of explanatory variables). Survival analysis is not simply about estimating model parameters, but also interpreting them and drawing out their implications to the fullest extent.

In the subsequent chapters, we move from concepts to estimation. The aim is to indicate the principles behind the methods used, rather than provide details (or proofs) about the statistical properties of estimators, or the numerical analysis methods required to actually derive them.

Chapter 4 discusses Kaplan-Meier (product-limit) and Lifetable estimators of the survivor and hazard functions. These are designed to fit functions for the sample as a whole, or for separate subgroups; they are not multivariate regression models. Multivariate regression models in which differences in characteristics are incorporated via differences in covariate values are the subject of Chapters 5–9. The problems with using OLS to model survival time data

are shown to be resolved if one uses instead estimation based on the maximum likelihood principle or the partial likelihood principle.

A continuing theme is that estimation has to take account of the sampling scheme that generates the observed survival time data. The chapters indicate how the likelihoods underlying estimation differ in each of the leading sampling schemes: random samples of spells (with or without right censoring), and left- and right-truncated spell data. We also examine how to incorporate time-varying covariates using 'episode-splitting'. Chapters 5 and 6 discuss continuous time and discrete time regression models respectively, estimated using maximum likelihood. Chapter 7 introduces Cox's semi-parametric proportional hazard model for continuous time data, estimated using partial likelihood.

The remainder of the book discusses a selection of additional topics. Chapter 8 addresses the subject of unobserved heterogeneity (otherwise known as 'frailty'). In the models considered in the earlier chapters, it is implicitly assumed that all relevant differences between individuals can be summarised by the observed explanatory variables. But what if there are unobserved or unobservable differences? The chapter discusses the impact that unobserved heterogeneity may have on estimates of regression coefficients and duration dependence, and outlines the methods that have been proposed to take account of the problem. Chapter 9 considers estimation of competing risks models. Earlier chapters consider models for exit to a single destination state; this chapter shows how one can model transitions to a number of mutually-exclusive destination states. Chapter 10 [not yet written!] discusses repeated spell data (the rest of the book assumes that the available data contain a single spell per subject).

A list of references and appendices complete the book.

Chapter 2

Basic concepts: the hazard rate and survivor function

In this chapter, we define key concepts of survival analysis, namely the hazard function and the survivor function, and show that there is a one-to-one relationship between them. For now, we shall ignore differences in hazard rates, and so on, across individuals in order to focus on how they vary with survival time. How to incorporate individual heterogeneity is discussed in the next chapter.

2.1 Continuous time

The length of a spell for a subject (person, firm, etc.) is a realisation of a continuous random variable T with a *cumulative distribution function* (cdf), $F(t)$, and *probability density function* (pdf), $f(t)$. $F(t)$ is also known in the survival analysis literature as the *failure function*. The *survivor function* is $S(t) \equiv 1 - F(t)$; t is the elapsed time since entry to the state at time 0.

Insert chart of pdf, showing cdf as area under curve

Failure function (cdf)

$$\Pr(T \leq t) = F(t) \tag{2.1}$$

which implies, for the Survivor function:

$$\Pr(T > t) = 1 - F(t) \equiv S(t). \tag{2.2}$$

Observe that some authors use $\overline{F}(t)$ to refer to the survivor function. I use $S(t)$ throughout.

Insert chart of pdf in terms of cdf

The pdf is the slope of the cdf (Failure) function:

$$f(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T \leq t + \Delta t)}{\Delta t} = \frac{\partial F(t)}{\partial t} = -\frac{\partial S(t)}{\partial t} \tag{2.3}$$

where Δt is a very small ('infinitesimal') interval of time. The $f(t)\Delta t$ is akin to the unconditional probability of having a spell of length exactly t, i.e. leaving state in tiny interval of time $[t, t + \Delta t]$.

The survivor function $S(t)$ and the Failure function $F(t)$ are each probabilities, and therefore inherit the properties of probabilities. In particular, observe that the survivor function lies between zero and one, and is a strictly decreasing function of t. The survivor function is equal to one at the start of the spell ($t = 0$) and is zero at infinity.

$$\begin{aligned} 0 &\leq S(t) \leq 1 && (2.4)\\ S(0) &= 1 && (2.5)\\ \lim_{t \to \infty} S(t) &= 0 && (2.6)\\ \frac{\partial S}{\partial t} &< 0 && (2.7)\\ \frac{\partial^2 S}{\partial t^2} &\gtrless 0. && (2.8) \end{aligned}$$

The density function is non-negative

$$f(t) \geq 0 \tag{2.9}$$

but may be greater than one in value (the density function does not summarize probabilities).

2.1.1 The hazard rate

The hazard rate is a difficult concept to grasp, some people find. Let us begin with its definition, and then return to interpretation. The continuous time hazard rate, $\theta(t)$, is defined as:

$$\theta(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}. \tag{2.10}$$

Suppose that we let $\Pr(A)$ be the probability of leaving the state in the tiny interval of time between t and $t+\Delta t$, and $\Pr(B)$ be the probability of survival up to time t, then the probability of leaving in the interval $(t, t+\Delta t]$, conditional on survival up to time t, may be derived from the rules of conditional probability:

$$\Pr(A|B) = \Pr(A \cap B)/\Pr(B) = \Pr(B|A)\Pr(A)/\Pr(B) = \Pr(A)/\Pr(B), \tag{2.11}$$

since $\Pr(B|A) = 1$. But $\Pr(A)/\Pr(B) = f(t)\Delta t/S(t)$. This expression is closely related to the expression that defines the hazard rate: compare it with (2.10):

$$\theta(t)\Delta t = \frac{f(t)\Delta t}{S(t)}. \tag{2.12}$$

Thus $\theta(t)\Delta t$, for tiny Δt, is akin to the conditional probability of having a spell length of exactly t, conditional on survival up to time t. It should be stressed, however, that the hazard rate is *not* a probability, as it refers to the exact time t and not the tiny interval thereafter. (Later we shall consider discrete time survival time data. We shall see that the discrete time hazard *is* a (conditional) probability.) The only restriction on the hazard rate, and implied by the properties of $f(t)$ and $S(t)$, is that:

$$\theta(t) \geq 0.$$

That is, $\theta(t)$ may be greater than one, in the same way that the probability density $f(t)$ may be greater than one.

The probability density function $f(t)$ summarizes the concentration of spell lengths (exit times) at each instant of time along the time axis. The hazard function summarizes the same concentration at each point of time, but conditions the expression on survival in the state up to that instant, and so can be thought of as summarizing the instantaneous *transition intensity*.

To understand the conditioning underpinning the hazard rate further, consider an unemployment example. Contrast (i) conditional probability $\theta(12)\Delta t$, for tiny Δt,and (ii) unconditional probability $f(12)\Delta t$. Expression (i) denotes the probability of (re-)employment in the interval $(12, 12 + \Delta t)$ for a person who has already been unemployed 12 months, whereas (ii) is the probability for an entrant to unemployment of staying unemployed for 12 months and leaving in interval $(12, 12 + \Delta t)$. Alternatively, take a longevity example. Contrast the unconditional probability of dying at age 12 (for all persons of a given birth cohort), and probability of dying at age 12, given survival up to that age.

Economists may recognise the expression for the hazard as being of the same form as the "inverse Mills ratio" that used when accounting for sample selection biases.

2.1.2 Key relationships between hazard and survivor functions

I now show that there is a one-to-one relationship between a specification for the hazard rate and a specification for the survivor function. I.e. whatever functional form is chosen for $\theta(t)$, one can derive $S(t)$ and $F(t)$ from it, and also $f(t)$ and $H(t)$. Indeed, in principle, one can start from any one of these different characterisations of the distribution, and derive the others from it. In practice,

one typically starts from considerations of the shape of the hazard rate.

$$\begin{aligned}
\theta(t) &= \frac{f(t)}{1-F(t)} && (2.13)\\
&= \frac{-\partial[1-F(t)]/\partial t}{1-F(t)} && (2.14)\\
&= \frac{\partial\{-\ln[1-F(t)]\}}{\partial t} && (2.15)\\
&= \frac{\partial\{-\ln[S(t)]\}}{\partial t} && (2.16)
\end{aligned}$$

using the fact that $\partial \ln[g(x)]/\partial x = g'(x)/g(x)$ and $S(t) = 1 - F(t)$. Now integrate both sides:

$$\int_0^t \theta(u)\mathrm{d}u = -\ln[1-F(t)] \quad |_0^t. \tag{2.17}$$

But $F(0) = 0$ and $\ln(1) = 0$, so

$$\begin{aligned}
\ln[1-F(t)] &= \ln[S(t)] = -\int_0^t \theta(u)du, \text{ i.e.} && (2.18)\\
S(t) &= \exp\left(-\int_0^t \theta(u)du\right) && (2.19)\\
S(t) &= \exp[-H(t)] && (2.20)
\end{aligned}$$

where the *integrated hazard function*, $H(t)$, is

$$\begin{aligned}
H(t) &\equiv \int_0^t \theta(u)du && (2.21)\\
&= -\ln[S(t)]. && (2.22)
\end{aligned}$$

Observe that

$$\begin{aligned}
H(t) &\geq 0\\
\frac{\partial H(t)}{\partial t} &= \theta(t).
\end{aligned}$$

We have therefore demonstrated the one-to-one relationships between the various concepts. In Chapter 3, we take a number of common specifications for the hazard rate and derive the corresponding survivor functions, integrated hazard functions, and density functions.

2.2 Discrete time

As discussed earlier, discrete survival time data may arise because either (i) the time scale is intrinsically discrete, or (ii) survival occurs in continuous time but spell lengths are observed only in intervals ('grouped' or 'banded' data). Let us consider the latter case first.

2.2.1 Survival in continuous time but spell lengths are interval-censored

*Insert chart (time scale) *

We suppose that the time axis may be partitioned into a number of contiguous non-overlapping ('disjoint') intervals where the interval boundaries are the dates $a_0 = 0, a_1, a_2, a_3, ..., a_k$.The intervals themselves are described by

$$[0 = a_0, a_1], (a_1, a_2], (a_2, a_3], ..., (a_{k-1}, a_k = \infty]. \tag{2.23}$$

This definition supposes that interval $(a_{j-1}, a_j]$ begins at the instant after the date marking the beginning and end of the interval (a_{j-1}). The time indexing the end of the interval (a_j) is included in the interval.[1] Observe that $a_1, a_2, a_3, ...,$ are *dates* (points in time), and the intervals need not be of equal length (though we will later suppose for convenience that they are).

The value of the survivor function at the time demarcating the start of the jth interval is

$$\Pr(T > a_{j-1}) = 1 - F(a_{j-1}) = S(a_{j-1}) \tag{2.24}$$

where $F(.)$ is the Failure function defined earlier. The value of the survivor function at the end of the jth interval is

$$\Pr(T > a_j) = 1 - F(a_j) = \overline{F}(a_j) = S(a_j). \tag{2.25}$$

The probability of exit *within* the jth interval is

$$\Pr(a_{j-1} < T \leq a_j) = F(a_j) - F(a_{j-1}) = S(a_{j-1}) - S(a_j). \tag{2.26}$$

The *interval hazard rate*, $h(a_j)$, also known as the discrete hazard rate, is the probability of exit in the interval $(a_{j-1}, a_j]$, and defined as:

$$\begin{aligned} h(a_j) &= \Pr(a_{j-1} < T \leq a_j | T > a_{j-1}) && (2.27)\\ &= \frac{\Pr(a_{j-1} < T \leq a_j)}{\Pr(T > a_{j-1})} && (2.28)\\ &= \frac{S(a_{j-1}) - S(a_j)}{S(a_{j-1})} && (2.29)\\ &= 1 - \frac{S(a_j)}{S(a_{j-1})} && (2.30) \end{aligned}$$

Note that the interval time hazard is a (conditional) probability, and so

$$0 \leq h(a_j) \leq 1. \tag{2.31}$$

[1] One could, alternatively, define the intervals as $[a_{j-1}, a_j)$, for $j = 1, ..., k$. The choice is largely irrelevant in development of the theory, though it can matter in practice, because different software packages use different conventions and this can lead to different results. (The differences arise when one splits spells into a sequence of sub-episodes – 'episode splitting'.) I have used the definition that is consistent with Stata's definition of intervals. The TDA package uses the other convention.

In this respect, the discrete hazard rate differs from the continuous time hazard rate, for which $\theta(u) \geq 0$ and may be greater than one.

Although the definition used so far refers to intervals that may, in principle, be of different lengths, in practice it is typically assumed that intervals are of equal *unit length*, for example a 'week' or a 'month'). In this case, the time intervals can be indexed using the positive integers. Interval $(a_{j-1}, a_j]$ may be relabelled $(a_j - 1, a_j]$, for $a_j = 1, 2, 3, 4, ...$, and we may refer to this as the jth interval, and to the interval hazard rate as $h(j)$ rather than $h(a_j)$.

We shall assume that intervals are of unit length from now on, unless otherwise stated.

The probability of survival until the *end* of interval j is the product of probabilities of not experiencing event in each of the intervals up to and including the current one. For example, S_3 = (probability of survival through interval 1) $\times$ (probability of survival through interval 2, given survival through interval 1) $\times$ (probability of survival through interval 3, given survival through interval 2). Hence, more generally, we have:

$$\begin{aligned} S(j) &\equiv S_j = (1-h_1)(1-h_2).....(1-h_{j-1})(1-h_j) && (2.32) \\ &= \prod_{k=1}^{j}(1-h_k) && (2.33) \end{aligned}$$

$S(j)$ refers to a *discrete time survivor function*, written in terms of interval hazard rates. We shall reserve the notation $S(a_j)$ or, more generally $S(t)$, to refer to the continuous time survivor function, indexed by a date – an instant of time – rather than an interval of time. (Of course the two functions imply exactly the same value since, by construction, the end of the jth interval is date a_j.) In the special case in which the hazard rate is constant over time (in which case survival times follow a Geometric distribution), i.e. $h_j = h$, all j, then

$$\begin{aligned} S(j) &= (1-h)^j && (2.34) \\ \log[\ S(j)] &= j\log(1-h). && (2.35) \end{aligned}$$

The discrete time failure function, $F(j)$, is

$$\begin{aligned} F(j) &\equiv F_j = 1 - S(j) && (2.36) \\ &= 1 - \prod_{k=1}^{j}(1-h_k). && (2.37) \end{aligned}$$

The discrete time density function for the interval-censored case, $f(j)$, is the probability of exit within the jth interval:

$$\begin{aligned} f(j) &= \Pr(a_{j-1} < T \leq a_j) && (2.38) \\ &= S(j-1) - S(j) \\ &= \frac{S(j)}{1-h_j} - S(j) \\ &= \left(\frac{1}{1-h_j} - 1\right) S(j) \\ &= \frac{h_j}{1-h_j} \prod_{k=1}^{j} (1-h_k). && (2.39) \end{aligned}$$

Thus the discrete density is the probability of surviving up to the end of interval $j-1$, multiplied by the probability of exiting in the j^{th} interval. Expression (2.39) is used extensively in later chapters when deriving expressions for sample likelihoods. Observe that

$$0 \leq f(j) \leq 1. \tag{2.40}$$

2.2.2 The discrete time hazard when time is intrinsically discrete

In the case in which survival times are instrinsically discrete, survival time T is now a discrete random variable with probabilities

$$f(j) \equiv f_j = \Pr(T = j) \tag{2.41}$$

where $j \in \{1, 2, 3, \ldots\}$, the set of positive integers. Note that j now indexes 'cycles' rather than intervals of equal length. But we can apply the same notation; we index survival times using the set of positive integers in both cases. The discrete time survivor function for cycle j, showing the probability of survival for j cycles, is given by:

$$S(j) = \Pr(T \geq j) = \sum_{k=j}^{\infty} f_k. \tag{2.42}$$

The discrete time hazard at j, $h(j)$, is the conditional probability of the event at j (with conditioning on survival until completion of the cycle immediately before the cycle at which the event occurs) is:

$$\begin{aligned} h(j) &= \Pr(T = j | T \geq j) && (2.43) \\ &= \frac{f(j)}{S(j-1)} && (2.44) \end{aligned}$$

It is more illuminating to write the discrete time survivor function analogously to the expression for the equal-length interval case discussed above (for survival

to the end of interval j), i.e.:

$$S_j = S(j) = (1-h_1)(1-h_2).....(1-h_{j-1})(1-h_j) \quad (2.45)$$

$$= \prod_{k=1}^{j}(1-h_k). \quad (2.46)$$

The discrete time failure function is:

$$F_j = F(j) = 1 - S(j) \quad (2.47)$$

$$= 1 - \prod_{k=1}^{j}(1-h_k). \quad (2.48)$$

Observe that the discrete time density function can also be written as in the interval-censored case, i.e.:

$$f(j) = h_j S_{j-1} = \frac{h_j}{1-h_j} S_j. \quad (2.49)$$

2.2.3 The link between the continuous time and discrete time cases

In the discrete time case, we have from (2.45) that:

$$\log S(j) = \sum_{k=1}^{j} \log(1-h_k). \quad (2.50)$$

For 'small' h_k, a first-order Taylor series approximation may be used to show that

$$\log(1-h_k) \approx -h_k \quad (2.51)$$

which implies, in turn, that

$$\log S(j) \approx -\sum_{k=1}^{j} h_k. \quad (2.52)$$

Now contrast this expression with that for the continuous time case, and note the parallels between the summation over discrete hazard rates and the integration over continuous time hazard rates:

$$\log S(t) = -H(t) = -\int_0^t \theta(u)du. \quad (2.53)$$

As h_k becomes smaller and smaller, the closer that discrete time hazard h_j is to the continuous time hazard $\theta(t)$ and, correspondingly, the discrete time survivor function tends to the continuous time one.

2.3 Choosing a specification for the hazard rate

The empirical analyst with survival time data to hand has choices to make before analyzing them. First, should the survival times be treated as observations on a continuous random variable, observations on a continuous random variable which is grouped (interval censoring), or observations on an intrinsically discrete random variable? Second, conditional on that choice, what is the shape of the all-important relationship between the hazard rate and survival time? Let us consider these questions in turn.

2.3.1 Continuous or discrete survival time data?

The answer to this question may often be obvious, and clear from consideration of both the underlying behavioural process generating survival times, and the process by which the data were recorded. Intrinsically discrete survival times are rare in the social sciences. The vast majority of the behavioural processes that social scientists study occur in continuous time, but it is common for the data summarizing spell lengths to be recorded in grouped form. Indeed virtually all data are grouped (even with survival times recorded in units as small as days or hours).

A key issue, then, is the length of the intervals used for grouping relative to the typical spell length: the smaller the ratio of the former to the latter, the more appropriate it is to use a continuous time specification.

If one has information about the day, month, and year in which a spell began, and also the day, month, and year, at which subjects were last observed – so survival times are measured in days – and the typical spell length is several months or years, then it is reasonable to treat survival times as observations on a continuous random variable (not grouped). But if spells length are typically only a few days long, then recording them in units of days implies substantial grouping. It would then make sense to use a specification that accounted for the interval censoring. A related issue concerns 'tied' survival times – more than one individual in the data set with the same recorded survival time. A relatively high prevalence of ties may indicate that the banding of survival times should be taken into account when choosing the specification. For some analysis of the effects of grouping, see Bergström and Edin (1992), Petersen (1991), and Petersen and Koput (1992).

Historically, many of the methods developed for analysis of survival time data assumed that the data set contained observations on a continuous random variable (and arose in applications where this assumption was reasonable). Application of these methods to social science data, often interval-censored, was not necessarily appropriate. Today, this is much less of a problem. Methods for handling interval-censored data (or intrinsically discrete data) are increasingly available, and one of the aims of this book is to promulgate them.

2.3.2 The relationship between the hazard and survival time

The only restriction imposed so far on the continuous time hazard rate is $\theta(t) \geq 0$, and on the discrete time hazard rate is $0 \leq h(j) \leq 1$. Nothing has been assumed about the pattern of duration dependence – how the hazard varies with survival times. The following desiderata suggest themselves:

- A shape that is empirically relevant (or suggested by theoretical models). This is likely to differ between applications – the functional form describing human mortality is likely to differ from that for transitions out of unemployment transitions, for failure of machine tools, and so on.
- A shape that has convenient mathematical properties e.g. closed form expressions for $\theta(t)$, and $S(t)$ and tractable expressions for summary statistics of survival time distributions such as the mean and median survival time.

2.3.3 What guidance from economics?

As an illustration of the first point, consider the extent to which economic theory might provides suggestions for what the shape of the hazard rate is like. Consider a two-state labour market, where the two states are (1) employment, and (2) unemployment. Hence the only way to leave unemployment is by becoming employed. To leave unemployment requires that an unemployed person both receives a job offer, and that that offer is acceptable. (The job offer probability is conventionally considered to be under the choice of firms, and the acceptance probability dependent on the choice of workers.) For a given worker, we may write the unemployment exit hazard rate $\theta(t)$ as the product of the job offer hazard $\xi(t)$ and the job acceptance hazard $A(t)$:

$$\theta(t) = \xi(t)A(t). \tag{2.54}$$

A simple structural model

In a simple *job search* framework, the unemployed person searches across the distribution of wage offers, and the optional policy is to adopt a reservation wage r, and accept a job offer with associated wage w only if $w \geq r$. Hence,

$$\theta(t) = \xi(t)\left[1 - W(t)\right] \tag{2.55}$$

where $W(t)$ is the cdf of the wage offer distribution facing the worker. How the re-employment hazard varies with duration thus depends on:

1. How the reservation wage varies with duration of unemployment. (In an infinite horizon world one would expect r to be constant; in a finite horizon world, one would expect r to decline with the duration of unemployment.)

2. How the job offer hazard ξ varies with duration of unemployment. (It is unclear what to expect.)

In sum, the simple search model provides some quite strong restrictions on the hazard rate (if influences via ξ are negligible).

Reduced form approach

An alternative interpretation is that the simple search model (or even more sophisticated variants) place too many restrictions on the hazard function, i.e. the restrictions may be consequence of the simplifying assumptions used to make the model tractable, rather than intrinsic. Hence one may want to use a reduced form approach, and write the hazard rate more generally as

$$\theta(t) = \theta\left(X\left(t, s\right), t\right), \tag{2.56}$$

where X is a vector of personal characteristics that may vary with unemployment duration (t) or with calendar time (s). That is we allow, in a more ad hoc way, for the fact that:

1. unemployment benefits may vary with duration t, and maybe also calendar time s (because of policy changes, for example); and
2. local labour market conditions may vary with calendar time (s); and
3. θ may also vary directly with survival time, t.

Examples of this include

- Employers screening unemployed applicants on the basis of how long each applicant has been unemployed, for example rejecting the longer-term unemployed) : $\partial\xi/\partial t < 0$.
- The reservation wage falling with unemployment duration: $\partial A/\partial t > 0$ (a resource effect);
- Discouragement (or a 'welfare culture' or 'benefit dependence' effect) may set in as the unemployment spell lengthens, leading to decline in search intensity: $\partial\xi/\partial t < 0$.
- Time limits on eligibility to Unemployment Insurance (UI) may lead to a benefit exhaustion effect, with the re-employment hazard (θ) rising as the time limit approaches.

Observe the variety of potential influences. Moreover, some of the influences mentioned would imply that the hazard rises with unemployment duration, whereas others imply that the hazard declines with duration. The actual shape of the hazard will reflect a mixture of these effects. This suggests that it is important not to pre-impose particular shape on the hazard function. What

one needs to do is strike a balance between what a theoretical (albeit simplified) model might suggest, and flexibility in specification and ease of estimation. This will improve model fit, though of course problems of interpretation may remain. (Without the structural model, one cannot identify which influence has which effect so easily.)

Although the examples above referred to modelling of unemployment duration, much the same issues for model selection are likely to arise in other contexts.

Chapter 3

Functional forms for the hazard rate

3.1 Introduction and overview: a taxonomy

The last chapter suggested that there is no single shape for the hazard rate that is appropriate in all contexts. In this chapter we review the functional forms for the hazard rate that have been most commonly used in the literature. What this means in terms of survival, density and integrated hazard functions is examined in the following chapter.

We begin with an overview and taxonomy, and then consider continuous time and discrete time specifications separately. See Table 3.1. The entries in the table in the *emphasized* font are the ones that we shall focus on. One further distinction between the specifications, discussed later in the chapter, refers to their *interpretation* – whether the models can be described as

- proportional hazards models (PH), or
- accelerated failure time models (AFT).

Think of the PH and AFT classifications as describing whole families of survival time distributions, where each member of a family shares a common set of features and properties.

Note that although we specified hazard functions simply as a function of survival time in the last chapter, now we add an additional dimension to the specification – allowing the hazard rate to vary between individuals, depending on their characteristics.

Continuous time parametric	**Continuous time semi-parametric**	**Discrete time**
Exponential	Piece-wise constant Exponential	*Logistic*
Weibull	*'Cox' model*	*Complementary log-log*
Log-logistic		
Lognormal		
Gompertz		
Generalized Gamma		

Table 3.1: Functional forms for the hazard rate: examples

3.2 Continuous time specifications

In the previous chapter, we referred to the continuous time hazard rate, $\theta(t)$ and ignored any potential differences in hazard rates between individuals. Now we remedy that. We suppose that the characteristics of a given individual may be summarized by a vector of variables X and, correspondingly, now refer to the hazard function $\theta(t, X)$ and survivor function $S(t, X)$, integrated hazard function $H(t, X)$, and so on.

The way is which heterogeneity is incorporated is as follows. We define a linear combination of the characteristics:

$$\beta' X \equiv \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_K X_K. \tag{3.1}$$

There are K variables observed for each person, and the βs are parameters, later to be estimated. Observe that the linear index $\beta' X$, when evaluated, is simply a single number for each individual. For the moment, suppose that the values of the Xs do not vary with survival or calendar time, i.e. there are no time-varying covariates.

3.2.1 Weibull model and Exponential model

The Weibull model is specified as:

$$\begin{aligned} \theta(t, X) &= \alpha t^{\alpha-1} \exp\left(\beta' X\right) && (3.2) \\ &= \alpha t^{\alpha-1} \lambda && (3.3) \end{aligned}$$

where $\lambda \equiv \exp\left(\beta' X\right)$, $\alpha > 0$, and $\exp(.)$ is the exponential function. The hazard rate either rises monotonically with time ($\alpha > 1$), falls monotonically with time ($\alpha < 1$), or is constant. The last case, $\alpha = 1$, is the special case of the Weibull model known as the *Exponential* model. For a given value of α, larger values of λ imply a larger hazard rate at each survival time. The α is the *shape parameter*.

Beware: different normalisations and notations are used by different authors. For example, you may see the Weibull hazard written as $\theta(t, X) =$

$(1/\sigma)t^{(1/\sigma)-1}\lambda$ where $\sigma \equiv 1/\alpha$. (The reason for this will become clearer in the next section.) Cox and Oakes (1984) characterise the Weibull model as $\theta(t, X) = \varkappa\lambda(\lambda t)^{\varkappa-1} = \varkappa\lambda^{\varkappa}t^{\varkappa-1}$.

Insert charts showing how hazard varies with (i) variations in α with fixed λ; (ii) variations in λ with fixed α.

3.2.2 Gompertz model

The Gompertz model has a hazard function given by:

$$\theta(t, X) = \lambda \exp(\gamma t) \tag{3.4}$$

and so the log of the hazard is a linear function of survival time:

$$\log \theta(t, X) = \beta' X + \gamma t. \tag{3.5}$$

If the shape parameter $\gamma > 0$, then the hazard is monotonically increasing; if $\gamma = 0$, it is constant; and if $\gamma < 0$, the hazard declines monotonically. *Insert charts* showing how hazard varies with (i) variations in γ with fixed λ; (ii) variations in λ with fixed γ.

3.2.3 Log-logistic Model

To specify the next model, we use the parameterisation $\psi \equiv \exp(-\beta^{*\prime} X)$, where

$$\beta^{*\prime} X \equiv \beta_0^* + \beta_1^* X_1 + \beta_2^* X_2 + \beta_3^* X_3 + \ldots + \beta_K^* X_K. \tag{3.6}$$

The reason for using a parameterization based on ψ and β^*, rather than on λ and β, as for the Weibull model, will become apparent later, when we compare proportional hazard and accelerated failure time models.

The Log-logistic model hazard rate is:

$$\theta(t, X) = \frac{\psi^{\frac{1}{\gamma}} t^{\left(\frac{1}{\gamma}-1\right)}}{\gamma\left[1 + (\psi t)^{\frac{1}{\gamma}}\right]} \tag{3.7}$$

where *shape parameter* $\gamma > 0$. An alternative representation is:

$$\theta(t, X) = \frac{\varphi\psi^{\varphi} t^{\varphi-1}}{1 + (\psi t)^{\varphi}} \tag{3.8}$$

where $\varphi \equiv 1/\gamma$. Klein and Moeschberger (1997) characterise the log-logistic model hazard function as $\theta(t, X) = \varphi\rho t^{\varphi-1}/(1 + \rho t^{\varphi})$, which is equivalent to the expression above with $\rho = \psi^{\varphi}$.

Insert chart of hazard rate against time.

Observe that the hazard rate is monotonically decreasing with survival time for $\gamma \geq 1$ (i.e. $\varphi \leq 1$). If $\gamma < 1$ (i.e. $\varphi > 1$), then the hazard first rises with time and then falls monotonically.

3.2.4 Lognormal model

This model has hazard rate

$$\theta(t,X) = \frac{\frac{1}{t\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left\{\frac{\ln(t)-\mu}{\sigma}\right\}^2\right]}{1-\Phi\left(\frac{\ln(t)-\mu}{\sigma}\right)} \tag{3.9}$$

where $\Phi(.)$ is the standard Normal cumulative distribution function and characteristics are incorporated with the parameterisation $\mu = \beta^{*\prime}X$. The hazard rate is similar to that for the Log-logistic model for the case $\gamma < 1$ (i.e. first rising and then declining).

3.2.5 Generalised Gamma model

This model has a rather complicated specification involving two shape parameters. Let us label them $\varkappa$ and σ, as in the Stata manual: they are the shape and scale parameters respectively. The hazard function is quite flexible in shape, even including the possibility of a U shaped or so-called 'bath-tub' shaped hazard (commonly cited as a plausible description for the hazard of human mortality looking at the lifetime as a whole). The Generalised Gamma incorporates several of the other models as special cases. If $\varkappa = 1$, we have the Weibull model; if $\varkappa = 1$, $\sigma = 1$, we have the Exponential model. With $\varkappa = 0$, the Lognormal model results. And if $\varkappa = \sigma$, then one has the standard Gamma distribution. These relationships mean that the generalised Gamma is useful for testing model specification: by estimating this general model, one can use a Wald (or likelihood ratio) test to investigate whether one of the nested models provides a satisfactory fit to the data.

Insert chart of hazard rate against time.

Note that a number of other parametric models have been proposed. For example, there is the generalized F distribution (see Kalbfleisch and Prentice 1980).

3.2.6 Proportional Hazards (PH) models

Let us now return to the general case of hazard rate $\theta(t,X)$, i.e. the hazard rate at survival time t for a person with fixed covariates summarised by the vector X.

The PH specification

Proportional hazards models are also known as 'multiplicative hazard' models, or 'log relative hazard' models for reasons that will become apparent shortly. The models are characterised by their satisfying a separability assumption:

$$\theta(t,X) = \theta_0(t)\exp(\beta'X) = \theta_0(t)\lambda \tag{3.10}$$

where

- $\theta_0(t)$: the 'baseline hazard' function, which depends on t (but not X). It summarizes the pattern of 'duration dependence', assumed to be common to all persons;
- $\lambda = \exp(\beta' X)$: a person-specific non-negative function of covariates X (which does not depend on t, by construction), which scales the baseline hazard function common to all persons. In principle, any non-negative function might be used to summarise the effects of differences in personal characteristics, but since exp(.) is the function that is virtually always used, we will use it.

Note that I have used one particular convention for the definition of the baseline hazard function. An alternative characterization writes the proportional hazard model as

$$\theta(t, X) = \theta_0^*(t)\,\lambda^* \tag{3.11}$$

where

$$\theta_0^*(t) = \theta_0(t)\exp(\beta_0) \text{ and } \lambda^* = \exp(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_K X_K). \tag{3.12}$$

The rationale for this alternative representation is that the intercept term (β_0) is common to all individuals, and is therefore not included in the term summarizing individual heterogeneity. By contrast, the first representation treats the intercept as a regression parameter like the other elements of β. In most cases, it does not matter which characterization is used.

Interpretation

The PH property implies that absolute differences in X imply proportionate differences in the hazard at each t. For some $t = \bar{t}$, and for two persons i and j with vectors of characteristics X_i and X_j,

$$\frac{\theta(\bar{t}, X_i)}{\theta(\bar{t}, X_j)} = \exp\left(\beta' X_i - \beta' X_j\right) = \exp\left[\beta'(X_i - X_j)\right]. \tag{3.13}$$

We can also write (3.13) in 'log relative hazard' form:

$$\log\left[\frac{\theta(\bar{t}, X_i)}{\theta(\bar{t}, X_j)}\right] = \beta'(X_i - X_j). \tag{3.14}$$

Observe that the right-hand side of these expressions does not depend on survival time (by assumption the covariates are not time dependent), i.e. the proportional difference result in hazards is constant.

If persons i and j are identical on all but the kth characteristic, i.e. $X_{im} = X_{jm}$ for all $m \in \{1, ..., K \backslash k\}$, then

$$\frac{\theta\left(\bar{t}, X_i\right)}{\theta\left(\bar{t}, X_j\right)} = \exp[\beta_k(X_{ik} - X_{jk})]. \tag{3.15}$$

If, in addition, $X_{ik} - X_{jk} = 1$, i.e. there is a *one unit change in* X_k, ceteris paribus, then

$$\frac{\theta\left(\bar{t}, X_i\right)}{\theta\left(\bar{t}, X_j\right)} = \exp(\beta_k). \tag{3.16}$$

The right hand side of this expression is known as the *hazard ratio.* It also shows the proportionate change in the hazard given a change in a dummy variable covariate from zero to one or, more precisely a change from $X_{ik} = 0$ to $X_{jk} = 1$, with all other covariates held fixed.

There is a nice interpretation of the regression coefficients in PH models. The coefficient on the kth covariate X, β_k, has the property

$$\beta_k = \partial \log \theta(t, X)/\partial X_k \tag{3.17}$$

which tells us that in a PH model, each regression coefficient summarises the proportional effect on the hazard of absolute changes in the corresponding covariate. This effect does not vary with survival time.

Alternatively, we can relate β_k to the *elasticity* of the hazard with respect to X_k. Recall that an elasticity summarizes the proportional response in one variable to a proportionate change in another variable, and so is given by $X_k \partial \log \theta(t, X)/\partial X_k = \beta_k X_k$. If $X_k \equiv \log(Z_k)$, so that the covariate is measured in logs, then it follows that β_k is the elasticity of the hazard with respect to Z_k.

If Stata is used to estimate a PH model, you can choose to report estimates of either $\widehat{\beta}_k$ or $\exp(\widehat{\beta}_k)$, for each k.

In addition, observe that, for two persons with the same $X = \overline{X}$, but different survival times:

$$\frac{\theta\left(t, \overline{X}\right)}{\theta\left(u, \overline{X}\right)} = \frac{\theta_0\left(t\right)}{\theta_0\left(u\right)} \tag{3.18}$$

so the ratio of hazards does not depend on X in this case.

Some further implications of the PH assumption

If $\theta\left(t, X\right) = \theta_0\left(t\right)\lambda$ and λ does not vary with survival time, i.e. the PH specification applies, then

$$\begin{aligned} S\left(t, X\right) &= \exp\left[-\int_0^t \theta\left(u\right) du\right] && (3.19)\\ &= \exp\left[-\lambda \int_0^t \theta_0\left(u\right) du\right] && (3.20)\\ &= \left[S_0\left(t\right)\right]^{\lambda} && (3.21) \end{aligned}$$

given the 'baseline' survivor function, $S_0(t)$, where

$$S_0(t) \equiv \exp\left[-\int_0^t \theta_0(u)\,du\right]. \tag{3.22}$$

Thus $\log S(t,X) = \lambda \log S_0(t)$. You can verify that it is also the case that $\log S(t,X) = \lambda^* \log S_0^*(t)$ where $S_0^*(t) \equiv \exp\left[-\int_0^t \theta_0^*(u)\,du\right]$.

Hence, in the PH model, differences in characteristics imply a scaling of the common baseline survivor function. Given the relationships between the survivor function and the probability density function, we also find that for a PH model

$$f(t) = f_0(t)\lambda\left[S_0(t)\right]^{\lambda-1} \tag{3.23}$$

where $f_0(t)$ is the baseline density function, $f_0(t) = f(t|X=0)$.

Alternatively, (3.19) may be re-written in terms of the integrated hazard function $H(t) = -\ln S(t)$, implying

$$H(t) = \lambda H_0(t). \tag{3.24}$$

where $H_0(t) = -\ln S_0(t)$.

If the baseline hazard does not vary with survival time, $\theta_0(u) = \theta$ for all u, then $H_0(t) = \int_0^t \theta_0(u)\,du = \theta\int_0^t du = \theta t$. So, in the constant hazard rate case, a plot of the integrated hazard against survival time should yield a straight line through the origin. Indeed, if the integrated hazard plot is concave to the origin (rising but at a decreasing rate), then this suggests that the hazard rate declines with survival time rather than constant. Similarly if the integrated hazard plot is convex to the origin (rising but at a increasing rate), then this suggests that the hazard rate increases with survival time.

The relationship between the survivor function and baseline survivor function also implies that

$$\begin{aligned} \ln[-\ln S(t)] &= \ln\lambda + \ln\left(-\ln\left[S_0(t)\right]\right) &(3.25)\\ &= \beta' X + \ln(-\ln[S_0(t)]) &(3.26)\end{aligned}$$

or, re-written in terms of the integrated hazard function,

$$\ln[H(t)] = \beta' X + \ln[H_0(t)] \tag{3.27}$$

The left-hand side is the log of the integrated hazard function at survival time t. The result suggests that one can check informally whether a model satisfies the PH assumption by plotting the log of the integrated hazard function against survival time (or a function of time) for groups of persons with different sets of characteristics. (The estimate of the integrated hazard would be derived using a non-parametric estimator such as the Nelson-Aalen estimator referred to in Chapter 4.) If the PH assumption holds, then the plots of the estimated $\ln[H(t)]$ against t for each of the groups should have different vertical intercepts

but a common shape, moving in parallel. To see this, suppose that there are two groups, labelled A and B. Hence,

$$\begin{aligned} \ln[H_A(t)] &= \beta' X_A + \ln[H_0(t)] \\ \ln[H_B(t)] &= \beta' X_B + \ln[H_0(t)]. \end{aligned}$$

The group-specific intercepts are the numbers implied by $\beta' X_A$ and $\beta' X_B$ whereas the common shape arises because the remaining term, the function of t is common to both groups.

Incorporating time-varying covariates

It is relatively straightforward in principle to generalise the PH specification to allow for time-varying covariates. For example, suppose that

$$\theta(t, X_t) = \theta_0(t) \exp(\beta' X_t) \tag{3.28}$$

where there is now a 't' subscript on the vector of characteristics X. Observe that for any given survival time $t = \bar{t}$, we still have that absolute differences in X correspond to proportional differences in the hazard. However the proportionality factor now varies with survival time rather than being constant (since the variables in X on the right-hand side of the equation are time-varying). The survivor function is more complicated too. From the standard relationship between the hazard and the survivor functions, we have:

$$\begin{aligned} S(t, X_t) &= \exp\left[-\int_0^t \theta(u)\, du\right] &(3.29)\\ &= \exp\left[-\int_0^t \theta_0(u) \exp(\beta' X_u) du\right]. &(3.30) \end{aligned}$$

In general, the survivor function cannot be simply factored as in the case when covariates are constant. Progress can be made, however, by assuming that *each covariate is constant within some pre-defined time interval*.

To illustrate this, suppose that there is a single covariate that takes on two values, depending on whether survival time is before or after some date s, i.e.

$$\begin{aligned} X &= X_1 \text{ if } t < s \\ X &= X_2 \text{ if } t \geq s. \end{aligned} \tag{3.31}$$

The survivor function (3.29) now becomes

$$S(t, X_t) = \exp\left[-\int_0^s \theta_0(u)\exp(\beta X_1)du - \int_s^t \theta_0(u)\exp(\beta X_2)du\right] \quad (3.32)$$

$$= \exp\left[-\lambda_1\int_0^s \theta_0(u)du - \lambda_2\int_s^t \theta_0(u)\,du\right] \quad (3.33)$$

$$= \exp\left[-\lambda_1\int_0^s \theta_0(u)\,du\right]\exp\left[-\lambda_2\int_s^t \theta_0(u)\,du\right] \quad (3.34)$$

$$= [S_0(s)]^{\lambda_1}\frac{[S_0(t)]^{\lambda_2}}{[S_0(s)]^{\lambda_2}}. \quad (3.35)$$

The probability of survival until time t is the probability of survival until time s, times the probability of survival until time t conditional on survival up until time s (the conditioning is accounted for by the expression in the denominator of the second term in the product). The density function can be derived in the usual way as the product of the survivor function and the hazard rate.

Expressions of this form underpin the process of estimation of PH models with time-varying covariates. As we shall see in Chapter 5, there are two parts to practical implementation: (a) reorganisation of one's data set to incorporate the time-varying covariates ('episode splitting'), and (b) utilisation of estimation routines that allow for conditioned survivor functions such as those in the second term in (3.32) – so-called 'delayed entry'.

3.2.7 Accelerated Failure Time (AFT) models

Let us again assume for the moment that personal characteristics do not vary with survival time.

AFT specification

The AFT model assumes a linear relationship between the log of (latent) survival time T and characteristics X :

$$\ln(T) = \beta^{*\prime}X + z \quad (3.36)$$

where β^* is a vector of parameters (cf. 3.6), and z is an error term. This expression may be re-written as

$$Y = \mu + \sigma u \quad (3.37)$$

or

$$\frac{Y-\mu}{\sigma} = u \quad (3.38)$$

where $Y \equiv \ln(T), \mu \equiv \beta^{*\prime}X$, and $u = z/\sigma$ is an error term with density function $f(u)$, and σ is a scale factor (which is related to the shape parameters for the hazard function – see below). This form is sometimes referred to as a generalized linear model (GLM) or log-linear model specification.

Distribution of u	Distribution of T
Extreme Value (1 parameter)	Exponential
Extreme Value (2 parameter)	Weibull
Logistic	Log-logistic
Normal	Lognormal
log Gamma (3 parameter Gamma)	Generalised Gamma

Table 3.2: Different error term distributions imply different AFT models

Distributional assumptions about u determine which sort of regression model describes the random variable T: see Table 3.2.

In the case of the Exponential distribution, the scale factor $\sigma = 1$, and $f(u) = \exp[u - \exp(u)]$. In the Weibull model, σ is a free parameter, and $\sigma = 1/\alpha$ to use our earlier notation. The density $f(u) = \alpha \exp[\alpha u - \exp(\alpha u)]$ in this case. For the Log-logistic model, $f(u) = \exp(u)/[1 + \exp(u)]$, and free parameter $\sigma = \varphi$, to use our earlier notation.

For the Lognormal model, the error term u is normally distributed, and we have a specification of a (log)linear model that appears to be similar to the standard linear model that is typically estimated using ordinary least squares. This is indeed the case: it is the example that we considered in the Introduction. The specification underscores the claim that the reason that OLS did not provide good estimates in the example was because it could not handle censored data rather than the specification itself. (The trick to handle censored data turns out to hinge on a different method of estimation, i.e. maximum likelihood. See Chapter 5.) The table also indicates that the normal linear model might be inappropriate for an additional reason: the hazard rate function may have a shape other than that assumed by the lognormal specification.

Interpretation

But why the 'Accelerated Failure Time' label for these models? From (3.36), and letting $\psi \equiv \exp\left(-\beta^{*\prime} X\right) = \exp(-\mu)$, it follows that:

$$\ln(T\psi) = z. \tag{3.39}$$

The term ψ, which is constant by assumption, acts like a time scaling factor. The two key cases are as follows:

- $\psi > 1$: it is as if the clock ticks faster (the time scale for someone with characteristics X is $T\psi$, whereas the time scale for someone with characteristics $X = 0$ is T). Failure is 'accelerated' (survival time shortened).
- $\psi < 1$: it is as if the clock ticks slower. Failure is 'decelerated' (survival time lengthened).

We can also see this time scaling property directly in terms of the survivor function. Recall the definition of the survivor function:

$$S(t, X) = \Pr[T > t|X]. \tag{3.40}$$

Expression (3.40) is equivalent to writing:

$$\begin{aligned} S(t, X) &= \Pr[Y > \ln(t)|X] && (3.41)\\ &= \Pr[\sigma u > \ln(t) - \mu] && (3.42)\\ &= \Pr[\exp(\sigma u) > t\exp(-\mu)]. && (3.43) \end{aligned}$$

Now define a 'baseline' survivor function, $S_0(t, X)$, which is the corresponding function when all covariates are equal to zero, i.e. $X = 0$, in which case then $\mu = \beta_0^*$, and $\exp(-\mu) = \exp(-\beta_0^*) \equiv \psi_0$. Thus:

$$S_0(t) = \Pr[T > t|X = 0]. \tag{3.44}$$

This expression can be rewritten using (3.43) as

$$S_0(t) = \Pr[\exp(\sigma u) > t\psi_0] \tag{3.45}$$

or

$$S_0(s) = \Pr[\exp(\sigma u) > s\psi_0] \tag{3.46}$$

for any s. In particular, let $s = t\exp(-\mu)/\psi_0$, and substitute this into the expression for $S_0(s)$. Comparisons with (3.43), show that we can rewrite the survivor function as:

$$\begin{aligned} S(t, X) &= S_0[t\exp(-\mu)] && (3.47)\\ &= S_0[t\psi] && (3.48) \end{aligned}$$

where $\psi \equiv \exp(-\mu)$. It follows that $\psi > 1$ is equivalent to having $\mu < 0$ and $\psi < 1$ is equivalent to having $\mu > 0$.

In sum, the effect of the covariates is to change the time scale by a constant (survival time-invariant) scale factor $\psi \equiv \exp(-\mu)$. When explaining this idea, Allison (1995, p. 62) provides an illustration based on longevity. The conventional wisdom is that one year for a dog is equivalent to seven years for a human, i.e. in terms of actual calendar years, dogs age faster than humans do. In terms of (3.47), if $S(t, X)$ describes the survival probability of a dog, then $S_0(t)$ describes the survival probability of a human, and $\psi = 7$.

How may we intrepret the coefficients β_k^*? Differentiation shows that

$$\beta_k^* = \frac{\partial \ln(T)}{\partial X_k}. \tag{3.49}$$

Thus an AFT regression coefficient relates proportionate changes in survival time to a unit change in a given regressor, with all other characteristics held fixed. (Contrast this with the interpretation of the coefficients in the PH model – they relate a one unit change in a regressor to a proportionate change in the

hazard rate, not survival time.) Stata allows you to choose to report either $\widehat{\beta}_k^*$or $\exp(\widehat{\beta}_k^*)$.

Recall from (3.36) that

$$T = \exp(\beta^{*\prime} X)\exp(z) \tag{3.50}$$

If persons i and j are identical on all but the kth characteristic, i.e. $X_{im} = X_{jm}$ for all $m \in \{1, ..., K\backslash k\}$, and they have the same z, then

$$\frac{T_i}{T_j} = \exp[\beta_k^*(X_{ik} - X_{jk})]. \tag{3.51}$$

If, in addition, $X_{ik} - X_{jk} = 1$, i.e. there is a one unit change in X_k, ceteris paribus, then

$$\frac{T_i}{T_j} = \exp(\beta_k^*). \tag{3.52}$$

The $\exp(\beta_k^*)$ is called the *time ratio*.

Some further implications

Using the general result that $\theta(t) = -[\partial S(t)/\partial t]/S(t)$, and applying it to (3.47), the relationship between the hazard rate for someone with characteristics X and the hazard rate for the case when $X = 0$ (i.e. the 'baseline' hazard $\theta_0(.)$), is given for AFT models by:

$$\theta(t, X) = \psi\theta_0(t\psi). \tag{3.53}$$

Similarly, the probability density function for AFT models is given by:

$$\begin{aligned} f(t, X) &= \psi\theta_0(t\psi)S_0(t\psi) \\ &= \psi f_0(t). \end{aligned} \tag{3.54}$$

The Weibull model is the only model that satisfies both the PH and AFT assumptions

The Weibull distribution is the only distribution for which, with constant covariates, the PH and AFT models coincide. To show this requires finding the functional form which satisfies the restriction that

$$[S_0(t)]^\lambda = S_0[t\psi]. \tag{3.55}$$

See Cox and Oakes (1984, p. 71) for a proof.

This property implies a direct correspondence between the parameters used in the Weibull PH and Weibull representations. What are the relationships? Recall that the PH version of Weibull model is:

$$\theta\,(t, X) = \alpha t^{\alpha-1}\lambda \tag{3.56}$$

$$= \theta_0\,(t)\,\lambda \tag{3.57}$$

with baseline hazard $\theta_0(t) = \alpha t^{\alpha-1}$, which is sometimes written $\theta_0(t) = (1/\sigma)t^{\frac{1}{\sigma}-1}$ with $\sigma = 1/\alpha$ (see earlier for reasons). Making the appropriate substitutions, one can show that the Weibull AFT coefficients (β^*) are related to the Weibull PH coefficients (β) as follows:

$$\beta_k^* = -\sigma\beta_k = -\beta_k/\alpha, \text{ for each } k. \tag{3.58}$$

To see this for yourself, substitute the PH and AFT Weibull survivor functions into (3.55):

$$[\exp(-t^\alpha)]^\lambda = \exp(-[t\exp(-\mu)]^\alpha).$$

By taking logs of both sides, and rearranging expressions, you will find that this equality holds only if $\beta_k^* = -\beta_k/\alpha$, for each k.

In Stata, you can choose to report either β or β^* (or the exponentiated values of each coefficient).

Incorporating time-varying covariates

Historically AFT models have been specified assuming that the vector summarizing personal characteristics is time-invariant. Clearly it is difficult to incorporate them directly into the basic equation that relates log survival time to charcteristics: see (3.36). However they can be incorporated via the hazard function and thence the survivor and density functions. A relatively straightforward generalisation of the AFT hazard to allow for time-varying covariates is to suppose that

$$\theta(t, X_t) = \psi_t\theta_0(t\psi_t) \tag{3.59}$$

where there is now a 't' subscript on $\psi_t \equiv \exp(-\beta^{*\prime}X_t)$. As in the PH case, we can factor the survivor function if we assume each covariate is constant within some pre-defined time interval. To illustrate this, again suppose that there is a single covariate that takes on two values, depending on whether survival time is before or after some date s, i.e.

$$\begin{array}{l} X = X_1 \text{ if } t < s \\ X = X_2 \text{ if } t \geq s. \end{array} \tag{3.60}$$

We therefore have $\psi_1 \equiv \exp(-\beta^* X_1)$ and $\psi_2 \equiv \exp(-\beta^* X_2)$. The survivor function now becomes

$$S(t, X_t) = \exp\left[-\int_0^s \theta_0(u\psi_1)\psi_1 du - \int_s^t \theta_0(u\psi_2)\psi_2 du\right] \tag{3.61}$$

$$= [S_0(s\psi_1)]^{\psi_1} \frac{[S_0(t\psi_2)]^{\psi_2}}{[S_0(s\psi_2)]^{\psi_2}}. \tag{3.62}$$

where manipulations similar to those used in the corresponding derivation for the PH case have been used. The density function can be derived as the product

Function	Proportional hazards	Accelerated failure time
Hazard, $\theta(t, X)$	$\theta_0(t)\lambda$	$\psi\theta_0(t\psi)$
Survivor, $S(t, X)$	$[S_0(t)]^{\lambda}$	$S_0(t\psi)$
Density, $f(t, X)$	$f_0(t)\lambda[S_0(t)]^{\lambda-1}$	$\psi f_0(t)$

Note: $\lambda = \exp(\beta' X)$; $\psi = \exp(-\beta^{*\prime} X)$. The characteristics vector X is time-invariant.

Table 3.3: Specification summary: proportional hazard versus accelerated failure time models

Model	PH	AFT
Exponential	✓	✓
Weibull	✓	✓
Log-logistic	×	✓
Lognormal	×	✓
Gompertz	✓	×
Generalized Gamma	×	✓

Table 3.4: Classification of models as PH or AFT: summary

of the hazard rate and this survivor function. As in the PH case, estimation of AFT models with time-varying covariates requires a combination of episode splitting and software that can handle conditioned survivor functions (delayed entry).

3.2.8 Summary: PH versus AFT assumptions for continuous time models

We can now summarise the assumptions about the functional forms of the hazard function, density function, and survivor function: see Table 3.3 which shows the case where there are no time-varying covariates.

Table 3.4 classifies parametric models according to whether they can be interpreted as PH or AFT. Note that the PH versus AFT description refers to the interpretation of parameter estimates and not to differences in how the model per se is estimated. As we see in later chapters, estimation of all the models cited so far is based on expressions for survivor functions and density functions.

3.2.9 A semi-parametric specification: the piecewise-constant Exponential (PCE) model

The *Piecewise-Constant Exponential* (PCE) model is an example of a semi-parametric continuous time hazard specification. By contrast with all the parametric models considered so far, the specification does not completely character-

ize the shape of the hazard function. Whether it generally increases or decreases with survival time is left to be fitted from the data, rather than specified a priori.

Insert chart shape of baseline hazard in this case.

The time axis is partitioned into a number of intervals using (researcher-chosen) cut-points. It is assumed that the hazard rate is constant within each interval but may, in principle, differ between intervals. An advantage of the model compared to the ones discussed so far is that the overall shape of the hazard function does not have to be imposed in advance. For example, with this model, one can explore whether the hazard does indeed appear to vary monotonically with survival time, and then perhaps later choose one of the parametric models in the light of this check.

It turns out that this is a special (and simple) case of the models incorporating time-varying covariates discussed earlier. The PCE model is a form of PH model for which we have, in general, $\theta(t, X_t) = \theta_0(t)\exp(\beta' X_t)$. In the PCE special case, we have:

$$\theta(t, X_t) = \begin{cases} \overline{\theta}_1 \exp(\beta' X_1) & t \in (0, \tau_1] \\ \overline{\theta}_2 \exp(\beta' X_2) & t \in (\tau_1, \tau_2] \\ \vdots & \vdots \\ \overline{\theta}_K \exp(\beta' X_K) & t \in (\tau_{K-1}, \tau_K] \end{cases} \tag{3.63}$$

The baseline hazard rate ($\overline{\theta}$) is constant within each of the K intervals but differs between intervals. Covariates may be fixed or, if time-varying, constant within each interval. This expression may be rewritten as

$$\theta(t, X_t) = \begin{cases} \exp\left[\log(\overline{\theta}_1) + \beta' X_1\right] & t \in (0, \tau_1] \\ \exp\left[\log(\overline{\theta}_2) + \beta' X_2\right] & t \in (\tau_1, \tau_2] \\ \vdots & \vdots \\ \exp\left[\log(\overline{\theta}_K) + \beta' X_K\right] & t \in (\tau_{K-1}, \tau_K] \end{cases} \tag{3.64}$$

or

$$\theta(t, X_t) = \begin{cases} \exp(\widetilde{\lambda}_1) & t \in (0, \tau_1] \\ \exp(\widetilde{\lambda}_2) & t \in (\tau_1, \tau_2] \\ \vdots & \vdots \\ \exp(\widetilde{\lambda}_K) & t \in (\tau_{K-1}, \tau_K] \end{cases} \tag{3.65}$$

so the constant interval-specific hazard rates are equivalent to having interval-specific intercept terms in the overall hazard. One can estimate these by defining binary (dummy) variables that refer to each interval, and the required estimates are the estimated coefficients on these variables. However, observe that in order to identify the model parameters, one cannot include all the interval-specific dummies and an intercept term in the regression. Either one includes all the dummies and excludes the intercept term, or includes all but one dummy and includes an intercept. (To see why, note that, for the first interval, and similarly for the others, $\widetilde{\lambda}_1 = \log(\overline{\theta}_1) + \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_K X_K$.)

The expression for the PCE survivor function is a special case of the earlier expression for survivor function where we assumed that covariates were constant within some pre-defined interval (cf. 3.32). If we consider the case where $K = 2$ (as earlier), then it can be shown that

$$S(t, X_t) = [S_0(s)]^{\widetilde{\lambda}_1} \frac{[S_0(t)]^{\widetilde{\lambda}_2}}{[S_0(s)]^{\widetilde{\lambda}_2}} \tag{3.66}$$

$$= [\exp(-s)]^{\widetilde{\lambda}_1} \frac{[\exp(-t)]^{\widetilde{\lambda}_2}}{[\exp(-s)]^{\widetilde{\lambda}_2}} \tag{3.67}$$

$$= \exp(-s\widetilde{\lambda}_1) \exp[-(t-s)\widetilde{\lambda}_2]. \tag{3.68}$$

It was stated earlier that expressions with this structure were the foundation for estimation of models with time-varying covariates – combining episode splitting and software that can handle delayed entry spell data. For the PCE model, however, one only needs to episode split; one does not need software that can handle spells with delayed entry (the spell from s to t in our example). With a constant hazard rate, the relevant likelihood contribution for this spell is the same as a spell from time 0 to $(t - s)$: note the second term on the right-hand side of (3.68).

The PCE model should be distinguished from Cox's Proportional Hazard model that is considered in Chapter 7. Both are continuous time models, can incorporate time-varying covariates, and allow for some flexibility in the shape of the hazard function. However the Cox model is more general, in the sense that it allows estimates of the slope parameters in the β vector to be derived regardless of what the baseline hazard function looks like. The PCE model requires researcher input in its specification (the cutpoints); the Cox model estimates are derived for a totally arbitrary baseline hazard function. On the other hand, if you desire flexibility and explicit estimates of the baseline hazard function, you might use the PCE model.

3.3 Discrete time specifications

We consider two models. The first is the discrete time representation of a continuous time proportional hazards model, and leads to the so-called complementary log-log specification. This model can also be applied when survival times are intrinsically discrete. The second model, the logistic model, was primarily developed for this second case but may also be applied to the first. It may be given an interpretation in terms of the proportional odds of failure. For expositional purposes we assume fixed covariates.

3.3.1 A discrete time representation of a continuous time proportional hazards model

The underlying continuous time model is summarised by the hazard rate $\theta(t, X)$, but the available survival time data are interval-censored – grouped or banded into intervals in the manner described earlier. That is, exact survival times are not known, only that they fall within some interval of time. What we do here is derive an estimate of parameters describing the continuous time hazard, but taking into account the nature of the banded survival time data that is available to us.

The survivor function at time a_j, the date marking the end of the interval $(a_{j-1}, a_j]$, is given by:

$$S(a_j, X) = \exp\left[-\int_0^{a_j} \theta(u, X)\, du\right]. \tag{3.69}$$

Suppose also that the hazard rate satisfies the PH assumption:

$$\theta(t, X) = \theta_0(t)\, e^{\beta' X} = \theta_0(t)\, \lambda \tag{3.70}$$

where, as before, $\beta' X \equiv \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_K X_K$ and $\lambda \equiv \exp(\beta' X)$. These assumptions imply that

$$\begin{aligned} S(a_j, X) &= \exp\left[-\int_0^{a_j} \theta_0(t)\, \lambda du\right] && (3.71)\\ &= \exp\left[-\lambda \int_0^{a_j} \theta_0(t)\, du\right] && (3.72)\\ &= \exp\left[-H_j \lambda\right] && (3.73) \end{aligned}$$

where $H_j \equiv H(a_j) = \int_0^{a_j} \theta_0(u, X)\, du$ is the integrated baseline hazard evaluated at the end of the interval. Hence, the baseline survivor function at a_j is:

$$S_0(a_j) = \exp(-H_j). \tag{3.74}$$

The discrete time (interval) hazard function, $h(a_j, X) \equiv h_j(X)$ is defined by

$$\begin{aligned} h_j(X) &= \frac{S(a_{j-1}, X) - S(a_j, X)}{S(a_{j-1}, X)} && (3.75)\\ &= 1 - \frac{S(a_j, X)}{S(a_{j-1}, X)} && (3.76)\\ &= 1 - \exp\left[\lambda\left(H_{j-1} - H_j\right)\right] && (3.77) \end{aligned}$$

which implies that

$$\log\left(1 - h_j(X)\right) = \lambda\left(H_{j-1} - H_j\right) \tag{3.78}$$

and hence

$$\log\left(-\log[1 - h_j(X)]\right) = \beta' X + \log(H_j - H_{j-1}).$$

Similarly, the discrete time (interval) *baseline* hazard for the interval (a_{j-1}, a_j) is

$$1 - h_{0j} = \exp\left(H_{j-1} - H_j\right) \tag{3.79}$$

and hence

$$\begin{aligned} \log\left[-\log\left(1 - h_{0j}\right)\right] &= \log\left(H_j - H_{j-1}\right) && (3.80)\\ &= \log\left[\int_{a_{j-1}}^{a_j} \theta_0\left(u\right) du\right] && (3.81)\\ &= \gamma_j, \text{ say} && (3.82) \end{aligned}$$

where γ_j is the log of the difference between the integrated baseline hazard $\theta_0(t)$ evaluated at the end of the interval $(a_{j-1}, a_j]$ and the beginning of the interval. We can substitute this expression back into that for $h(a_j, X)$ and derive an expression for the interval hazard rate:

$$\begin{aligned} \log(-\log[1 - h_j\left(X\right)]) &= \beta' X + \gamma_j, \text{ or} && (3.83)\\ h\left(a_j, X\right) &= 1 - \exp[-\exp\left(\beta' X + \gamma_j\right)]. && (3.84) \end{aligned}$$

The $\log(-\log(.))$ transformation is known as the *complementary log-log* transformation; hence the discrete-time PH model is often referred to as a *cloglog model*.

Observe that, if each interval is of unit length, then we can straightforwardly index time intervals in terms of the interval number rather than the dates marking the end of each interval. In this case, we can write the discrete time hazard equivalently as

$$h\left(j, X\right) = 1 - \exp[-\exp\left(\beta' X + \gamma_j\right)].$$

The cloglog model is a form of generalized linear model with particular link function: see (3.83). When estimated using interval-censored survival data, one derives estimates of the regression coefficients β, and of the parameters γ_j. The β coefficients are the same ones as those characterizing the continuous time hazard rate $\theta\left(t\right) = \theta_0\left(t\right)\exp(\beta' X)$. However the parameters characterizing the baseline hazard function $\theta_0\left(t\right)$ cannot be identified without further assumptions: the γ_j summarize differences in values of the integrated hazard function, and are consistent with a number of different shapes of the hazard function within each interval. To put things another way, the γ_j summarize the pattern of duration dependence in the *interval* hazard, but one cannot identify the precise pattern of duration dependence in the *continuous* time hazard without further assumptions.

Restrictions on the specification of the γ_j can lead to discrete time models corresponding directly to the parametric PH models considered in the previous chapter. For example, if the continuous time model is the Weibull one, and we have unit-length intervals, then evaluation of the integrated baseline hazard

– see later in chapter for the formula – reveals that $\gamma_j = (j)^\alpha - (j-1)^\alpha$. In practice, however, most analysts do not impose restrictions such as these on the γ_j when doing empirical work. Instead, either they do not place any restrictions on how the γ_j vary from interval to interval (in which case the model is a type of semi-parametric one) or, alternatively, they specify duration dependence in terms of the discrete time hazard (rather than the continuous time one). That is, the variation in γ_j from interval to interval is specified using a parametric functional form. Examples of these specifications are given below.

Note, finally, that the cloglog model is not the only model that is consistent with a continuous time model and interval-censored survival time data (though it is the most commonly-used one). Sueyoshi (1995) has shown how, for example, a logistic hazard model (as considered in the next sub-section) with interval-specific intercepts may be consistent with an underlying continuous time model in which the within-interval durations follow a loglogistic distribution.

3.3.2 A model in which time is intrinsically discrete

When survival times are intrinsically discrete, we could of course apply the discrete time PH model that we have just discussed, though of course it would not have the same interpretation. An alternative, also commonly used in the literature, is a model that can be labelled the *proportional odds model*. (Beware that here the odds refer to the hazard, unlike in the case of the Loglogistic model where they referred to the survivor function.) Let us suppose, for convenience, that survival times are recorded in 'months'.

The proportional odds model assumes that the relative odds of making a transition in month j, given survival up to end of the previous month, is summarised by an expression of the form:

$$\frac{h\left(j,X\right)}{1-h\left(j,X\right)} = \left[\frac{h_0\left(j\right)}{1-h_0\left(j\right)}\right] \exp\left(\beta' X\right) \tag{3.85}$$

where $h\left(j,X\right)$ is the discrete time hazard rate for month j, and $h_0\left(j,X\right)$ is the corresponding baseline hazard arising when $X=0$. The relative odds of making a transition at any given time is given by the product of two components: (a) a relative odds that is common to all individuals, and (b) an individual-specific scaling factor. It follows that

$$\log \text{it}\left[h\left(j,X\right)\right] = \log\left[\frac{h\left(j,X\right)}{1-h\left(j,X\right)}\right] = \alpha_j + \beta' X \tag{3.86}$$

where $\alpha_j = \text{logit}[\ h_0\left(j\right)]$. We can write this expression, alternatively, as

$$h(j,X) = \frac{1}{1+\exp\left(-\alpha_j - \beta' X\right)}. \tag{3.87}$$

This is the *logistic hazard model* and, given its derivation, has a proportional odds interpretation. In principle the α_j may differ for each month. Often however the pattern of variation in the α_j, and the γ_j in the cloglog model, are characterized using some function of j.

3.3.3 Functional forms for characterizing duration dependence in discrete time models

Examples of duration dependence specifications include:

- $r \log(j)$, which can be thought of a discrete-time analogue to the continuous time Weibull model, because the shape of the hazard monotonically increases if $r > 0$, decreases if $r < 0$, or is constant if $r = 0$ ($r + 1 = q$ is thus analogous to the Weibull shape parameter α). If this specification were combined with a logistic hazard model, then the full model specification would be $\text{logit}[h(j, X)] = r \log(j) + \beta' X$, and r is a parameter that would be estimated together with intercept and slope parameters within the vector β.

- $z_1 j + z_2 j^2 + z_3 j^3 + ... + z_p j^p$, i.e. a p-th order polynomial function of time, where the shape parameters are $z_1, z_2, z_3, ...z_p$. If $p = 2$ (a quadratic function of time), the interval hazard is U-shaped or inverse-U-shaped. If the quadratic specification were combined with a cloglog hazard model, then the full model specification would be $\text{cloglog}[h(j, X)] = z_1 j + z_2 j^2 + \beta' X$, and $z_1 j$ and z_2 are parameters that would be estimated together with intercept and slope parameters within the vector β.

- piecewise constant, i.e. groups of months are assumed to have the same hazard rate, but the hazard differs between these groups. If this specification were combined with a logistic hazard model, then the full model specification would be $\text{cloglog}[h(j, X)] = \gamma_1 D_1 + \gamma_2 D_2 + ... + \gamma_J D_J + \beta' X$, where D_l is a binary variable equal to one if $j = l$ and equal to zero otherwise. I.e. the researcher creates dummy variables corresponding to each interval (or group of intervals). When estimating the full model, one would not include an intercept term within the vector β, or else as it would be collinear. (Alternatively one could include an intercept term but drop one of the dummy variables.)

The choice of shape of hazard function in these models is up to the investigator – just as one can choose between different parametric functional forms in the continuous time models.

In practice, cloglog and logistic hazard models that share the same duration dependence specification and the same X yield similar estimates – as long as the hazard rate is relatively 'small'. To see why this is so, note that

$$\log \text{it}(h) = \log\left(\frac{h}{1-h}\right) = \log(h) - \log(1-h). \tag{3.88}$$

As $h \to 0$, then $\log(1-h) \to 0$ also. That is,

$$\log \text{it}(h) \approx \log(h) \text{ for 'small' } h.$$

With a sufficiently small hazard rate, the proportional odds model (a linear function of duration dependence and characteristics) is a close approximation

to a model with the log of the hazard rate as dependent variable. The result follows from the fact that the estimates of β from a discrete time proportional hazards model correspond to those of a continuous time model in which $\log(\theta)$ is a linear function of characteristics.

3.4 Deriving information about survival time distributions

The sorts of questions that one might wish to know, given estimates of a given model, include:

- How long are spells (on average)?
- How do spell lengths differ for persons with different characteristics?
- What is the pattern of duration dependence (the shape of the hazard function with survival time)?

To provide answers to these questions, we need to know the shape of the survivor function for each model. We also need expressions for the density and survivor functions in order to construct expressions for sample likelihoods in order to derive model estimates using the maximum likelihood principle, as we shall see in Chapters 5 and 6. The integrated hazard function also has several uses, including being used for specification tests. We derived a number of general results for PH and AFT models in an earlier section; in essence what we are doing here is illustrating those results, with reference to the specific functional forms for distributions that were set out in the earlier sections of this chapter.

To illustrate how one goes about deriving the relevant expressions, we focus on a relatively small number of models, and assume no time-varying covariates.

3.4.1 The Weibull model

Hazard rate

From Section 3.2, we know that the Weibull model hazard rate is given by

$$\theta(t, X) = \alpha t^{\alpha-1} \exp\left(\beta' X\right) \tag{3.89}$$

$$= \alpha t^{\alpha-1} \lambda \tag{3.90}$$

where $\lambda \equiv \exp\left(\beta' X\right)$. The baseline hazard function is $\theta_0(t) = \alpha t^{\alpha-1}$ or, using the alternative characterization mentioned earlier, $\theta_0^*(t) = \alpha t^{\alpha-1} \exp(\beta_0)$. The expression for the hazard rate implies that:

$$\log[\theta(t, X)] = \log \alpha + (\alpha - 1) \log(t) + \beta' X \tag{3.91}$$

We can easily demonstrate the Weibull model satisfies the general properties of PH model, discussed earlier. Remember that all these results also apply to the Exponential model, as that is simply the Weibull model with $\alpha = 1$. First,

$$\frac{\partial \theta(t, X)}{\partial X_k} = \theta \beta_k \tag{3.92}$$

$$\text{or } \frac{\partial \log \theta(t, X)}{\partial X_k} = \beta_k \tag{3.93}$$

where X_k is the kth covariate in the vector of characteristics X. Thus each coefficient summarises the proportionate response of the hazard to a small change in the relevant covariate.

From this expression can be derived the *elasticity of the hazard rate with respect to changes in a given characteristic*:

$$\frac{X_k}{\theta} \frac{\partial \theta}{\partial X_k} = \frac{\partial \log \theta}{\partial \log X_k} = \beta_k X_k. \tag{3.94}$$

If $X_k \equiv \log(Z_k)$, then the elasticity of the hazard with respect to changes in Z_k is

$$\frac{\partial \log \theta}{\partial \log Z_k} = \beta_k. \tag{3.95}$$

The Weibull model also has the PH property concerning the relative hazard rates for two persons at same t, but with different X :

$$\frac{\theta(\bar{t}, X_1)}{\theta(\bar{t}, X_2)} = \exp\left[\beta'(X_1 - X_2)\right]. \tag{3.96}$$

The Weibull shape parameter α can also be interpreted with reference to the elasticity of the hazard with respect to survival time:

$$\frac{\partial \log \theta(t, X)}{\partial \log(t)} = \alpha - 1. \tag{3.97}$$

Insert graphs

The relative hazard rates for two persons with same $X = \overline{X}$, but at different survival times t and u, where $t > u$, are given by

$$\theta(t, \overline{X}) / \theta(u, \overline{X}) = \left(\frac{t}{u}\right)^{\alpha - 1}. \tag{3.98}$$

Thus failure at time t is $(t/u)^{\alpha-1}$ times more likely than at time u, unless $\alpha = 1$ in which case failure is equally likely. Constrast the cases of $\alpha > 1$ and $\alpha < 1$.

Survivor function

Recall that $\theta(t,X) = \alpha t^{\alpha-1}\lambda$. But we know that for all models, $S(t,X) = 1 - F(t,X) = \exp\left[-\int_0^t \theta(u,X)\,\mathrm{d}u\right]$. So substituting the Weibull expression for the hazard rate into this expression:

$$\begin{aligned} S(t,X) &= \exp\left(-\int_0^t \alpha u^{\alpha-1}\lambda \mathrm{d}u\right) && (3.99)\\ &= \exp\left(-\lambda\alpha\left\{\frac{u^\alpha}{\alpha}\right]_0^t\right\}\right) && (3.100)\\ &= \exp\left(-\lambda\alpha\left[\frac{t^\alpha}{\alpha} - \frac{0^\alpha}{\alpha}\right]\right) && (3.101)\end{aligned}$$

using the fact that there are no time-varying covariates. The expression in the curly brackets {.} refers to the evaluation of the definite integral. Hence the Weibull survivor function is:

$$S(t,X) = \exp\left(-\lambda t^\alpha\right). \tag{3.102}$$

Insert graphs to show how varies with X and with α

Density function

Recall that $f(t) = \theta(t)S(t)$, and so in the Weibull case:

$$f(t,X) = \alpha t^{\alpha-1}\lambda \exp\left(-\lambda t^\alpha\right). \tag{3.103}$$

Integrated hazard function

Recall that $H(t,X) = -\ln S(t,X)$. Hence, in the Weibull case,

$$H(t,X) = \lambda t^\alpha. \tag{3.104}$$

It follows that

$$\begin{aligned} \log H(t,X) &= \log(\lambda) + \alpha\log(t) && (3.105)\\ &= \beta' X + \alpha\log(t). && (3.106)\end{aligned}$$

This expression suggests that a simple graphical specification test of whether observed survival times follow the Weibull distribution can be based on a plot of the log of the integrated hazard function against the log of survival time. If the distribution is Weibull, the graph should be a straight line (with slope equal to the Weibull shape parameter α) or, for groups characterised by combinations of X, the graphs should be parallel lines.

Quantiles of the survivor function, including median

The *median duration* is survival time m such that $S(m) = 0.5$.

*Insert chart * $S(t)$ against t, showing m

From (3.102) we have

$$\log S(t, X) = -\lambda t^{\alpha} \tag{3.107}$$

and hence

$$t = \left[\frac{1}{\lambda}\left[-\log S(t)\right]\right]^{1/\alpha}. \tag{3.108}$$

Thus, at the median survival time, we must also have that

$$m = \left[\frac{1}{\lambda}[-\log(0.5)]\right]^{1/\alpha} \tag{3.109}$$

$$= \left[\frac{\log(2)}{\lambda}\right]^{1/\alpha}. \tag{3.110}$$

Expressions for the upper and lower quartiles (or any other quantile) may be derived similarly. Substitute $t = 0.75$ or 0.25 in the expression for m rather than 0.5.

How does the median duration vary with differences in characteristics? Consider the following expression for the proportionate change in the median given a change in a covariate:

$$\frac{\partial \log m}{\partial X_k} = \frac{1}{\alpha}\frac{\partial[\log(2) - \log(\lambda)]}{\partial X_k} = -\frac{1}{\alpha}\frac{\partial \log(\lambda)}{\partial X_k} = -\frac{\beta_k}{\alpha}. \tag{3.111}$$

Observe also that this elasticity is equal to β_k^*. Also, using the result that $\partial \log m/\partial X_k = (1/m)\partial m/\partial X_k$, the elasticity of the median duration with respect to a change in characteristics is

$$\frac{X_k}{m}\frac{\partial m}{\partial X_k} = \frac{\partial \log(m)}{\partial \log(X_k)} = \frac{-\beta_k X_k}{\alpha} = \beta_k^* X_k. \tag{3.112}$$

If $X_k \equiv \log(Z_k)$, then the elasticity of the median with respect to changes in Z_k is $-\beta_k/\alpha = \beta_k^*$.

Mean (expected) survival time

In general, the mean or expected survival time is given by

$$\mathcal{E}(T) \equiv \int_0^{\infty} t f(t)\, dt = \int_0^{\infty} S(t)\, dt \tag{3.113}$$

where the expression in terms of the survivor function was derived using integration by parts. Hence in the Weibull case, $\mathcal{E}(T) = \int_0^{\infty} \exp(-\lambda t^{\alpha})\, dt$. Evaluation of this integral yields (Klein and Moeschberger 1990, p.37):

α	$\Gamma\left(1+\frac{1}{\alpha}\right)$	$[\log(2)]^{1/\alpha}$	Ratio of mean to median
4.0	0.906	0.912	0.99
3.0	0.893	0.885	1.01
2.0	0.886	0.832	1.06
1.1	0.965	0.716	1.35
1.0	1.000	0.693	1.44
0.9	1.052	0.665	1.58
0.8	1.133	0.632	1.79
0.7	1.266	0.592	2.14
0.6	1.505	0.543	2.77
0.5	2.000	0.480	4.17

Table 3.5: Ratio of mean to median survival time: Weibull model

$$\mathcal{E}(T) = \left(\frac{1}{\lambda}\right)^{\frac{1}{\alpha}} \Gamma\left(1+\frac{1}{\alpha}\right) \tag{3.114}$$

where $\Gamma(.)$ is the Gamma function. The Gamma function has a simple formula when its argument is an integer: $\Gamma(n) = (n-1)!$, for integer-valued n. For example, $\Gamma(5) = 4! = 4 \times 3 \times 2 \times 1 = 24$. $\Gamma(4) = 3! = 6$. $\Gamma(3) = 2! = 2$. $\Gamma(2) = 1$. For non-integer arguments, one has to use special tables (or a built-in function in one's software package). Observe that for Exponential model ($\alpha = 1$), the mean duration is simply $1/\lambda$.

So, if $\alpha = 0.5$, i.e. there is negative duration dependence (a monotonically declining hazard), then $\mathcal{E}(T) = (1/\lambda)^{\frac{1}{\alpha}} \Gamma(3) = 2/\lambda^2$. Compare this with the median $m \approx 0.480/\lambda^2$. More generally, the ratio of the mean to the median in the Weibull case is

$$ratio = \frac{\Gamma\left(1+\frac{1}{\alpha}\right)}{[\log(2)]^{1/\alpha}} \tag{3.115}$$

Table 3.5 summarises the ratio of the mean to median for various values of α. Observe that, unless the hazard is increasing at a particularly fast rate (large $\alpha > 0$), then the mean survival time is longer than the median.

The elasticity of the mean duration with respect to a change in characteristics is

$$\frac{X_k}{\mathcal{E}(T)} \frac{\partial \mathcal{E}(T)}{\partial X_k} = \frac{-\beta_k X_k}{\alpha} \tag{3.116}$$

which is exactly the same as the corresponding elasticity for the median. Note that $\partial \log \mathcal{E}(T) / \partial X_k = -\beta_k/\alpha$, so that if $X_k \equiv \log(Z_k)$, then the elasticity of the median with respect to changes in Z_k is $-\beta_k/\alpha$, i.e. the same expression as for the corresponding elasticity for the median!

3.4.2 Gompertz model

From Section 3.2, the Gompertz model hazard rate is given by

$$\theta\left(t,X\right)=\lambda\exp(\gamma t). \tag{3.117}$$

By integrating this expression, one derives the integrated hazard function

$$H\left(t,X\right)=\frac{\lambda}{\gamma}[\exp(\gamma t)-1]. \tag{3.118}$$

Since $H(t)=-\log S(t)$, it follows that the expression for the survivor function is:

$$S\left(t,X\right)=\exp\left(\frac{\lambda}{\gamma}[1-\exp(\gamma t)]\right). \tag{3.119}$$

The density function can be derived as $f(t,X)=\theta\left(t,X\right)S\left(t,X\right)$. Observe that if $\gamma=0$, then we have the Exponential hazard model.

To derive an expression for the median survival time, we applying the same logic as above, and find that

$$m=\frac{1}{\gamma}\log\left(1+\frac{\gamma\log 2}{\lambda}\right). \tag{3.120}$$

There is no closed-form expression for the mean.

3.4.3 Log-logistic model

Hazard rate

From Section 3.2, the Log-logistic model hazard rate is given by

$$\begin{aligned}\theta\left(t,X\right) &= \frac{\psi^{\frac{1}{\gamma}}t^{\left(\frac{1}{\gamma}-1\right)}}{\gamma\left[1+\left(\psi t\right)^{\frac{1}{\gamma}}\right]} &(3.121)\\ &= \frac{\varphi\psi^{\varphi}t^{(\varphi-1)}}{1+\left(\psi t\right)^{\varphi}} &(3.122)\end{aligned}$$

for $\varphi\equiv 1/\gamma$.where *shape parameter* $\gamma>0$ and $\psi=\exp\left(-\beta^{*\prime}X\right)$.

Survivor function

The log-logistic survivor function is given by

$$\begin{aligned}S\left(t,X\right) &= \frac{1}{1+\left(\psi t\right)^{1/\gamma}} &(3.123)\\ &= \frac{1}{1+\left(\psi t\right)^{\varphi}}. &(3.124)\end{aligned}$$

Beware: this expression differs slightly from that shown by e.g. Klein and Moeschberger (1997), since they parameterise the hazard function differently.

Insert graphs to show how varies with X and with γ

Density function

Recall that $f(t,X) = \theta(t,X)S(t,X)$, and so in the Log-logistic case:

$$f(t,X) = \frac{\frac{1}{\gamma}\psi^{\frac{1}{\gamma}}t^{\left(\frac{1}{\gamma}-1\right)}}{\left[1+(\psi t)^{1/\gamma}\right]^2} \tag{3.125}$$

$$= \frac{\varphi\psi^{\varphi}t^{(\varphi-1)}}{\left[1+(\psi t)^{\varphi}\right]^2} \tag{3.126}$$

Integrated hazard function

From above, we have

$$H(t,X) = \log\left[1+(\psi t)^{\frac{1}{\gamma}}\right] = \log\left[1+(\psi t)^{\varphi}\right] \tag{3.127}$$

and hence

$$\ln[\exp[H(t,X)]-1] = -\varphi\beta^{*\prime}X + \varphi\log(t)\,. \tag{3.128}$$

One might therefore graph an estimate of the left-hand side against $\log(t)$ as a specification check for the model: the graph should be a straight line. A more common check is expressed in terms of the log odds of survival, however: see below.

Quantiles of the survivor function, including median

The median survival time is

$$m = \psi^{-1}. \tag{3.129}$$

Hence it can be derived that $\partial m/\partial X_k = \beta_k/\psi^2$, and so the elasticity of the median with respect to a change in X_k is

$$\frac{\partial\log(m)}{\partial\log(X_k)} = \frac{\beta_k X_k}{\psi}. \tag{3.130}$$

Mean (expected) survival time

There is no closed form expression for mean survival time unless $\gamma < 1$ or equivalently $\varphi > 1$ (i.e. the hazard rises and then falls). In this case, the mean is (Klein and Moeschberger, 1997, p. 37):

$$\mathcal{E}(T) = \frac{1}{\psi} \frac{\gamma\pi}{\sin(\gamma\pi)}, \gamma < 1. \tag{3.131}$$

where π is the trigonometric constant 3.1415927.... (Klein and Moeschberger's expression is in terms of the relatively unfamiliar cosecant function, $\mathrm{Csc}(x)$. Note however that $\mathrm{Csc}(x) = 1/\sin(x)$.) The elasticity of the mean is given by

$$\frac{\partial \log \mathcal{E}(T)}{\partial \log(X_k)} = \frac{\beta_k X_k}{\psi} = \frac{\partial \log(m)}{\partial \log(X_k)}, \tag{3.132}$$

i.e. the same expression as for the elasticity of the median.

The ratio of the mean to the median is given by

$$\frac{\mathcal{E}(T)}{m} = \frac{\gamma\pi}{\sin(\gamma\pi)}, \gamma < 1.$$

Log-odds of survival interpretation

The Log-logistic model can be given a log-odds of survival interpretation. From (3.123), the (conditional) odds of survival to t is

$$\frac{S(t,X)}{1-S(t,X)} = (\psi t)^{-\frac{1}{\gamma}} \tag{3.133}$$

and at $X = 0$ (i.e. $\psi = \exp(-\beta_0^*) \equiv \psi_0$, also, so

$$\frac{S(t,X|X=0)}{1-S(t,X|X=0)} = (t\psi_0)^{-\frac{1}{\gamma}} \tag{3.134}$$

and therefore

$$\frac{S(t,X)}{1-S(t,X)} = \left[\frac{S(t,X|X=0)}{1-S(t,X|X=0)}\right] (\frac{\psi}{\psi_0})^{-\frac{1}{\gamma}}. \tag{3.135}$$

Hence the (conditional) log odds of survival at each t for any individual depend on a 'baseline' conditional odds of survival (common to all persons) and a person-specific factor depending on characteristics (and the shape parameter γ) that scales those 'baseline' conditional odds: $(\psi/\psi_0) = \exp(-\beta_1^{*\prime}X_1 - \beta_2^{*\prime}X_2 - ... - \beta_K^{*\prime}X_K)$.

The log-odds property suggests a graphical specification check for the Log-logistic model. From (3.133), we have

$$\log\left[\frac{S(t,X)}{1-S(t,X)}\right] = \beta^{*\prime}X - \varphi\log(t). \tag{3.136}$$

The check is therefore based on a graph of $\log[(1 - S(t, X))/S(t, X)]$ against $\log(t)$ – it should be a straight line if the Log-logistic model is appropriate. Or give rise to parallel lines if plotted separately for different groups classified by combinations of values of X.

3.4.4 Other continuous time models

See texts for discussion of other models. Most continuous time parametric models have closed form expressions for the median but not for the mean.

3.4.5 Discrete time models

The expressions for the median and mean, etc., depend on the baseline hazard specification that the researcher chooses. For discrete time models, there is typically no closed form expressions for median and mean and they have to be derived numerically (see web Lessons for examples).

Recall that survival up to the end of the jth interval (or completion of the jth cycle) is given by:

$$S(j) = S_j = \prod_{k=1}^{j} (1 - h_k) \tag{3.137}$$

and h_k is a logit or complementary log-log function of characteristics and survival time. In general this function is not invertible so as to produce suitable closed form expressions. The median is defined implicitly by finding the value of j such that $S(t_j) = 0.5$.

Things are simpler in the special case in which the discrete hazard rate is constant over time, i.e. $h_j = h$ all j (the case of a Geometric distribution of survival times). In this case,

$$S_j = (1 - h)^j \tag{3.138}$$

$$\log S_j = j \log(1 - h) \tag{3.139}$$

and so the median is

$$m = \frac{\log(0.5)}{\log(1 - h)} = \frac{-\log(2)}{\log(1 - h)}. \tag{3.140}$$

In the general case, in which the hazard varies with survival time, the mean survival time $\mathcal{E}(T)$ satisfies

$$\mathcal{E}(T) = \sum_{k=1}^{K} k f(k) \tag{3.141}$$

or, alternatively,

$$\mathcal{E}(T) = \sum_{k=1}^{K} S(k) \tag{3.142}$$

where $f(k) = \Pr(T = k)$, and K is the maximum survival time. In the special case in which the hazard rate is constant at all survival times, then

$$\mathcal{E}(T) = \left(\frac{1-h}{h}\right)\left[1-(1-h)^K\right] \tag{3.143}$$

implying that the mean survival time is approximated well by $(1-h)/h$ if K is 'large'.

Chapter 4

Estimation of the survivor and hazard functions

In this chapter, we consider estimators of the survivor and hazard functions – the empirical counterparts of the concepts that we considered in the previous chapter. One might think of the estimators considered here as non-parametric estimators, as no prior assumptions are made about the shapes of the relevant functions. The methods may be applied to a complete sample, or to subgroups of subjects within the sample, where the groups are defined by some combination of observable characteristics. The analysis of subgroups may, of course, be constrained by cell size considerations – which is one of the reasons for taking account of the effects of differences in characteristics on survival and hazard functions by using multiple regression methods. (These are discussed in the chapters following.) In any case, because the non-parametric analysis is informative about the pattern of duration dependence, it may also assist with the choice of parametric model.

We shall assume that we have a random sample from the population of spells, and allow for right-censoring (but not truncation). We consider first the Kaplan-Meier product-limit estimator and then the lifetable estimator. The main difference between them is that the former is specified assuming continuous time measures of spells, whereas for the latter the survival times are banded (grouped).

4.1 Kaplan-Meier (product-limit) estimators

Let $t_1 < t_2 < ... < t_j < ... < t_k < \infty$ represent the survival times that are observed in the data set. From the data, one can also determine the following quantities:

d_j : the number of persons observed to 'fail' (make a transition out of the state) at t_j

Failure time	# failures	# censored	# at risk of failure
t_1	d_1	m_1	n_1
t_2	d_2	m_2	n_2
t_3	d_3	m_3	n_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_j	d_j	m_j	n_j
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_k	d_k	m_k	n_k

Table 4.1: Example of data structure

m_j : the number of persons whose observed duration is censored in the interval $[t_j, t_{j+1})$, i.e. still in state at time t but not in state by $t+1$

n_j : the number of persons at risk of making a transition (ending their spell) immediately prior to t_j, which is made up of those who have a censored or completed spell of length t_j or longer:

$$n_j = (m_j + d_j) + (m_{j+1} + d_{j+1}) + ... + (m_k + d_k)$$

Table 4.1 provides a summary of the data structure.

4.1.1 Empirical survivor function

The proportion of those entering a state who survive to the first observed survival time t_1, $\widehat{S}(t_1)$, is simply one minus the proportion who made a transition out of the state by that time, where the latter can be estimated by the number of exits divided by the number who were at risk of transition: $d_1/(d_1 + m_1) = d_1/n_1$. Similarly the proportion surviving to the second observed survival time t_2 is $\widehat{S}(t_1)$ multiplied by one minus the proportion who made a transition out of the state between t_1 and t_2. More generally, at survival time t_j,

$$\widehat{S}(t_j) = \prod_{j|t_j<t} \left(1 - \frac{d_j}{n_j}\right). \tag{4.1}$$

So, the Kaplan-Meier estimate of the survivor function is given by the product of one minus the number of exits divided by the number of persons at risk of exit, i.e. the product of one minus the 'exit rate' at each of the survival times. Standard errors can be derived for the estimates using Greenwood's formulae (see texts).

From this, one can also derive an estimate of the failure function $\widehat{F}(t_j) = 1 - \widehat{S}(t_j)$ and the integrated hazard function $\widehat{H}(t_j)$, since

$$S(t) = \exp\left[-\int_0^t \theta(u)\,du\right] = \exp\left[-H(t)\right] \tag{4.2}$$

$$\Longrightarrow \quad \widehat{H}(t_j) = -\log \widehat{S}(t_j). \tag{4.3}$$

It is important to note that one can only derive estimates of the survivor function (and the integrated hazard function) at the dates at which there are transitions, and the last estimate depends on the largest non-censored survival time. To derive estimates of survival probabilities at dates beyond the maximum observed failure time, or at dates between within-sample failure times, requires additional assumptions. Reflecting this, the estimated survivor function is conventionally drawn as a step function, with the first 'step', at $t = 0$, having a height of one and width t_1, with the next 'step' beginning at time $t_1$with height $\widehat{S}(t_1)$ and width $t_2 - t_1$, and so on. The shape of this step function is not like a regular staircase: the height between steps varies (depending on the survivor function estimates), so too does the width of the steps (depending on the times at which failures were observed). To put things another way, one has a set of estimates at $\{\widehat{S}(t_j), t_j\}$ at various t_j but, to draw the survivor function, one cannot simply 'connect the dots' (as in children's drawing books). This would be making assumptions about the shape of the survivor function or, equivalently, the hazard function, that would most likely be unwarranted. This is a price that one has to pay for using a non-parametric estimator.

Indeed, the non-parametric step function nature of the survivor function and the integrated hazard function has implications for estimation of the hazard function. One might think that one could derive an estimate of the hazard rate $\widehat{\theta}(t)$ directly from the estimates of the integrated hazard function, using the property that $\partial H(t)/\partial t = \theta(t)$. To do this one would need estimates of the slope of the integrated hazard function at a series of survival times. But the estimated integrated hazard function ('cumulative hazard') is also a step function (with higher steps at longer observed survival times). Trying to estimate the slope of the integrated hazard function at each of the observed survival time is equivalent to trying to find the slope at the corner of each of the steps. Clearly, the slope is not well-defined: so, nor is a non-parametric estimate of the hazard rate.

What is one to do if an estimate of the hazard function is required? One possibility is to divide the survival time axis into a number of regular *intervals* of times and to derive estimates of the *interval* hazard rate (h rather than θ): see the discussion of life-table estimators below. Another possibility is to return to the estimated integrated hazard function, and to smooth this step function (for example using a kernel smoother). In effect, this is a statistically sophisticated method of 'connecting the dots'. See e.g. Klein and Moeschberger (1997, section 6.2) for a full discussion.

The advantage is, of course, that once one has a smooth curve, one can also derive its slope. (The estimated slope is the 'smoothed hazard'.) The disadvantage is that smoothing incorporates assumptions that need not be appropriate.

What if, for example, there are genuine step changes in survival probabilities? (Think about exit from employment into retirement: there are likely to be large – and genuine – changes in survival probabilities at the age at which people qualify for state pensions.)

An alternative to the Kaplan-Meier estimator of the survival and integrated hazard functions is the Nelson-Aalen estimator. In essence, with the Kaplan-Meier estimator, one estimates the survivor function and derives the cumulative hazard function from that; with the Nelson-Aalen estimator, it is vice-versa. The Nelson-Aalen estimator of the cumulative hazard is

$$\widehat{H}(t_j) = \sum_{j|t_j<t} \left(\frac{d_j}{n_j}\right) \tag{4.4}$$

and this provides an estimate of the survivor function which is $\exp(-\widehat{H}(t_j))$, sometimes called the Fleming-Harrington estimator. It turns out that the estimators are asymptotically equivalent (they provide the same estimates as the sample size becomes infinitely large), but have different small sample properties. The Nelson-Aalen estimator of the cumulative hazard function has better small-sample properties that the Kaplan-Meier estimator, whereas for estimation of the survivor function, the relative advantage is reversed. (In practice, with datasets with sample sizes typical in the social sciences, the difference between corresponding estimates is often negliglible.)

The **sts** command in Stata provides estimates of the Kaplan-Meier (product-limit) estimator of the survivor function and Nelson-Aalen estimator of the cumulative hazard function; **sts graph** includes an option for graphing smoothed hazards, with flexibility about the degree of smoothing (bandwidth).

4.2 Lifetable estimators

The lifetable method uses broadly the same idea as the Kaplan-Meier one, but it was explicitly developed to handle situation where the observed information about the number of exits and the number at risk are those that pertain to intervals of time, i.e. there is grouped survival time data.

Define intervals of time I_j where $j = 1, ..., J : I_j : [t_j, t_{j+1})$, where

d_j : the number of failures observed in interval I_j

m_j : the number of censored spell endings observed in interval I_j

N_j : the number at risk of failure at start of interval.

Because the survival time axis is grouped into bands, one wants to adjust the observed number of persons at risk in a given interval to take account of the fact that during the period, some people will have left. The idea is to produce an 'averaged' estimate centred on the midpoint of the interval. Suppose that transitions are evenly spread over each interval, in which case half of the total

number of events for each interval will have left half-way through the relevant interval. We can therefore define an adjusted number at risk of exit:

n_j : the adjusted number at risk of failure used for midpoint of interval

$$n_j = N_j - \frac{d_j}{2} \tag{4.5}$$

and hence

$$\widehat{S}(j) = \prod_{k=1}^{j} \left(1 - \frac{d_k}{n_k}\right). \tag{4.6}$$

From this expression can be derived an estimate of the density function (recall $S(t) = 1 - F(t)$)

$$\widehat{f}(j) = \frac{\widehat{F}(j+1) - \widehat{F}(j)}{t_{j+1} - t_j} = \frac{\widehat{S}(j) - \widehat{S}(j+1)}{t_{j+1} - t_j} \tag{4.7}$$

and an estimate of the hazard rate

$$\widehat{\theta}(j) = \frac{\left[\widehat{f}(j)\right]}{\widetilde{S}(j)} \tag{4.8}$$

where

$$\widetilde{S}(k) = \frac{\widehat{S}(k) + \widehat{S}(k+1)}{2} \tag{4.9}$$

and is taken as applying to the time corresponding to the midpoint of the interval. See Stata's **ltable** command.

Note that the 'adjustment' is used because the underlying survival times are continuous, but the observed survival time data are grouped.

If the time axis were intrinsically discrete, so that the intervals referred to basic units of time or 'cycles', then the 'adjustment' is not required. The same is true if one simply wanted to derive an estimate of the interval hazard rate. The formula for the estimator of the survivor function in this case is the same as that for the Kaplan-Meier one in the continuous time case. Observe that if there are no events within some interval I_j, then the estimator of the interval hazard is equal to zero.

** to be added: discussion of formulae for standard errors; stratification, and tests of homogeneity **

Chapter 5

Continuous time multivariate models

In this chapter and the next, we return to the problem that was raised in Chapter 1. We drew attention to problems with using OLS (and some other commonly used econometric methods) to estimate multivariate models of survival times given the issues of censoring and time-varying covariates. The principal lesson of this chapter (and the next) is that these problems may be addressed by using estimation based on *maximum likelihood* methods (ML). ML methods per se are not considered at all here – see any standard econometrics text, e.g. Greene (2003), for a discussion of the ML principle and the nice properties of ML estimates. For a very useful (and free) introduction, see chapter 1 of Gould et al. (2003), downloadable from http://www.stata-press.com/books/ml-ch1.pdf. In this chapter we consider parametric regression models formulated in continuous time; in the next chapter we consider discrete time regression models. The continuous time Cox regression model is considered in Chapter 7 (it uses an estimation principle that differs from ML).

A second lesson of this and succeeding chapters is that the sample likelihood used needs to be appropriate for the type of process that generated the data, i.e. the type of sampling scheme. The main sampling schemes in social science applications are as follows:

1. a random sample from the inflow to the state (or of the population or some other group – as long as sample selection is unrelated to survival) and each spell is monitored (from start) until completion;

2. a random sample as in (1) and either (a) each spell is monitored until some common time t^*, or (b) the censoring point varies;

3. a sample from the stock at point of time who are interviewed some time later (this corresponds to the *delayed entry* or *left truncated spell data* case discussed in the biostatistics literature);

4. a sample from the stock with no re-interview.

5. a sample from the outflow (the case of *right truncated spell data*).

Cases (3) – (5) differ from the first two in that one has to take account of 'selection effects' when characterising the sample likelihood. For continuous time data, Stata's **streg** modules handle cases (1)–(3) straightforwardly with no problem, as long as the data have been **stset** data correctly. Cases (4) and (5) require one to write one's own maximisation routines.

Let us consider the likelihood contributions relevant in each case. The likelihood contribution per spell is $\mathcal{L}_i$ and for the sample as a whole,

$$\mathcal{L} = \prod_{i=1}^{n} \mathcal{L}_i$$

or, equivalently, in terms of the log-likelihood function,

$$\log \mathcal{L} = \sum_{i=1}^{n} \log \mathcal{L}_i.$$

Software for maximization likelihood estimation typically works by requiring specification of expressions of $\log \mathcal{L}_i$ for each individual i (corresponding to a row in the analysis data set), and $\log \mathcal{L}$ is derived by summing down the rows.

5.0.1 Random sample of inflow and each spell monitored until completed

Suppose we have a sample of completed spells (or persons – since we assume one spell per person), indexed by $i = 1, \ldots, n$. Then each individual contribution to the likelihood is given by the relevant density function. (There are no censored spells, by construction.). Hence the overall sample likelihood function is given by

$$\mathcal{L} = \prod_{i=1}^{n} f\left(T_i\right) \tag{5.1}$$

where T_i denotes completed spell length for person i, and hence

$$\log \mathcal{L} = \sum_{i=1}^{n} \log f\left(T_i\right).$$

with $\log \mathcal{L}_i = \log f\left(T_i\right)$ .To complete the model specification, one has to choose a specific parametric model for the survival time distribution, e.g. Weibull, and substitute the relevant expression for into the expression for $\log f\left(T_i\right)$.

5.0.2 Random sample of inflow with (right) censoring, monitored until t^*

Now the sample consists of

- completed spells, indexed $j = 1, ..., J$, with T_j such that $T_j \leq t^*$: $\mathcal{L}_j = f(T_j)$, and
- (right) censored spells, indexed $k = 1, ..., K$, with T_k such that $T_k > t^*$: $\mathcal{L}_k = S(t^*)$

It follows that

$$\mathcal{L} = \prod_{j=1}^{J} f(T_j) \prod_{k=1}^{K} S(t^*). \tag{5.2}$$

5.0.3 Random sample of population, right censoring but censoring point varies

This is almost the same as the case just considered. It is probably the most common likelihood used in empirical applications.

$$\mathcal{L} = \prod_{j=1}^{J} f(T_j) \prod_{k=1}^{K} S(T_k) \tag{5.3}$$

The likelihood $\mathcal{L}$ is often written differently in this case, in terms of the hazard rate. Taking logs of both sides implies:

$$\begin{aligned}
\log \mathcal{L} &= \sum_{j=1}^{J} \log f(T_j) + \sum_{k=1}^{K} \log S(T_k) && (5.4)\\
&= \sum_{j=1}^{J} \log \left[\left(\frac{f(T_j)}{S(T_j)} \right) S(T_j) \right] + \sum_{k=1}^{K} \log S(T_k) && (5.5)\\
&= \sum_{j=1}^{J} \log \left[\theta(T_j) S(T_j) \right] + \sum_{k=1}^{K} \log S(T_k) && (5.6)\\
&= \sum_{j=1}^{J} \log \ \theta(T_j) + \sum_{j=1}^{J} \log S(T_j) + \sum_{k=1}^{K} \log S(T_k) && (5.7)\\
&= \sum_{j=1}^{J} \log \theta(T_j) + \sum_{i=1}^{N} \log S(T_i) && (5.8)\\
&= \sum_{i=1}^{N} \left[c_i \log \theta(T_i) + \log S(T_i) \right] && (5.9)
\end{aligned}$$

where c_i is a censoring indicator defined such that

$$c_i = \begin{cases} 1 \text{ if spell complete} \\ 0 \text{ if spell censored.} \end{cases} \tag{5.10}$$

In this case, we have

$$\log \mathcal{L}_i = c_i \log \theta(T_i) + \log S(T_i).$$

Given the relationship between the survivor function and the integrated hazard function, one can also write the log-likelihood contribution for each individual as

$$\begin{aligned} \log \mathcal{L}_i &= c_i \log \theta(T_i) - H(T_i) \\ &= c_i \log \theta(T_i) - \int_0^{T_i} \theta(u) du. \end{aligned} \tag{5.11}$$

Again, it is by choice of different parametric specifications for $\theta(t)$ that the model is fully specified.

If all survival times are censored ($c_i = 0$, for all i), then the model cannot be fitted. If there are no exits, then the sample provides no information about the nature of duration dependence in the hazard rate.

5.0.4 Left truncated spell data (delayed entry)

The most common social science example of this type of sampling scheme is when there is a sample from the stock of individuals at a point in time, who are interviewed some time later ('stock sampling with follow-up'). Spell start dates are assumed to be known (these dates are of course before the date of the stock sample), so the total time at risk of exit can be calculated, together with the time between sampling and interview (last observation). Entry is 'delayed' because the observation of the subjects under study occurs some time after they are first at risk of the event.

A sample from the stock is a non-random sample (see below), but we can handle the 'selection bias' using information about the elapsed time between sampling and interview. We have to analyse outcomes that have occurred by the time of interview *conditional* on surviving in the state up to the sampling time.

Insert chart time line indicating sampling time and interview times

Let us index spells that were completed by the time of the interview by $j = 1, ..., J$, and index spells still in progress at the time of the interview (right-censored) by $k = 1, ..., K$. Then we may define

the incomplete spell length for each person at the time that the spell was sampled from the stock: T_j for completed spells, T_k for censored spells.

the total observed spell length is $T_j + \Delta t_j$ for spells that were completed by the time of the interview, and $T_k + Z_k$ for censored spells, where Z_k is the length of time between sampling and interview.

For both completed and censored spells, we have to condition on the fact that the person survived sufficiently long in the state to be at risk of being sampled in the stock. For example, suppose that the sample was of the stock of unemployed persons at 1 May 2001. Of all those who entered unemployment on 1 January 2001, some will have left unemployment by May; only the 'slower' exiters will still be unemployed in May and at risk of being sampled in the stock. If we ignored this problem, we would not take account of the length-biased sampling. How do we do this?

Recall the expression for a conditional probability:

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}. \tag{5.12}$$

By analogy, we deflate the likelihood contribution for each individual by the probability of survival from entry to the state until the stock sampling date. But this probability is given for each i by the survival function: $S(T_i)$. Hence,

Likelihood contribution for leavers (type j):

$$\mathcal{L}_j = \frac{f\left(T_j + \Delta t_j\right)}{S\left(T_j\right)}. \tag{5.13}$$

Likelihood for stayers (type k):

$$\mathcal{L}_k = \frac{S(T_k + Z_k)}{S\left(T_k\right)}. \tag{5.14}$$

Hence the overall sample likelihood is

$$\mathcal{L} = \prod_{j=1}^{J} \frac{f\left(T_j + \Delta t_j\right)}{S\left(T_j\right)} \prod_{k=1}^{K} \frac{S\left(T_k + Z_k\right)}{S\left(T_k\right)}. \tag{5.15}$$

To write this expression in a form closer to that used in the previous subsection, let us simply define T_i as the total spell length, and use the censoring indicator c_i to distinguish between censored and complete spells, and let τ_i be the date at which the stock sample was drawn.Then

$$\mathcal{L} = \prod_{i=1}^{N} \left[\frac{f\left(T_i\right)}{S\left(\tau_i\right)}\right]^{c_i} \left[\frac{S\left(T_i\right)}{S\left(\tau_i\right)}\right]^{1-c_i} \tag{5.16}$$

$$= \prod_{i=1}^{N} \left[\theta\left(T_i\right)\right]^{c_i} \left[\frac{S\left(T_i\right)}{S\left(\tau_i\right)}\right] \tag{5.17}$$

or

$$\log \mathcal{L} = \sum_{i=1}^{N} \left\{ c_i \log \theta(T_i) + \log \left[\frac{S\left(T_i\right)}{S\left(\tau_i\right)}\right] \right\} \tag{5.18}$$

which is directly comparable with the (log)likelihood derived for the case without left-truncation. This expression can be re-written as

$$\log \mathcal{L} = \sum_{i=1}^{N} \{c_i \log \theta(T_i) + \log [w_i S(T_i)]\} \tag{5.19}$$

where $w_i \equiv 1/S(\tau_i)$. Think of the w_i as being like an weighting variable: one weights the delayed entry observations by a type of inverse-probability weight to account for the left truncation. The later in time that τ_i is (the closer to T_i), the larger the weight. If there is no left truncation, then $S(\tau_i) = 1 = w_i$.

This model can also be estimated straightforwardly in Stata using **streg** as long as the data have been properly **stset** first (use the **enter** option to indicate the 'entry' time, i.e. stock sampling date).

Lancaster (1979) also considered a variation on this sampling scheme. In his data set, he did not have the exact date of exit for those spells that ended in the interval between stock sampling and interview: he only knew that there had been an exit. To estimate a model with this structure, one simply replaces the expression $f(T_j + \Delta t_j)$ in (5.15) with $S(T_j) - S(T_j + \Delta t_j)$.

5.0.5 Sample from stock with no re-interview

This is a difficult and awkward case. It used to be relatively common because of a lack of suitable longitudinal surveys. This model could be fitted using data from standard cross-sectional household surveys: these surveys include samples of unemployed people, and typically include a single question asking how long the respondent had been unemployed. See Nickell (1979) and Atkinson *et al.* (1984) for applications.

In this situation we have no information that we can use to condition survival on (as in the previous subsection). So, in order to derive the sample likelihood, we have to write down the probability of observing such a spell including taking account of the different chances of entering unemployment at different dates. Someone unemployed when surveyed at the start of October 2001 could have entered unemployment on 1 January and remained unemployed 9 months, or entered unemployment on 1 February and remained unemployed 8 months, or unemployment on 1 March and remained unemployed 7 months. (Similarly for the unemployed people who were surveyed during a different month.)

We want to model the chance of having an incomplete spell of length t at calendar time s, conditional on being unemployed at date s – where the chance of the latter itself depends on the chances of having entered unemployment at some date in the past, r, and then remaining unemployed between r and s.

Distinguish a series of different event types:

A : incomplete spell of length t at calendar time s

B : unemployed at date s (may differ by person – be s_i – if from survey with interviews through year)

C : survival of spell until time s given entry at time r (i.e. spell length $t = s-r$)

D : entry to unemployment at date r.

$$\begin{aligned}
\mathcal{L}_i &= \Pr(A|B) && (5.20)\\
&= \Pr(A \cap B) / \Pr(B) && (5.21)\\
&= \Pr(A) / \Pr(B) && (5.22)\\
&= \frac{\Pr(C)\Pr(D)}{\Pr(B)}, \text{ since } P(A) = \Pr(C)\Pr(D) && (5.23)
\end{aligned}$$

What are the expressions for the component probabilities? Look first at:

$$\Pr(B) = \sum_{\tau=-\infty}^{s} S_i(s-\tau|\text{entry at } \tau)\, u_i(\tau). \tag{5.24}$$

This says that the probability of being unemployed at s is the sum over all dates before the interview (indexed by τ) of the product of the probability of entering unemployment at date τ and the probability of remaining unemployed between τ and s.

We now need to make an assumption about unemployment inflow rates in order to evaluate $\Pr(D)$. Assume that

$$\Pr(D) = u_i(r) = \kappa_i u(r), \tag{5.25}$$

i.e. the probability of entry at r factors into an individual fixed effect (this varies with i but not time) and a function depending on time but not i (i.e. common to all persons).

$$\begin{aligned}
\mathcal{L}_i &= \frac{S(s_i|\text{entry at } r_i)\kappa_i u(r_i)}{\sum_{\tau=-\infty}^{s_i} S(s_i - \tau|\text{entry at } \tau)\kappa_i u(\tau)} && (5.26)\\
&= \frac{S(s_i|\text{entry at } r_i) u(r_i)}{\sum_{\tau=-\infty}^{s_i} S(s_i - \tau|\text{entry at } \tau) u(\tau)}. && (5.27)
\end{aligned}$$

Observe the cancelling of the individual fixed effects. Implementing the model one can use 'external' data about unemployment entry rates for the $u(.)$ term. This likelihood cannot be fitted using a canned routine in standard packages and has to be specially written (tricky!).

5.0.6 Right truncated spell data (outflow sample)

An example of this sampling scheme is where one has a sample of all the persons who left unemployment at a particular date and one wants to study the hazard of (re)employment. Or one wants to study longevity, and has a sample based

on death records. In these sorts of cases, there is an over-representation of short spells relative to long spells. Of all people beginning a spell at a particular date, those who are more likely to survive (have relatively long spells) are less likely to be found in the outflow at a particular date – a form of 'selection bias'. To control for this, one has to condition on failure at the relevant survival time when specifying each individual's likelihood contribution.

The sample likelihood is given by

$$\mathcal{L} = \prod_{i=1}^{n} \frac{f(T_i)}{F(T_i)}. \tag{5.28}$$

There are no censored spells, of course, by definition. The numerator in (5.28) is the density function for a spell of length T_i. This must be conditioned: hence the expression in the denominator in (5.28) which gives the probability of failure at $t = T_i$, i.e. the probability of entering at $t = 0$, surviving up until the instant just before time $t = T_i$, and then making the transition at $t = T_i$.

5.1 Episode splitting: time-varying covariates and estimation of continuous time models

To illustrate this, let us return to the most commonly assumed sampling scheme, i.e. a random sample of spells with right censoring but the censoring point varies. In our sample likelihood derivation above, we implicitly assumed

- explanatory variables were all constant – there were no time-varying covariates;
- the data set was organised so that there was one row for each individual at risk of transition.

Estimation of continuous time parametric regression models incorporating time-varying covariates requires *episode splitting*. One has to *split* the survival time (*episode*) for each individual into subperiods within which each time-varying covariate is constant. I.e. one has to create multiple records for each individual, with one record per subperiod.[1] What is the logic behind this?

Consider a person i with two different values for a covariate:

$$\begin{array}{l} X_1 \text{ if } t < u \\ X_2 \text{ if } t \geq u \end{array} \tag{5.29}$$

Recall that the log-likelihood contribution for person i in the data structure that we have is:

[1] Episode splitting to incorporate time-varying covariates is also required for other types of model. For the Cox model it need only be done at the survival times at which transitions occur (see Chapter 7) and, for discrete-time models, episode splitting is done at each time point (see Chapter 6).

Record #	**Censoring indicator**	**Survival time**	**Entry time**	**TVC value**
Single data record for i				
1	$c_i = 0$ or 1	T_i	0	-
Multiple data records for i (after episode splitting)				
1	$c_i = 0$	u	0	X_1
2	$c_i = 0$ or 1	T_i	u	X_2

Table 5.1: Example of episode splitting

$$\log \mathcal{L}_i = c_i \log \left[\theta \left(T_i\right)\right] + \log \left[S \left(T_i\right)\right] \tag{5.30}$$

where i's observed survival time is T_i and the censoring indicator $c_i = 1$ if i's spell is complete (transition observed) and 0 if the spell is censored. But

$$\begin{aligned} \log \left[S \left(T_i\right)\right] &= \log \left[S \left(u\right) \frac{S \left(T_i\right)}{S \left(u\right)}\right] &(5.31)\\ &= \log \left[S \left(u\right)\right] + \log \left[\frac{S \left(T_i\right)}{S \left(u\right)}\right]. &(5.32) \end{aligned}$$

(This expression, incorporating some notational liberties to facilitate exposition, follows directly from the discussion about the incorporation of time-varying covariates into PH and AFT models in Chapter 3.) Thus the log of the probability of survival until $T =$ (log of probability of survival to time u) + (log of probability of survival to T_i, *conditional on entry at* u).

So what we do is create one new record with $c_i = 0$, $t = u$ (a right censored episode), plus one new record summarising an episode with 'delayed entry' at time u and censoring indicator c_i has the value as in the original data. In the first episode and record, the time-varying covariate takes on the value X_1 and in the second record the time-varying covariate takes on the value X_2. See Table 5.1 for a summary of the old and new data structures.

Thus episode splitting (when combined with software that can handle left truncated spell data) gives the correct log-likelihood contribution. In Stata, episode splitting is easily accomplished using **stsplit** (applied after the data have been **stset**, and which then cleverly updates **stset**). Then one creates the appropriate time-varying covariate values for each record and, finally, multivariate continuous time models can be straightforwardly estimated using the **streg** or **stcox** commands.

Chapter 6

Discrete time multivariate models

6.1 Inflow sample with right censoring

We now measure time in discrete intervals indexed by the positive integers, and let us suppose that each interval is a month long. We observe a person i's spell from month $k = 1$ through to the end of the j^{th} month, at which point i's spell is either complete ($c_i = 1$), or right censored ($c_i = 0$). The discrete hazard is

$$h_{ij} = \Pr\left(T_i = j | T_i \geq j\right). \tag{6.1}$$

The likelihood contribution for a censored spell is given by the discrete time survivor function

$$\begin{aligned} \mathcal{L}_i &= \Pr(T_i > j) = S_i(j) && (6.2)\\ &= \prod_{k=1}^{j}(1 - h_{ik}) && (6.3) \end{aligned}$$

and the likelihood contribution for each completed spell is given by the discrete time density function:

$$\begin{aligned} \mathcal{L}_i &= \Pr(T_i = j) = f_i(j) && (6.4)\\ &= h_{ij} S_i(j-1) && (6.5)\\ &= \frac{h_{ij}}{1 - h_{ij}} \prod_{k=1}^{j}(1 - h_{ik}) && (6.6) \end{aligned}$$

The likelihood for the whole sample is

$$\mathcal{L} = \prod_{i=1}^{n} \left[\Pr(T_i = j)\right]^{c_i} \left[\Pr\left(T_i > j\right)\right]^{1-c_i} \tag{6.7}$$

$$= \prod_{i=1}^{n} \left[\left(\frac{h_{ij}}{1-h_{ij}}\right) \prod_{k=1}^{j} (1-h_{ik})\right]^{c_i} \left[\prod_{k=1}^{j} (1-h_{ik})\right]^{1-c_i} \tag{6.8}$$

$$= \prod_{i=1}^{n} \left[\left(\frac{h_{ij}}{1-h_{ij}}\right)^{c_i} \prod_{k=1}^{j} (1-h_{ik})\right] \tag{6.9}$$

where c_i is a censoring indicator defined such that $c_i = 1$ if a spell is complete and $c_i = 0$ if a spell is right-censored (as in the previous chapter). This implies that

$$\log \mathcal{L} = \sum_{i=1}^{n} c_i \log\left(\frac{h_{ij}}{1-h_{ij}}\right) + \sum_{i=1}^{n}\sum_{k=1}^{j} \log\left(1-h_{ik}\right). \tag{6.10}$$

Now define a new binary indicator variable $y_{ik} = 1$ if person i makes a transition (their spell ends) in month k, and $y_{ik} = 0$ otherwise. That is,

$$c_i = 1 \Longrightarrow y_{ik} = 1 \text{ for } k = T_i,\ y_{ik} = 0 \text{ otherwise} \tag{6.11}$$

$$c_i = 0 \Longrightarrow y_{ik} = 0 \text{ for all } k \tag{6.12}$$

Hence, we can write

$$\log \mathcal{L} = \sum_{i=1}^{n}\sum_{k=1}^{j} y_{ik} \log\left(\frac{h_{ik}}{1-h_{ik}}\right) + \sum_{i=1}^{n}\sum_{k=1}^{j} \log\left(1-h_{ik}\right) \tag{6.13}$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{j} \left[y_{ik} \log h_{ik} + (1-y_{ik}) \log\left(1-h_{ik}\right)\right]. \tag{6.14}$$

But this expression has exactly the same form as the standard likelihood function for a binary regression model in which y_{ik} is the dependent variable *and* in which the data structure has been reorganized from having one record per spell to having one record for each month that a person is at risk of transition from the state (so-called person-month data or, more generally, *person-period data*).

Table 6.1 summarises the two data structures.

This is just like the episode splitting described earlier for continuous time models which also lead to the creation of data records, except that here there is episode splitting on a much more extensive basis, and done regardless of whether there are any time-varying covariates among the X. As long as the hazard is not constant, then the variable summarizing the pattern of duration dependence will itself be a time-varying covariate in this re-organised data set.

This result implies that there is an *easy estimation method* available for discrete time hazard models using data with this sampling scheme (see *inter alia* Allison, 1984; Jenkins, 1995). It has four steps:

Person data			Person-month data				
person id, i	c_i	T_i	person id, i	c_i	T_i	y_{ik}	person-month id, k
1	0	2	**1**	**0**	**2**	**0**	**1**
			1	**0**	**2**	**0**	**2**
2	1	3	**2**	**1**	**3**	**0**	**1**
⋮	⋮	⋮	**2**	**1**	**3**	**0**	**2**
			2	**1**	**3**	**1**	**3**
			⋮	⋮	⋮	⋮	⋮

Table 6.1: Person and person-period data structures: example

1. Reorganize data into person-period format;
2. Create any time-varying covariates – at the very least this includes a variable describing duration dependence in the hazard rate (see the discussion in Chapter 3 of alternative specifications for the γ_j terms);
3. Choose the functional form for h_{ik} (logistic or cloglog);
4. Estimate the model using any standard binary dependent regression package: logit for a logistic model; cloglog for complementary log-log model.

The easy estimation method is not the only way of estimating the model. In principle, one could estimate the sequence likelihood given at the start without reorganising the data. Indeed before the advent of cheap computer memory or storage, one might have had to do this because the episode splitting required for the easy estimation method could create very large data sets with infeasible storage requirements.

That the likelihood for discrete time hazard model can be written in the same form as the likelihood for a binary dependent model also has some other implications. For example, the latter models can only be estimated if the data contains both 'successes' and 'failures' in the binary dependent variable. In the current context, it means that in order to estimate discrete time hazard models in this way, the data must contain both censored and complete spells. (To see this, look at first order condition for a maximum of the log-likelihood function.)

The easy estimation method also can be used when there is stock sample with follow-up ('delayed entry'), as we shall now see.

6.2 Left-truncated spell data ('delayed entry')

We proceed in the discrete time case in an analogous manner to that employed with continuous time data. Recall that with no delayed entry

$$\mathcal{L}_i = \left(\frac{h_{ij}}{1-h_{ij}}\right)^{c_i} \prod_{k=1}^{j} (1-h_{ik}) = \left(\frac{h_{ij}}{1-h_{ij}}\right)^{c_i} S_i(j). \tag{6.15}$$

With delayed entry at time u_i (say) for person i, we have to condition on survival up to time u_i (corresponding to the end of the u_ith interval or cycle), which means dividing the expression above by $S(u_i)$. Hence with left-truncated data, the likelihood contribution for i is:

$$\mathcal{L}_i = \frac{\left(\frac{h_{ij}}{1-h_{ij}}\right)^{c_i} \prod_{k=1}^{j} (1-h_{ik})}{S_i(u_i)}. \tag{6.16}$$

But

$$S_i(u_i) = \prod_{k=1}^{u_i} (1-h_{ik}) \tag{6.17}$$

and this leads to a 'convenient cancelling' result (Guo, 1993; Jenkins, 1995):

$$\mathcal{L}_i = \left(\frac{h_{ij}}{1-h_{ij}}\right)^{c_i} \left[\frac{\prod_{k=1}^{j} (1-h_{ik})}{\prod_{k=1}^{u_i} (1-h_{ik})}\right] \tag{6.18}$$

$$= \left(\frac{h_{ij}}{1-h_{ij}}\right)^{c_i} \prod_{k=u_i+1}^{j} (1-h_{ik}). \tag{6.19}$$

Taking logarithms, we have

$$\log \mathcal{L}_i = \sum_{k=u_i+1}^{j} [y_{ik} \log h_{ik} + (1-y_{ik}) \log (1-h_{ik})] \tag{6.20}$$

which is very similar to the expression that we had in the no-delayed-entry case, except that the summation now runs over the months from the month of 'delayed entry' (e.g. when the stock sample was drawn) to the month when last observed.

Implementation of the easy estimation method using data with this sampling scheme now has four steps (Jenkins, 1995):

1. Reorganize data into person-period format;
2. Throw away the records for all the periods prior to the time of 'delayed entry' (the months up to and including u in this case), retaining the months during which each individual is observed at risk of experiencing the event;
3. Create any time-varying covariates (at the very least this includes a variable describing duration dependence in the hazard rate – see above);
4. Choose the functional form for h_{ik} (logistic or cloglog);
5. Estimate the model using any standard binary dependent regression package: logit for a logistic model; cloglog for complementary log-log model.

The only difference from before is step 2 (throwing data away). Actually, in practice, steps one and two might be combined. If one has a panel data set, for instance, one can create a data set with the correct structure directly.

6.3 Right-truncated spell data (outflow sample)

Again the specification for the sample likelihood closely parallels that for the continuous time case. There are no censored cases; all spells are completed, by construction. But one has to account for the selection bias arising from outflow sampling by conditioning each individual's likelihood contribution on failure at the observed survival time. Each spell's contribution to the sample likelihood is given by the discrete density function, $f(j)$, divided by the discrete time failure function, $F(j) = 1 - S(j)$. Hence i's contribution to sample likelihood is:

$$\mathcal{L}_i = \frac{\left(\frac{h_{ij}}{1-h_{ij}}\right) \prod_{k=1}^{j} (1 - h_{ik})}{1 - \left[\prod_{k=1}^{j} (1 - h_{ik})\right]}. \tag{6.21}$$

Unfortunately the likelihood function does not conveniently simplify in this case (as it did in the left-truncated spell data case).

Chapter 7

Cox's proportional hazard model

This model, proposed by Cox (1972), is perhaps the most-often cited article in survival analysis. The distinguishing feature of Cox's proportional hazard model, sometimes simply referred to as the 'Cox model', is its demonstration that one could estimate the relationship between the hazard rate and explanatory variables without having to make any assumptions about the shape of the baseline hazard function (cf. the parametric models considered earlier). Hence the Cox model is sometimes referred to as a semi-parametric model. The result derives from innovative use of the proportional hazard assumption together with several other insights and assumptions, and a *partial likelihood (PL)* method of estimation rather than maximum likelihood. Here follows an intuitive demonstration of how the model works, based on the explanation given by Allison (1984).

We are working in continuous time, and suppose that we have a random sample of spells, some of which are censored and some are complete. (The models can be estimated using left-truncated data, but we shall not consider that case here.)

Recall the PH specification

$$\begin{aligned} \theta(t, X_i) &= \theta_0(t) \exp\left(\beta' X_i\right) && (7.1) \\ &= \theta_0(t) \lambda_i. && (7.2) \end{aligned}$$

or, equivalently,

$$\theta(t, X_i) = \theta_0^*(t) \lambda_i^*. \quad (7.3)$$

We shall suppose, for the moment, that the X vector is constant, but note that the Cox model can in fact also handle time-varying covariates.

Cox (1972) proposed a method for estimating the slope coefficients in β (i.e. excluding the intercept) without having to specify any functional form for the

Person # i	Time t_i	Event # k
1	2	1
2	4	2
3	5	3
4	5*	
5	6	4
6	9*	
7	11	5
8	12*	

*: censored observation

Table 7.1: Data structure for Cox model: example

baseline hazard function, using the method of PL. PL works in terms of the *ordering of events* by constrast with the focus in ML on persons (spells).

Consider the illustrative data set shown in Table 7.1. We assume that there is a maximum of one event at each possible survival time (this rules out ties in survival times – for which there are various standard ways of adapting the Cox model). In the table persons are arranged in order of survival times.

The sample Partial Likelihood is given by

$$PL = \prod_{k=1}^{K} \mathcal{L}_k \tag{7.4}$$

This is quite different from the sample likelihood expressions we had before in the maximum likelihood case. The k indexes *events*, not persons. But what then is each $\mathcal{L}_k$? Think of it as the following:

$$\begin{aligned} L_k &= \Pr(\text{person } i \text{ has event at } t = t_i \text{ conditional on being in the risk set at } t = t_i) \quad (7.5) \\ &= \Pr(\text{this } \textit{particular} \text{ person } i \text{ experiences the event at } t = t_i, \\ &\quad \text{given that one observation amongst many at risk experiences the event}) \end{aligned}$$

Or, of the people at risk of experiencing the event, which one in fact does experience it? To work out this probability, use the rules of conditional probability together with the result that the probability density for survival times is the product of the hazard rate and survival function, i.e. $f(t) = \theta(t)S(t)$, and so the probability that an event occurs in the tiny interval $[t, t + \Delta t)$ is $f(t)\,dt = \theta(t)S(t)dt$.

Consider event $k = 5$ with risk set $i \in \{7, 8\}$. We can define

$$A = \Pr \text{ (event experienced by } i = 7 \text{ and not } i = 8) = [\theta_7(11)S_7(11)dt]\,[S_8(11)]$$

$$B = \Pr \text{ (event experienced by } i = 8 \text{ and not } i = 7) = [\theta_8(11)S_8(11)dt]\,[S_7(11)]$$

Now consider the expression for probability A conditional on the probability of either A or B (the chance that either could have experienced event, which is a sum of probabilities). Using the standard conditional probability formula, we find that

$$\mathcal{L}_5 = \frac{A}{A+B} = \frac{\theta_7(11)}{\theta_7(11) + \theta_8(11)} \tag{7.6}$$

Note that the survivor function terms cancel. This same idea can be applied to derive all the other $\mathcal{L}_k$. For example,

$$\mathcal{L}_1 = \frac{\theta_1(2)}{\theta_1(2) + \theta_2(2) + ... + \theta_8(2)}. \tag{7.7}$$

Everyone is in the risk set for the first event.

Now let us apply the PH assumption $\theta(t, X_i) = \theta_0(t)\lambda_i$, starting for illustration with event $k = 5$ with associated risk set $i \in \{7, 8\}$. We can substitute into the expression for $\mathcal{L}_5$ above, and find that

$$\mathcal{L}_5 = \frac{\theta_0(11)\lambda_7}{\theta_0(11)\lambda_7 + \theta_0(11)\lambda_8} \tag{7.8}$$

$$= \frac{\lambda_7}{\lambda_7 + \lambda_8}. \tag{7.9}$$

The baseline hazard contributions cancel. (The intercept term, which is common to all, $\exp(\beta_0)$, also cancels from both the numerator and the denominator since it appears in each of the λ terms.) By similar arguments,

$$\mathcal{L}_1 = \frac{\lambda_1}{\lambda_1 + \lambda_2 + ... + \lambda_8} \tag{7.10}$$

and so on. Given each $\mathcal{L}_k$ expression, one can construct the complete PL expression for the whole sample of events, and then maximize it to derive estimates of the slope coefficients within β. It has been shown that these estimates have 'nice' properties.

Note that the baseline hazard function is completely unspecified, which can be seen as a great advantage (one avoids potential problems from specifying the wrong shape), but some may also see it as a disadvantage if one is particularly interested in the shape of the baseline hazard function for its own sake.

One can derive an estimate of the baseline survivor and cumulative hazard functions (and thence of the baseline hazard using methods analogous to the Kaplan-Meier procedure discussed earlier – with the same issues arising).

If there are tied survival times, then the neat results above do not hold exactly. In this case, either one has to use various approximations (several standard ones are built into software packages by default), or modify the expressions to derive the 'exact' partial likelihood – though this may increase computational time substantially. (It turns out that the specification in the latter case is closely related to the conditional logit model.) On the other hand, if one finds that the

incidence of tied survival times is relatively high in one's data set, then perhaps one should ask whether a continuous time model is suitable. One could instead apply the discrete time proportional hazards model.

The Cox PL model can incorporate time-varying covariates. In empirical implementation, one uses episode splitting again. By constrast with the approaches to this employed in the last two chapters, observe that now the splitting need only be done at the failure times. This is because the PL estimates are derived only using the information about the risk pool at each failure time. Thus covariates are only 'evaluated' during the estimation at the failure times, and it does not matter what happens to their values in between.

Finally observe that the expression for each $\mathcal{L}_k$ does not depend on the precise survival time at which the kth event occurs. Only the order of events affects the PL expression. Check this yourself: multiply all the survival times in the table by two and repeat the derivations (you should find that the estimates of the slope coefficients in β are the same). What is the intuition? Remember that the PH assumption means implies that the hazard function for two different individuals has the same shape, differing only by a constant multiplicative scaling factor that does not vary with survival time. To estimate that constant scaling factor, given the common shape, one does not need the exact survival times.

Chapter 8

Unobserved heterogeneity ('frailty')

In the multivariate models considered so far, all differences between individuals were assumed to be captured using observed explanatory variables (the X vector). We now consider generalisations of the earlier models to allow for unobserved individual effects. There are usually referred to as 'frailty' in the bio-medical sciences. (If one is modelling human survival times, then frailty is an unobserved propensity to experience an adverse health event.) There are several reasons why these variables might be relevant. For example:

- omitted variables (unobserved in the available data, or intrinsically unobservable such as 'ability')
- measurement errors in observed survival times or regressors (see Lancaster, 1990, chapter 4).

What if the effects are important but 'ignored' in modelling? The literature suggests several findings:

- The 'no-frailty' model will over-estimate the degree of negative duration dependence in the hazard (i.e. under-estimate the degree of positive duration dependence); * Insert figure *
- The proportionate response of the hazard rate to a change in a regressor k is no longer constant (it was given by β_k in the models without unobserved heterogeneity), but declines with time;
- one gets an under-estimate of the true proportionate response of the hazard to a change in a regressor k from the no-frailty-model β_k.

Let us now look at these results in more detail.

8.1 Continuous time case

For convenience we suppress the subscript indexing individuals, and assume for now that there are no time-varying covariates. We consider the model

$$\theta\left(t, X \mid v\right) = v\theta\left(t, X\right) \tag{8.1}$$

where

$\theta(t, X)$ is the hazard rate depending on observable characteristics X, and

v is an unobservable individual effect that scales the no-frailty component. Random variable v is assumed to have the following properties:

- $v > 0$
- $\mathcal{E}(v) = 1$ (unit mean, a normalisation required for identification)
- finite variance $\sigma^2 > 0$, and is
- distributed independently of t and X.

This model is sometimes referred to as a 'mixture' model – think of the two components being 'mixed' together – or as a 'mixed proportional hazard' model (MPH). It can be shown, using the standard relationship between the hazard rate and survivor function (see chapter 3), that the relationship between the frailty survivor function and the no-frailty survivor function is

$$S\left(t, X \mid v\right) = \left[S\left(t, X\right)\right]^v \tag{8.2}$$

Thus the individual effect v scales no-frailty component survivor function. Individuals with above-average values of v leave relatively fast (their hazard rate is higher, other things being equal, and their survival times are smaller), and the opposite occurs for individuals with below-average values of v.

If the no-frailty hazard component has Proportional Hazards form, then:

$$\begin{aligned}
\theta\left(t, X\right) &= \theta_0\left(t\right) e^{\beta' X} && (8.3)\\
\theta\left(t, X \mid v\right) &= v\theta_0\left(t\right) e^{\beta' X} && (8.4)\\
\log\left[\theta\left(t, X \mid v\right)\right] &= \log\theta_0\left(t\right) + \beta' X + u && (8.5)
\end{aligned}$$

where $u \equiv \log\left(v\right)$ and $\mathcal{E}(u) = \phi$. In the no-frailty model, the log of the hazard rate at each survival time t equals the log of the baseline hazard rate at t (common to all persons) plus an additive individual-specific component ($\beta' X$). The frailty model for the log-hazard adds an additional additive 'error' term (u). Alternatively, think of this as a random intercept model: the intercept is $\beta_0 + u$.

How does one estimate frailty models, given that the individual effect is unobserved? Clearly we cannot estimate the values of v themselves since, by

construction, they are unobserved. Put another way, there are as many individual effects as individuals in the data set, and there are not enough degrees of freedom left to fit these parameters. However if we suppose the distribution of v has a shape whose functional form is summarised in terms of only a few key parameters, then we can estimate those parameters with the data available. The steps are as follows:

- Specify a distribution for the random variable v, where this distribution has a particular parametric functional form (e.g. summarising the variance of v).
- Write the likelihood function so that it refers to the distributional parameter(s) (rather than each v), otherwise known as 'integrating out' the random individual effect.

This means that one works with some survivor function

$$S_v(t, X) = S(t, X|\beta, \sigma^2), \tag{8.6}$$

and not $S(t, X|\beta, v)$. Then

$$S_v(t, X) = \int_0^\infty [S(t, X)]^v \, g(v) \, dv \tag{8.7}$$

where $g(v)$ is the probability density function (pdf) for v. The $g(v)$ is the 'mixing' distribution. But what shape is appropriate to use for the distribution of v (among the potential candidates satisfying the assumptions given earlier)?

The most commonly used specification for the mixing distribution is the *Gamma* distribution, with unit mean and variance σ^2. Making the relevant substitutions into $S_v(t, X)$ and evaluating the integral implies a specific functional form for the *frailty survivor function:*

$$S\left(t, X \mid \beta, \sigma^2\right) = \left[1 - \sigma^2 \ln S(t, X)\right]^{-(1/\sigma^2)} \tag{8.8}$$

$$= \left[1 + \sigma^2 H(t, X)\right]^{-(1/\sigma^2)} \tag{8.9}$$

where $S(t, X)$ is the no-frailty survivor function, and using the property that the integrated hazard function $H(t, X) = -\ln S(t, X)$. (Note that $\lim_{\sigma^2 \to 0} S\left(t, X \mid \beta, \sigma^2\right) = S(t, X)$.)

Now consider what this implies if the no-frailty part follows a *Weibull* model, in which case: $S(t) = \exp(-\lambda t^\alpha)$, $H(t) = \lambda t^\alpha$, $\lambda = \exp(\beta' X)$, and so

$$S\left(t, X \mid \beta, \sigma^2\right) = \left[1 + \sigma^2 \lambda t^\alpha\right]^{-(1/\sigma^2)} \tag{8.10}$$

for which one may calculate that the median survival time is given by

$$m_v = \left(\frac{(2^{\sigma^2} - 1)}{\sigma^2 \lambda}\right)^{\frac{1}{\alpha}}. \tag{8.11}$$

This may be compared with the expression for the median in the no-frailty case for the Weibull model, $[(\log(2)/\lambda]^{1/\alpha}$, which is the corresponding expression as $v \to 0$.

This survivor function (and median) is an unconditional one, i.e. not conditioning on a value of v (that has been integrated out). An alternative approach is to derive the survivor function (and its quantiles) for specific values of v (these are conditional survivor functions). Of the specific values, the most obvious choice is $v = 1$ (the mean value), but other values such as the upper or lower quartiles could also be used. (The estimates of the quartiles of the heterogeneity distribution can be derived using the estimates of σ^2.)

An alternative mixing distribution to the Gamma distribution is the *Inverse Gaussian* distribution. This is less commonly used (but available in Stata along with the Gamma mixture model). For further discussion, see e.g. Lancaster (1990), the Stata manuals under **streg**, or EC968 Web Lesson 8.

8.2 Discrete time case

Recall that in the continuous time proportional hazard model, then the log of the frailty hazard is:

$$\log\left[\theta\left(j, X \mid v\right)\right] = \log\left[\theta_0\left(j\right)\right] + \beta' X + u. \tag{8.12}$$

Recall too the discrete time model for the situation where survival times are grouped (leading to the cloglog model). By the same arguments as those, one can have a discrete time PH model with unobserved heterogeneity:

$$\operatorname{cloglog}\left[h\left(j, X \mid v\right)\right] = D\left(j\right) + \beta' X + u \tag{8.13}$$

where $u \equiv \log\left(v\right)$, as above. $D\left(j\right)$ characterises the baseline hazard function. If v has a Gamma distribution, as proposed by Meyer (1990), there is a closed form expression for the frailty survivor function: see (8.9) above. If we have an inflow sample with right censoring, the contribution to the sample likelihood for a censored observation with spell length j intervals is $S\left(j, X \mid \beta, \sigma^2\right)$, and contribution of someone who makes a transition in the jth interval is $S\left(j-1, X \mid \beta, \sigma^2\right) - S\left(j, X \mid \beta, \sigma^2\right)$, with the appropriate substitutions made. This model can be estimated using my Stata program **pgmhaz8** (or **pgmhaz** for users of Stata versions 5–7). The data should be organized in person-period form, as discussed in the previous chapter, and time-varying covariates may be incorporated.

Alternatively, one may suppose that u has a Normal distribution with mean zero. In this case, there is no convenient closed form expression for the survivor function and hence likelihood contributions: the 'integrating out' must be done numerically. Estimation may be done using the built-in Stata program **xtcloglog**, using data organized in person-period form.

For the logit model, one might suppose now that log odds of the hazard takes the form

$$\frac{h\left(j,X\mid e\right)}{1-h\left(j,X\mid e\right)}=\left[\frac{h_0\left(j\right)}{1-h_0\left(j\right)}\right]\exp\left(\beta'X+e\right) \tag{8.14}$$

which leads to the model with

$$\text{logit}\left[h\left(j,X\mid e\right)\right]=D\left(j\right)+\beta'X+e \tag{8.15}$$

and e is an 'error' term with mean zero, and finite variance. If one assumes that e has a Normal distribution with mean zero, one has a model that can be estimated using the Stata program **xtlogit** applied to data organized in person-period form.

All the approaches mentioned so far use parametric approaches. A *non-parametric* approach to characterising the frailty distribution was pioneered in the econometrics literature by Heckman and Singer (1984). The idea is essentially that one fits an *arbitrary* distribution using a set of parameters. These parameters comprise a set of 'mass points' and the probabilities of a person being located at each mass point. We have a discrete (multinomial) rather than a continuous mixing distribution. Sociologists might recognize the specification as a form of 'latent class' model: the process describing time-to-event now differs between a number of classes (groups) within the population.

To be concrete, consider the discrete time proportional hazards model, for which the interval hazard rate is given (see Chapter 3) by

$$h(j,X)=1-\exp[-\exp(\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_KX_K+\gamma_j)].$$

Suppose there are two types of individual in the population (where this feature is not observed). We can incorporate this idea by allowing the intercept β_0 to vary between the two classes, i.e.

$$h_1(j,X)=1-\exp[-\exp(\mu_1+\beta_1X_1+\beta_2X_2+...+\beta_KX_K+\gamma_j)] \text{ for Type 1} \tag{8.16}$$

$$h_2(j,X)=1-\exp[-\exp(\mu_2+\beta_1X_1+\beta_2X_2+...+\beta_KX_K+\gamma_j)] \text{ for Type 2.} \tag{8.17}$$

If $\mu_2>\mu_1$, then Type 2 people are fast exiters relatively to Type 1 people, other things being equal. Once again we have a random intercept model, but the randomness is characterised by a discrete distribution rather than a continuous (e.g. Gamma) distribution, as earlier.

If we have an inflow sample with right censoring, the contribution to the sample likelihood for a person with spell length j intervals is the probability-weighted sum of the contributions arising were he a Type 1 or a Type 2 person, i.e.

$$\mathcal{L}=\pi\mathcal{L}(\mu_1)+(1-\pi)\mathcal{L}(\mu_2) \tag{8.18}$$

where

$$\mathcal{L}_1=\left(\frac{h_1(j,X)}{1-h_1(j,X)}\right)^c\prod_{k=1}^{j}\left[1-h_1(k,X)\right] \tag{8.19}$$

$$\mathcal{L}_2 = \left(\frac{h_2(j,X)}{1-h_2(j,X)} \right)^c \prod_{k=1}^{j} \left[1 - h_2(k,X) \right] \tag{8.20}$$

where π is the probability of belonging to Type 1, and c is the censoring indicator. Where there are M latent classes, the likelihood contribution for a person with spell length j is

$$\mathcal{L} = \sum_{m=1}^{M} \pi_m \mathcal{L}(\mu_m). \tag{8.21}$$

The μ_m are the M 'mass point' parameters describing the support of the discrete multinomial distribution, and the π_m are their corresponding probabilities, with with $\sum_m \pi_m = 1$. The number of mass points to choose is not obvious a priori. In practice, researchers have often used a small number (e.g. $M = 2$ or $M = 3$) and conducted inference conditional on this choice. For a discussion of the 'optimal' number of mass points, using so-called Gâteaux derivatives, see e.g. Lancaster (1990, chapter 9).

This model can be estimated in Stata using my program **hshaz** (get it by typing **ssc install hshaz** in an up-to-date version of Stata) or using Sophia Rabe-Hesketh's program **gllamm** (**ssc install gllamm**).

8.3 What if unobserved heterogeneity is 'important' but ignored?

In this section, we provide more details behind the results that were stated at the beginning of the chapter. See Lancaster (1979, 1990) for a more extensive treatment (on which I have relied heavily).

We suppose that the *'true' model*, with no omitted regressors, takes the PH form and there is a continuous mixing distribution:

$$\theta\left(t, X \mid v\right) = v\theta_0\left(t\right)\lambda, \quad \lambda \equiv e^{\beta' X} \tag{8.22}$$

which implies

$$\frac{\partial \log\left[\theta\left(t, X \mid v\right)\right]}{\partial X_k} = \beta_k. \tag{8.23}$$

I.e. the proportionate response of the hazard to some included variable X_k is given by the coefficient β_k. This parameter is constant and does not depend on t (or X). The model also implies

$$\frac{\partial \log\left[\theta\left(t, X\right) v\right]}{\partial t} = \frac{\partial \log \theta_0\left(t, X\right)}{\partial t}. \tag{8.24}$$

We suppose that the *'observed' model*, i.e. with omitted regressors, is

$$\theta\left(t, X \mid v\right) = v\theta_0\left(t\right)\lambda_1 \tag{8.25}$$

where $\lambda_1 = e^{\beta_1' X_1}$, X_1 is a subset of X.

Assuming $v \sim \text{Gamma}(1, \sigma^2)$, then

$$S_1\left(t \mid \sigma^2\right) = \left[1 - \sigma^2 S\left(t\right)\right]^{-\frac{1}{\sigma^2}} = \left[1 + \sigma^2 H\left(t\right)\right]^{-\frac{1}{\sigma^2}} \tag{8.26}$$

and

$$\theta_1\left(t \mid \sigma^2\right) = \left[S_1\left(t \mid \sigma^2\right)\right]^{\sigma^2} \theta_0\left(t\right) \lambda_1. \tag{8.27}$$

8.3.1 The duration dependence effect

Look at the ratio of the hazards from the two models:

$$\frac{\theta_1}{\theta} = \frac{\theta_0\left(t\right) \lambda_1 \left[S_1\left(t \mid \sigma^2\right)\right]^{\sigma^2}}{\theta_0\left(t\right) v} \tag{8.28}$$

$$\propto \left[S_1\left(t \mid \sigma^2\right)\right]^{\sigma^2} \tag{8.29}$$

which is monotonically decreasing with t, for all $\sigma^2 > 0$.

In sum, the hazard rate from a model with omitted regressors increases less fast, or falls faster, that does the 'true' hazard (from the model with no omitted regressors).

The intuition is of a selection or 'weeding out' effect. Controlling for observable differences, people with unobserved characteristics associated with higher exit rates leave the state more quickly than others. Hence 'survivors' at longer t increasingly comprise those with low v which, in turn, implies a lower hazard, and the estimate of hazard is an underestimate of 'true' one.

Bergström and Eden (1992) illustrate the argument nicely in the context of a Weibull model for the non-frailty component. Recall that the Weibull model in the AFT representation is written $\log T = \beta^{*\prime} X + \sigma u$, with the variance of the 'residuals' given by $\sigma^2 \text{var}(u)$. If one added in (unobserved) heterogeneity to the systematic part of the model, then one would expect the variance of the residuals to fall correspondingly. This is equivalent to a decline in σ. But recall $\sigma \equiv 1/\alpha$ where α is the Weibull shape parameter we discussed earlier. A decline in σ is equivalent to a rise in α. Thus the 'true' model exhibits more positive duration dependence than the model without this extra heterogeneity incorporated.

8.3.2 The proportionate response of the hazard to variations in a characteristic

We consider now the proportionate response of the hazard to a variation in X_k where X_k is an included regressor (part of X_1). We know that the proportionate response in the 'true' model is

$$\frac{\partial \log \theta}{\partial X_k} = \beta_k. \tag{8.30}$$

In the observed model, we have

$$\frac{\partial \log \theta_1}{\partial X_k} = \frac{\partial \left[\sigma^2 S_1\left(t \mid \sigma^2\right) + \beta_1' X_1\right]}{\partial X_k} \tag{8.31}$$

$$= \beta_k + \frac{\sigma^2}{S_1\left(t \mid \sigma^2\right)} \frac{\partial \left[S_1\left(t \mid \sigma^2\right)\right]}{\partial X_k}. \tag{8.32}$$

But

$$\frac{\partial \left[S_1\left(t \mid \sigma^2\right)\right]}{\partial X_k} = -\frac{1}{\sigma^2}[1 + \sigma^2 H_1(t)]^{-\frac{1}{\sigma^2}-1} \sigma^2 \frac{\partial \left[H_1\left(t\right)\right]}{\partial X_k} \tag{8.33}$$

$$= \frac{-S_1\left(t \mid \sigma^2\right)}{1 + \sigma^2 H_1(t)} \frac{\partial \left[H_1\left(t\right)\right]}{\partial X_k} \tag{8.34}$$

and

$$\frac{\partial \left[H_1\left(t\right)\right]}{\partial X_k} = \beta_k H_1\left(t\right) \tag{8.35}$$

so

$$\frac{\sigma^2}{S_1} \frac{\partial \left[S_1\right]}{\partial X_k} = \sigma^2 \frac{-\beta_k H_1\left(t\right)}{1 + \sigma^2 H_1(t)}. \tag{8.36}$$

Hence, substituting back into the earlier expression, we have that

$$\frac{\partial \log \theta_1}{\partial X_k} = \beta_k \left[1 - \frac{\sigma^2 H_1\left(t\right)}{1 + \sigma^2 H_1(t)}\right] \tag{8.37}$$

$$= \frac{\beta_k}{1 + \sigma^2 H_1(t)} \tag{8.38}$$

$$= \beta_k \left[S_1\left(t \mid \sigma^2\right)^{\sigma^2}\right]. \tag{8.39}$$

Note that $1 + \sigma^2 H_1(t) > 1$ or, alternatively, that $0 \leq S_1 \leq 1$ and tends to zero as $t \to \infty$. The implications are two-fold:

1. Suppose one wants to estimate β_k (with its nice interpretation as being the proportionate response of the hazard in the 'true' model), then *the omitted regressor model provides an under-estimate* (in the modulus) *of the proportionate response.*

2. Variation in H_1 causes equi-proportionate variation in the hazard rate at all t in the 'true' model. But with omitted regressors, the proportionate effect tends to zero as $t \to \infty$.

The intuition is as follows. Suppose there are two types of person A, B. On *average* the hazard rate is higher for A than for B at each time t. In general the elasticity of the hazard rate at any t varies with the ratio of the average group hazards, θ_A and θ_B. Group A members with high v (higher hazard) leave first, implying that the average θ_A among the survivors falls and becomes more similar to the average θ_B. Thus the ratio of θ_A to θ_B declines as t increases and the proportionate response will fall. It is a 'weeding out' effect again.

8.4 Empirical practice

These results suggest that taking accounting of unobserved heterogeneity is a potentially important part of empirical work. It has often not been so in the past, partly because of a lack of available software, but that constraint is now less relevant. One can estimate frailty models and test whether unobserved heterogeneity is relevant using likelihood ratio tests based on the restricted and unrestricted models. (Alternatively there are 'score' tests based on the restricted model.)

Observe that virtually commonly-available estimation software programs assume that one has data from a inflow or population sample. To properly account for left (or right) truncation requires special programs. For example, the 'convenient cancelling' results used to derive an easy estimation method for discrete time models no longer apply. The likelihood contribution for a person with a left truncated spell still requires conditioning on survival up to the truncation date, but the expression for the survivor function now involves factors related to the unobserved heterogeneity.

The early empirical social science literature found that conclusions about whether or not frailty was 'important' (effects on estimate of duration dependence and estimates of β) appeared to be sensitive to choice of shape of distribution for v. Some argued that the choice of distributional shape was essentially 'arbitrary', and this stimulated the development of non-parametric methods, including the discrete mixture methods briefly referred to.

Subsequent empirical work suggests, however, that the effects of unobserved heterogeneity are mitigated, and thence estimates more robust, if the analyst uses a flexible baseline hazard specification. (The earlier literature had typically used specifications, often the Weibull one, that were not flexible enough.) See e.g. Dolton and van der Klaauw (1995).

All in all, the topic underscores the importance of getting good data, including a wide range of explanatory variables that summarize well the differences between individuals.

Chapter 9

Competing risks models

Until now we have considered modelled transitions out the current state (exit to any state from the current one). Now we consider the possibility of exit to one of several destination states. For illustration, we will suppose that there are two destination states A, B, but the arguments generalise to any number of destinations. The continuous time case will be considered first, and then the discrete time one (for both the interval-censored and 'pure' discrete cases). We shall see that the assumption of *independence in competing risks* makes several of the models straightforward to estimate.

9.1 Continuous time data

Define

$\theta_A(t)$: the latent hazard rate of exit to destination A, with survival times characterised by density function $f_A(t)$, and latent failure time T_A;

$\theta_B(t)$: the latent hazard rate of exit to destination B, with survival times characterised by density function $f_B(t)$, and latent failure time T_B.

$\theta(t)$: the hazard rate for exit to *any* destination.

Each destination-specific hazard rate can be thought of as the hazard rate that would apply were transitions to all the other destinations not possible. If this were so, we would be able to link the observed hazards with that destination-specific hazard. However, as there are competing risks, the hazard rates are 'latent' rather than observed in this way. What we observe in the data is either (i) no event at all (a censored case, with a spell length T_C), or (ii) an exit which is to A or to B (and we know which is which). The observed failure time $T = \min\{T_A, T_B, T_C\}$. What we are seeking is methods to estimate the destination-specific hazard rates given data in this form.

Assume θ_A and θ_B are *independent.* This implies that

$$\theta(t) = \theta_A(t) + \theta_B(t). \tag{9.1}$$

I.e. the hazard rate for exit to any destination is the sum of the destination-specific hazard rates. (Cf. probabilities: $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B)$ if A, B independent. So $\theta(t)\,dt = \theta_A(t)\,dt + \theta_B(t)\,dt$.)

Independence also means that the survivor function for exit to any destination can factored into a product of destination-specific survivor functions:

$$\begin{aligned} S(t) &= \exp\left[-\int_0^t \theta(u)\,du\right] && (9.2)\\ &= \exp\left[-\int_0^t [\theta_A(u) + \theta_B(u)]\,du\right] && (9.3)\\ &= \exp\left[-\int_0^t \theta_A(u)\,du\right]\exp\left[-\int_0^t \theta_B(u)\,du\right] && (9.4)\\ &= S_A(t)\,S_B(t). && (9.5) \end{aligned}$$

The derivation uses the property $e^{a+b} = e^a e^b$.

The individual sample likelihood contribution in the independent competing risk model with two destinations is of three types:

$\mathcal{L}^A$: exit to A, where $\mathcal{L}^A = f_A(T)\,S_B(T)$

$\mathcal{L}^B$: exit to B, where $\mathcal{L}^B = f_B(T)\,S_A(T)$

$\mathcal{L}^C$: censored spell, where $\mathcal{L}^C = S(T) = S_A(T)\,S_B(T)$

In the $\mathcal{L}^A$ case, the likelihood contribution summarises the chances of a transition to A combined with no transition to B, and vice versa in the $\mathcal{L}^B$ case.

Now define destination-specific censoring indicators

$$\begin{aligned} \delta^A &= \begin{cases} 1 \text{ if } i \text{ exits to } A \\ 0 \text{ otherwise (exit to } B \text{ or censored)} \end{cases} && (9.6)\\ \delta^B &= \begin{cases} 1 \text{ if } i \text{ exits to } B \\ 0 \text{ otherwise (exit to } A \text{ or censored)} \end{cases} && (9.7) \end{aligned}$$

The overall contribution from the individual to the likelihood, $\mathcal{L}$, is

$$\begin{aligned} \mathcal{L} &= (\mathcal{L}^A)^{\delta^A}(\mathcal{L}^B)^{\delta^B}\left(\mathcal{L}^C\right)^{1-\delta^A-\delta^B} && (9.8)\\ &= [f_A(T)\,S_B(T)]^{\delta^A}\,[f_B(T)\,S_A(T)]^{\delta^B}\,[S_A(T)\,S_B(T)]^{1-\delta^A-\delta^B} && (9.9)\\ &= \left[\frac{f_A(T)}{S_A(T)}\right]^{\delta^A} S_A(T)\left[\frac{f_B(T)}{S_B(T)}\right]^{\delta^B} S_B(T) && (9.10)\\ &= \left\{[\theta_A(T)]^{\delta^A} S_A(T)\right\}\left\{[\theta_B(T)]^{\delta^B} S_B(T)\right\} && (9.11) \end{aligned}$$

or

$$\ln \mathcal{L} = \left\{ \delta^A \ln \theta_A (T) + \ln S_A (T) \right\} + \left\{ \delta^B \ln \theta_B (T) + \ln S_B (T) \right\}. \qquad (9.12)$$

The log-likelihood for the sample as a whole is the sum of this expression over all individuals in the sample.

In other words, the (log) likelihood for the continuous time competing risk model with two destination states factors into two parts, each of which depends only on parameters specific to that destination. Hence one can maximise the overall (log)likelihood by maximising the two component parts separately. These results generalize straightforwardly to the situation with more than two independent competing risks.

This means that the model can be estimated very easily. Simply define new destination-specific censoring variables (as above) and then estimate separate models for each destination state. The overall model likelihood value is the sum of the likelihood values for each of the destination-specific models.

These results are extremely convenient if one is only interested in estimating the competing risks model. Often, however, one is also interested in testing hypotheses involving restrictions across the destination-specific hazards. In particular one is typically interested in testing whether the same model applies for transitions to each destination or whether the models differ (and how). But introducing such restrictions means that, in principle, one has to jointly estimate the hazards: under the null hypotheses with restrictions, the likelihood is no longer separable.

Fortunately there are some simple methods for testing a range of hypotheses that can be implemented using estimation of single-risk models, as long as one assumes that hazard rates have a proportional hazard form. See Narendranathan and Stewart (1991). The first hypothesis they consider is equality across the destination-specific hazards of all parameters except the intercepts, which is equivalent to supposing that the ratios of destination-specific hazards are independent of survival time and the same for all individuals. The second, and weaker, hypothesis is that the hazards need not be constant over t but are equal across individuals. This is equivalent to supposing that conditional probabilities of exit to a particular state at a particular elapsed survival time are the same for all individuals. Implementation of the first test involves use of likelihood values from estimation of the unrestricted destination-specific models and the unrestricted single risk model (not distinguishing between exit types), and the number of exits to each destination. Implementation of the second test requires similar information, including the number of exits to each destination at each observed survival time.

9.2 Intrinsically discrete time data

Recall that in the continuous time case, exits to only one destination are feasible at any given instant. Hence, assuming independent risks, the overall (continuous

time) hazard equals the sum of the destination-specific hazards. With discrete time, things get more complicated, and the neat separability result that applies in the continuous time case no longer holds. As before let us distinguish between the cases in which survival times are intrinsically discrete and in which they arise from interval censoring (survival times are intrinsically continuous, but are observed grouped into intervals), beginning with the former.

In this case, there is a parallel with the continuous time case: the discrete hazard rate for exit at time j to any destination is the sum of the destination-specific discrete hazard rates. That is,

$$h(j) = h_A(j) + h_B(j). \tag{9.13}$$

Because survival times are intrinsically discrete, if there is an exit to one of the destinations at a given survival time, then there cannot be an exit to the other destination at the same survival time. However this property does not lead to a neat separability result for the likelihood analogous to that for the continuous time case. To see why, consider the likelihood contributions for the discrete time model. There are three types: that for an individual exiting to A ($\mathcal{L}^A$), that for an individual exiting to B ($\mathcal{L}^B$), and that for a censored case ($\mathcal{L}^C$). Supposing that the observed survival time for an individual is j cycles, then:

$$\begin{aligned}
\mathcal{L}^A &= h_A(j)\, S(j-1) && (9.14)\\
&= \left[\frac{h_A(j)}{1-h(j)}\right] S(j) && (9.15)\\
&= \left[\frac{h_A(j)}{1-h_A(j)-h_B(j)}\right] S(j) && (9.16)
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathcal{L}^B &= h_B(j)\, S(j-1) && (9.17)\\
&= \left[\frac{h_B(j)}{1-h(j)}\right] S(j) && (9.18)\\
&= \left[\frac{h_B(j)}{1-h_A(j)-h_B(j)}\right] S(j) && (9.19)
\end{aligned}$$

and

$$\mathcal{L}^C = S(j). \tag{9.20}$$

There is a common term in each expression summarising the overall probability of survival of survival for j cycles, i.e. $S_i(j)$. In the continuous time case, the analogous expression could be rewritten as the product of the destination-specific survivor functions. But this result does not carry over to here:

$$S(j) = \prod_{k=1}^{j} [1-h(k)] = \prod_{k=1}^{j} [1-h_A(k)-h_B(k)]. \tag{9.21}$$

The overall likelihood contribution for an individual with an observed spell length of j cycles is:

$$\begin{aligned}\mathcal{L} &= (\mathcal{L}^A)^{\delta^A}(\mathcal{L}^B)^{\delta^B}\left(\mathcal{L}^C\right)^{1-\delta^A-\delta^B} \\ &= \left[\frac{h_A(j)}{1-h_A(j)-h_B(j)}\right]^{\delta^A}\left[\frac{h_B(j)}{1-h_A(j)-h_B(j)}\right]^{\delta^B} \\ &\quad \times \prod_{k=1}^{j}\left[1-h_A(k)-h_B(k)\right] \end{aligned} \tag{9.22}$$

Another way of writing the likelihood, which we refer back to later on, is

$$\mathcal{L} = S(j)\left[\frac{h(j)}{1-h(j)}\right]^{\delta^A+\delta^B}\left[\frac{h_A(j)}{h(j)}\right]^{\delta^A}\left[\frac{h_B(j)}{h(j)}\right]^{\delta^B}. \tag{9.23}$$

Although there is no neat separability result in this case, it turns out that there is still a straightforward means of estimating an independent competing risk model, as Allison (1982) has demonstrated. The 'trick' is to assume a particular form for the destination-specific hazards:

$$h_A(k) = \frac{\exp(\beta'_A X)}{1+\exp(\beta'_A X)+\exp(\beta'_B X)} \tag{9.24}$$

$$h_B(k) = \frac{\exp(\beta'_B X)}{1+\exp(\beta'_A X)+\exp(\beta'_B X)} \tag{9.25}$$

and hence

$$1-h_A(k)-h_B(k) = \frac{1}{1+\exp(\beta'_A X)+\exp(\beta'_B X)} \tag{9.26}$$

With destination-specific censoring indicators δ^A and δ^A defined as before, the likelihood contribution for the individual with spell length j can be written:

$$\begin{aligned}\mathcal{L} &= \left[\frac{\exp(\beta'_A X)}{1+\exp(\beta'_A X)+\exp(\beta'_B X)}\right]^{\delta^A}\left[\frac{\exp(\beta'_B X)}{1+\exp(\beta'_A X)+\exp(\beta'_B X)}\right]^{\delta^B} \\ &\quad \times \left[\frac{1}{1+\exp(\beta'_A X)+\exp(\beta'_B X)}\right]^{1-\delta^A-\delta^B} \\ &\quad \times \prod_{k=1}^{j-1}\left[\frac{1}{1+\exp(\beta'_A X)+\exp(\beta'_B X)}\right]. \end{aligned} \tag{9.27}$$

However, as Allison (1982) pointed out, this likelihood has the same form as the likelihood for a standard multinomial logit model applied to re-organised data. To estimate the model, there are four steps:

1. expand the data into person-period form (as discussed in Chapter 6 for single-destination discrete time models).

2. Construct an dependent variable for each person-period observation. This takes the value 0 for all censored observations in the reorganised data set. (For persons with censored spells, all observations are censored; for persons with a completed spell, all observations are censored except the final one.) For persons with an exit to destination A in the final period observed, set the dependent variable equal to 1, and for those with an exit to destination B in the final period observed, set the dependent variable equal to 2. (If there are more destinations, create additional categories of the dependent variable.)

3. Construct any other variables required, in particular variables summarising duration-dependence in the destination-specific hazards. Other time-varying covariates may also be constructed.

4. Estimate the model using a multinomial logit program, setting the reference (base) category equal to 0.

Observe that the particular values for the dependent variable that are chosen do not matter. What is important is that they are distinct (and also that one's software knows which value corresponds to the base category). In effect, censoring is being treated as another type of destination, call it C. However in the multinomial logit model, there is an identification issue. One cannot estimate a set of coefficients (call them β_C) for the third process in addition to the other two sets of parameters (β_A, β_B). There is more than one set of estimates that would led to the same probabilities of the outcomes observed. As a consequence, one of the sets of parameters is set equal to zero. In principle, it does not matter which, but in the current context it is intuitively appealing to set $\beta_C = 0$. The other coefficients (β_A, β_B) then measure changes in probabilities relative to the censored (no event) outcome.

The ratio of the probability of exit to destination A to the probability of no exit at all is $\exp(\beta_A)$, with an analogous interpretation for $\exp(\beta_B)$. This is what the Stata Reference Manuals (in the -mlogit- entry) refer to as the 'relative risk'. And as the manual explains, 'the exponentiated value of a coefficient is the relative risk ratio for a one unit change in the corresponding variable, it being understood that risk is measured as the risk of the category relative to the base category' (no event in this case).

A nice feature of this 'multinomial logit' hazard model is that it is straightforward to test hypotheses about whether the destination-specific discrete hazards have common determinants. Using software packages such as Stata it is straightforward to apply Wald tests of whether, for example, corresponding coefficients are equal or not. One can also estimate models in which some components of the destination-specific hazards, for example the duration-dependence specifications, are constrained to take the same form. One potential disadvantage of estimating the model using the multinomial logit programs in standard software

is that these typically require that the same set of covariates appears in each equation.

9.3 Interval-censored data

We now consider the case in which survival times are generated by some continuous time process, but observed survival times are grouped into intervals of unit length (for example the 'month' or 'week'). One way to proceed would simply be to apply the 'multinomial logit' hazard model to these data, as described above, i.e. making assumptions about the discrete time (interval) hazard and eschewing any attempt to relate the model to an underlying process in continuous time. The alternative is to do as we did in Chapter 3, and to explicitly relate the model to the continuous time hazard(s). As we shall now see, the models are complicated relative to those discussed earlier in this chapter, for two related reasons. First, the likelihood is not separable as it was for the continuous time case. Second, and more fundamentally, the shape of the (continuous time) hazard rate *within* each interval cannot be identified from the grouped data that is available. To construct the sample likelihood, assumptions have to be made about this shape, and alternative assumptions lead to different econometric models.

In the data generation processes discussed so far in this chapter, the overall hazard was equal to the sum of the destination-specific hazards. This is not true in the interval-censored case. With grouped survival times, more than one latent event is possible in each interval (though, of course, only one is actually observed). Put another way, when constructing the likelihood and considering the probability of observing an exit to a specific destination in a given interval, we have to take account of the fact that, not only was there an exit to that destination, but also that that exit occurred before an exit to the other potential destinations.

Before considering the expressions for the likelihood contributions, let us explore the relationship between the overall discrete (interval) hazard and the destination-specific interval hazards, and the relationship between these discrete interval hazards and the underlying continuous time hazards. Using the same notation as in Chapter 2, we note that the jth interval (of unit length) runs between dates $a_j - 1$ and a_j. The overall discrete hazard for the jth interval is given by

$$\begin{aligned} h(j) &= 1 - \frac{S(a_j)}{S(a_j - 1)} \\ &= 1 - \frac{\exp\left[-\int_0^{a_j} [\theta_A(t) + \theta_B(t)]dt\right]}{\exp\left[-\int_0^{a_j - 1} [\theta_A(t) + \theta_B(t)]dt\right]} \\ &= 1 - \exp\left[-\int_{a_j - 1}^{a_j} [\theta_A(t) + \theta_B(t)]dt\right] \end{aligned} \tag{9.28}$$

where we have used on the property $\theta(t) = \theta_A(t) + \theta_B(t)$. The destination-specific discrete hazards for the same interval are

$$h_A(j) = 1 - \exp\left[-\int_{a_j-1}^{a_j} \theta_A(t)\, dt\right] \tag{9.29}$$

and

$$h_B(j) = 1 - \exp\left[-\int_{a_j-1}^{a_j} \theta_B(t)\, dt\right]. \tag{9.30}$$

It follows that

$$h(j) = 1 - \{[1 - h_A(j)][1 - h_B(j)]\} \tag{9.31}$$

or

$$1 - h(j) = [1 - h_A(j)][1 - h_B(j)]. \tag{9.32}$$

Thus the overall discrete interval hazard equals one minus the probability of not exiting during the interval to any of the possible destinations, and this latter probability is the product of one minus the destination-specific discrete hazard rates. Rearranging the expressions, we also have

$$\begin{aligned} h(j) &= h_A(j) + h_B(j) + h_A(j)\, h_B(j) && (9.33)\\ &\approx h_A(j) + h_B(j) \text{ if } h_A(j)\, h_B(j) \approx 0. && (9.34) \end{aligned}$$

This tells us that the overall interval hazard is only approximately equal to the sum of the destination-specific interval hazards, with the accuracy of the approximation improving the smaller that the destination-specific hazards are.

Now consider the relationship between the survivor function for exit to *any* destination and the survivor functions for exits to *each* destination:

$$\begin{aligned} S(j) &= (1-h_1)(1-h_2)(....)(1-h_j) && (9.35)\\ &= (1-h_{A1})(1-h_{B1})(1-h_{A2})(1-h_{B2}) \\ &\quad \times ... \times (1-h_{A2})(1-h_{Bj}) \\ &= (1-h_{A1})(1-h_{B2})(.....)(1-h_{Aj}) \\ &\quad \times (1-h_{B1})(1-h_{B2})(....)(1-h_{Bj}). && (9.36) \end{aligned}$$

In other words,

$$S(j) = S_A(j)\, S_B(j) \tag{9.37}$$

so that there is a factoring of the overall grouped data survivor function, analogous to the continuous time case. (The same factoring result can also be derived by expressing the survivor functions in terms of the continuous time hazards, rather than the discrete time ones.)

As before, there are three types of contribution to the likelihood: that for an individual exiting to A ($\mathcal{L}^A$), that for an individual exiting to B ($\mathcal{L}^B$), and that for a censored case ($\mathcal{L}^C$). The latter is straightforward (we have just derived it). For a person with a censored spell length of j intervals,

$$\begin{aligned}\mathcal{L}^C &= S(j) = S_A(j)\, S_B(j) \\ &= \prod_{k=1}^{j}[1-h_A(k)][1-h_B(k)] \qquad (9.38)\end{aligned}$$

What is the likelihood contribution if, instead of being censored, the individual's spell completed with an exit to destination A during interval j? We need to write down the *joint* probability that the exact spell length lay between the lengths implied by the boundaries of the interval *and* that latent exit time to destination B was after the latent exit time to destination A. According to the convention established in Chapter 2, the jth interval is defined as $(a_j - 1, a_j]$. I.e. the interval, of unit length, begins just after date $a_j - 1$, and it finishes at (and includes) date a_j, the start of the next interval. The expression for the joint probability that we want is

$$\begin{aligned}\mathcal{L}^A &= \Pr(a_j - 1 < T_A \leq a_j\ , T_B > T_A) \qquad (9.39)\\ &= \int_{a_j-1}^{a_j}\int_{u}^{\infty} f(u,v)dvdu \qquad (9.40)\\ &= \int_{a_j-1}^{a_j}\int_{u}^{\infty} f_A(u)f_B(v)dvdu \qquad (9.41)\\ &= \int_{a_j-1}^{a_j}\left[\int_{u}^{a_j} f_A(u)f_B(v)dv + \int_{a_j}^{\infty} f_A(u)f_B(v)dv\right]du \qquad (9.42)\end{aligned}$$

where $f(u,v)$ is the joint probability density function for latent spell lengths T_A and T_B, and the lower integration point in the second integral, u, is the (unobserved) time within the interval at which the exit to A occurred. Because we assumed independence of competing risks, $f(u,v) = f_A(u)f_B(v)$. We cannot proceed further without making some assumptions about the shape of the within-interval density functions or, equivalently, the within-interval hazard rates.

Five main assumptions have been used in the literature to date. The first is that transitions can only occur at the boundaries of the intervals. The second is that the destination-specific density functions are constant within each interval (though may vary between intervals), and the third is that destination-specific hazard rates are constant within each interval (though may vary between intervals). The fourth is that the the hazard rate takes a particular proportional hazards form, and the fifth is that the log of the integrated hazard changes linearly over the interval.

9.3.1 Transitions can only occur at the boundaries of the intervals.

This was the assumption made by Narendrenathan and Stewart (1993). They considered labour force transitions by British unemployed men using spell data

grouped into weekly intervals. At the time, there was a requirement for unemployed individuals to register ('sign on'), weekly, at the unemployment office in order to receive unemployment benefits, and so there is some plausibility in the idea that transitions would only occur at weekly intervals. The assumption has powerful and helpful implications. If transitions can only occur at interval boundaries then, if a transition to A occurred in interval $j = (a_j - 1, a_j]$, it occurred at date a_j, and it must be the case that $T_B > a_j$ (i.e. beyond the jth interval). This, in turn, means that $f_B(v) = 0$ between dates u and a_j. This allows substantial simplification of (9.42):

$$\begin{aligned}
\mathcal{L}^A &= \int_{a_j-1}^{a_j} \int_{a_j}^{\infty} f_A(u) f_B(v) dv du && (9.43)\\
&= \int_{a_j-1}^{a_j} f_A(u) du \int_{a_j}^{\infty} f_B(v) dv && (9.44)\\
&= [F_A(a_j) - F_A(a_j - 1)] [1 - F_B(a_j)] && (9.45)\\
&= h_A(j) S_A(j-1) S_B(j) && (9.46)\\
&= \left[\frac{h_A(j)}{1 - h_A(j)}\right] S_A(j) S_B(j). && (9.47)
\end{aligned}$$

By similar arguments, we may write

$$\mathcal{L}^B = \left[\frac{h_B(j)}{1 - h_B(j)}\right] S_A(j) S_B(j). \tag{9.48}$$

In both expressions, the Chapter 3 definitions of the discrete (interval) hazard function (h) and survivor function (S) in the interval-censored case have been used. Of course, empirical evaluation of the expressions also requires selection of a functional form for the destination-specific continuous hazards. A natural choice for these is a proportional hazard specification, in which case $h_A(j)$ and $h_B(j)$ would each take the complementary log-log form discussed in Chapter 3:

$$h_A(j) = 1 - \exp[-\exp(\beta_A' X + \gamma_{Aj})] \tag{9.49}$$

and

$$h_B(j) = 1 - \exp[-\exp(\beta_B' X + \gamma_{Bj})] \tag{9.50}$$

where γ_{Aj} is the log of the integrated baseline hazard for destination A over the jth interval, and γ_{Bj} is interpreted similarly.

If we define destination-specific censoring indicators δ^A and δ^B, as above, then the overall likelihood contribution for the person with a spell length of j intervals is given by

$$\begin{aligned}
\mathcal{L} &= (\mathcal{L}^A)^{\delta^A} (\mathcal{L}^B)^{\delta^B} (\mathcal{L}^C)^{1-\delta^A-\delta^B}\\
&= \left[\frac{h_A(j)}{1 - h_A(j)}\right]^{\delta^A} S_A(j) \left[\frac{h_B(j)}{1 - h_B(j)}\right]^{\delta^B} S_B(j). && (9.51)
\end{aligned}$$

Thus in the case where transitions can only occur at the interval boundaries, the likelihood contribution partitions into a product of terms, each of which is a function of a single destination-specific hazard only. In other words, the result is analogous to that continuous time models (and also generalises to a greater number of destination states). And, as in the continuous time case, one can estimate the overall independent competing risk model by estimating separate destination-specific models having defined suitable destination-specific censoring variables.[1]

9.3.2 Destination-specific densities are constant within intervals

This was the assumption made by, for example, Dolton and van der Klaauw (1999). We begin again with (9.42), but now apply a different set of assumptions. The assumption of constant within-interval densities (as used for the 'acturial adjustment' in the context of lifetables) implies that

$$f_A(u)f_B(v) = \overline{f}_{Aj}\overline{f}_{Bj} \text{ when } a_j - 1 < u \leq a_j \text{ and } a_j - 1 < v \leq a_j. \tag{9.52}$$

Substituting this result into (9.42), we derive

$$\mathcal{L}^A = \int_{a_j-1}^{a_j} \left[\int_u^{a_j} \overline{f}_{Aj}\overline{f}_{Bj} dv + \frac{1}{2}\int_{a_j}^{\infty} f_A(v)f_B(u)dv + \frac{1}{2}\int_{a_j}^{\infty} f_A(v)f_B(u)dv \right] du. \tag{9.53}$$

Now, evaluating the first term of the three within the square brackets,

$$\int_u^{a_j} \overline{f}_{Aj}\overline{f}_{Bj} dv = \overline{f}_{Aj}\overline{f}_{Bj}[a_j - u] \tag{9.54}$$

and then substituting back into $\mathcal{L}^A$, we get

$$\begin{aligned} \mathcal{L}^A &= \int_{a_j-1}^{a_j} \overline{f}_{Aj}\overline{f}_{Bj}[a_j - u]du \\ &\quad + \frac{1}{2}\int_{a_j-1}^{a_j}\int_{a_j}^{\infty} f_A(u)f_B(v)dvdu + \frac{1}{2}\int_{a_j-1}^{a_j}\int_{a_j}^{\infty} f_A(u)f_B(v)dvdu \end{aligned} \tag{9.55}$$

The first term in (9.55) is

$$\begin{aligned} \int_{a_j-1}^{a_j} \overline{f}_{Aj}\overline{f}_{Bj}[a_j - u]du &= \overline{f}_{Aj}\overline{f}_{Bj}[a_j - \frac{1}{2}(a_j)^2 + \frac{1}{2}(a_j - 1)^2] \\ &= \frac{1}{2}\overline{f}_{Aj}\overline{f}_{Bj} \qquad (9.56) \\ &= \frac{1}{2}\int_{a_j-1}^{a_j}\int_{a_j-1}^{a_j} \overline{f}_{Aj}\overline{f}_{Bj} dvdu \qquad (9.57) \end{aligned}$$

[1] In an earlier version of these Lecture Notes, I claimed (incorrectly) that this result held universally for interval-censored spell data. I am grateful for clarificatory discussions with Paul Allison, Chandra Shah, Mark Stewart, Peter Dolton, and especially Wilbert van der Klaauw. My derivations reported for case (2) are based on unpublished notes by Wilbert.

Now, substituting this back into (9.55), the expression for $\mathcal{L}^A$, we can combine the first two terms in square brackets, to derive

$$\begin{aligned}\mathcal{L}^A &= \frac{1}{2}\int_{a_j-1}^{a_j}\int_{a_j-1}^{\infty} f_A(u)f_B(v)dvdu + \frac{1}{2}\int_{a_j-1}^{a_j}\int_{a_j}^{\infty} f_A(u)f_B(v)dvdu \qquad (9.58)\\ &= \frac{1}{2}\Pr(a_j-1<T_A\le a_j\ ,T_B>a_j-1)+\frac{1}{2}\Pr(a_j-1<T_A\le a_j\ ,T_B>a_j) \qquad (9.59)\end{aligned}$$

This is the result cited by Dolton and van der Klaauw (1999, footnote 9). Since survival times T_A and T_B are independent (by assumption), the joint probabilities can be calculated using expressions for the marginal probabilities. That is,

$$\begin{aligned}\mathcal{L}^A &= \frac{1}{2}\left[\Pr(a_j-1<T_A\le a_j)\Pr(T_B>a_j-1)\right]\\ &\quad +\frac{1}{2}\left[\Pr(a_j-1<T_A\le a_j)\Pr(T_B>a_j)\right] \qquad (9.60)\\ &= \Pr(a_j-1<T_A\le a_j)\times\frac{1}{2}\left[\Pr(T_B>a_j-1)+\Pr(T_B>a_j)\right] \qquad (9.61)\\ &= \left[\frac{h_A(j)}{1-h_A(j)}\right]S_A(j)\times\frac{1}{2}\left[S_B(j-1)+S_B(j)\right] \qquad (9.62)\\ &= \left[\frac{h_A(j)}{1-h_A(j)}\right]S_A(j)\times\frac{S_B(j)}{2}\left[\frac{1}{1-h_B(j)}+1\right] \qquad (9.63)\\ &= \left[\frac{h_A(j)}{1-h_A(j)}\right]S_A(j)S_B(j)\left[\frac{1-h_B(j)/2}{1-h_B(j)}\right]. \qquad (9.64)\end{aligned}$$

Similarly, the expression for the likelihood contribution for the case when there is a transition to destination B in interval j is

$$\begin{aligned}\mathcal{L}^B &= \frac{1}{2}\left[\Pr(a_j-1<T_B\le a_j)\Pr(T_A>a_j-1)\right]\\ &\quad +\frac{1}{2}\left[\Pr(a_j-1<T_B\le a_j)\Pr(T_A>a_j)\right] \qquad (9.65)\\ &= \left[\frac{h_B(j)}{1-h_B(j)}\right]S_B(j)\times\frac{1}{2}\left[S_A(j-1)+S_A(j)\right] \qquad (9.66)\\ &= \left[\frac{h_B(j)}{1-h_B(j)}\right]S_B(j)S_A(j)\left[\frac{1-h_A(j)/2}{1-h_A(j)}\right]. \qquad (9.67)\end{aligned}$$

Thus the constant within-interval densities assumption leads to expressions for the likelihood contributions that have rather nice interpretations. The two component probabilities comprising equation (9.60), and similarly in (9.65), provide bounds on the joint probability of interest and, with the assumption, we simply take the average of them. This in turn involves a simple averaging of survival functions that refer to the beginning and end of the relevant interval: see equations (9.62) and (9.66). The averaging property also holds when there are more

than two destinations. If there were three destinations, call them A, B, and D, then expression corresponding to (9.66) is

$$\begin{aligned}\mathcal{L}^B &= \left[\frac{h_B(j)}{1-h_B(j)}\right] S_B(j) \\ &\quad \times \frac{1}{3}\left[S_A(j-1)S_D(j-1)\right] \times \frac{2}{3}\left[S_A(j)S_D(j)\right]. \end{aligned} \tag{9.68}$$

It is as if, with a greater number of competing risks, the fact that a transition to B rather than the other destinations occurred, suggests that the transition times for the other risks were more likely to have been after the end of the interval. The likelihood contribution gives greater weight to end-of-interval survival functions.

To sum up, with two destinations A, B, the overall likelihood contribution for an individual with a spell of j intervals is

$$\begin{aligned}\mathcal{L} &= (\mathcal{L}^A)^{\delta^A}(\mathcal{L}^B)^{\delta^B}\left(\mathcal{L}^C\right)^{1-\delta^A-\delta^B} \\ &= \left[\frac{h_A(j)}{1-h_A(j)}\right]^{\delta^A}\left[\frac{1-h_B(j)/2}{1-h_B(j)}\right]^{\delta^A} S_A(j) \qquad (9.69) \\ &\quad \times \left[\frac{h_B(j)}{1-h_B(j)}\right]^{\delta^B}\left[\frac{1-h_A(j)/2}{1-h_A(j)}\right]^{\delta^B} S_B(j). \qquad (9.70)\end{aligned}$$

where the precise specification depends on the choice of functional form for the discrete (interval) hazard, for example the cloglog form as in (9.49) and (9.50) above. Whichever is used, the expression is not separable into destination-specific components that can be maximized separately. That is, to estimate the independent competing risk model in this case, one could not use standard single-risk model software – special programs are required.

On the other hand, equations (9.64) and (9.67) imply that, if one did (incorrectly) use the single risk approach as in Case 1, then computations of the individual likelihood contributions would be under-estimates, with the magnitude of the error depending on the size of the destination-specific hazard rates. If the hazard rate $h_A(j) = 0.1$, then $\left(\frac{1-h_A(j)/2}{1-h_A(j)}\right) \approx 1.06$, whereas if the hazard equalled 0.01, the ratio is about 1.01. In other words, as the destination-specific hazards become infinitesimally small, then we have the likelihood contributions tend to expressions that are the same as those derived for the case when transitions could only occur at the interval boundaries. The size of the discrete hazards will depend on the application in question, and how wide the intervals are (for example are they one week or one year?).

9.3.3 Destination-specific hazard rates are constant within intervals

With this assumption, the distribution of survival times for each destination (and overall) takes the Exponential form within each interval. Let

$$\begin{aligned}
\theta_A(t) &= \overline{\theta}_{Aj} \text{ if } a_j - 1 < t \leq a_j, \text{ and} && (9.71)\\
\theta_B(t) &= \overline{\theta}_{Bj} \text{ if } a_j - 1 < t \leq a_j, \text{ implying} && (9.72)\\
\theta(j) &= \overline{\theta}_{Aj} + \overline{\theta}_{Bj}, \text{ if } a_j - 1 < t \leq a_j. && (9.73)
\end{aligned}$$

One obvious parameterisation would be $\overline{\theta}_{Aj} = \exp(\beta_{0Aj} + \beta_{1A}X_1 + \beta_{2A}X_2 + ... + \beta_{KA}X_K)$, and $\overline{\theta}_{Bj} = \exp(\beta_{0Bj} + \beta_{1B}X_1 + \beta_{2B}X_2 + ... + \beta_{KB}X_K)$. I.e. each destination-specific hazard rate has a piece-wise constant exponential (PCE) form.

Given these assumptions, the destination-specific interval hazards are, making the appropriate substitutions into (9.29) and (9.30),

$$\begin{aligned}
h_A(j) &= 1 - \exp\left[-\overline{\theta}_{Aj}\right] && (9.74)\\
h_B(j) &= 1 - \exp\left[-\overline{\theta}_{Bj}\right] && (9.75)\\
h(j) &= 1 - \exp\left[-\overline{\theta}_{j}\right]. && (9.76)
\end{aligned}$$

Now consider the likelihood contribution for an individual who made a transition to destination A during interval j. We have explicit functional forms for the density functions within the relevant interval: $f_A(u) = \overline{\theta}_{Aj}S_A(u) = \overline{\theta}_{Aj}S_A(j-1)\exp\left[\overline{\theta}_{Aj}(a_j - 1 - u)\right]$ when $a_j - 1 \leq u < a_j$, and similarly for $f_B(v)$. These can be substituted into (9.42), and the expression evaluated. There are two terms in the double integral. The second is

$$\begin{aligned}
\mathcal{L}^A(\text{second term}) &= \int_{a_j-1}^{a_j}\int_{a_j}^{\infty} f_A(u)f_B(v)dvdu\\
&= h_A(j)S_A(j-1)S_B(j)\\
&= \frac{h_A(j)}{1-h_A(j)}S_A(j)S_B(j) && (9.77)
\end{aligned}$$

$\mathcal{L}^A$(second term) is the likelihood contribution if one knew that no transitions to B could have occurred within the interval. But we have to allow for this possibility; hence the adjustment summarised in the first double integral term in the expression for $\mathcal{L}^A$. Tedious manipulations show that

$$\mathcal{L}^A(\text{first term}) = S(j-1)h(j)\left[\frac{\overline{\theta}_{Aj}}{\overline{\theta}_j} - \frac{h_A(j)}{h(j)}\exp[-\overline{\theta}_{Bj}]\right]. \qquad (9.78)$$

The first two terms in (9.78) represent the probability of survival to beginning of the jth interval followed by exit to any destination within interval j. This probability is then multiplied by a term (in the square brackets), which is the difference between two components. The first component is the ratio of the

destination-specific instantaneous hazard to the overall instantaneous hazard. The second component is the ratio of the destination-specific discrete hazard to the overall discrete hazard, scaled by a factor ($\exp[-\overline{\theta}_{Bj}]$) which is a destination-specific conditional probability of survival (non-exit to B) over an interval of unit length.Observe that if $\overline{\theta}_{Bj} = 0$, then the term in square brackets is equal to zero, and $\mathcal{L}^A$ takes the same form as we derived when transitions could only occur at the boundaries of intervals.

Combining $\mathcal{L}^A$(first term) and $\mathcal{L}^A$(second term), further manipulations reveal that

$$\mathcal{L}^A = S(j-1)h(j)\left(\frac{\overline{\theta}_{Aj}}{\overline{\theta}_{Aj}+\overline{\theta}_{Bj}}\right) = S(j)\left(\frac{h(j)}{1-h(j)}\right)\left(\frac{\overline{\theta}_{Aj}}{\overline{\theta}_{Aj}+\overline{\theta}_{Bj}}\right) \tag{9.79}$$

The contribution in this case is the probability of survival to the beginning of interval j, multiplied by the probability that an event of any type occurred within the interval conditional on survival to the beginning of the interval (i.e. $h(j)$), multiplied by the relative chance that the event was A rather than B at each instant during the interval.

The overall likelihood in this case is

$$\begin{aligned}\mathcal{L}_3 &= S(j)\left(\frac{h(j)}{1-h(j)}\right)^{\delta^A+\delta^B}\left(\frac{\overline{\theta}_{Aj}}{\overline{\theta}_{Aj}+\overline{\theta}_{Bj}}\right)^{\delta^A}\left(\frac{\overline{\theta}_{Bj}}{\overline{\theta}_{Aj}+\overline{\theta}_{Bj}}\right)^{\delta^B} \\ &= S(j-1)\left[1-h(j)\right]^{1-\delta^A-\delta^B}\left[h(j)\right]^{\delta^A+\delta^B}\left(\frac{\overline{\theta}_{Aj}}{\overline{\theta}_j}\right)^{\delta^A}\left(\frac{\overline{\theta}_{Bj}}{\overline{\theta}_j}\right)^{\delta^B}\end{aligned} \tag{9.80}$$

which corresponds to equation 4 of Røed and Zhang (2002a, p. 14). The expression generalises straightforwardly when there are more than two potential destinations. Whatever the number of destinations, observe that the expression for the likelihood contribution is not separable into destination-specific components that can be maximized separately. Observe also that as the destination-specific hazards become infinitesimally small, then we have the likelihood contributions tend to expressions that are the same as those derived for the case when transitions could only occur at the interval boundaries. The size of the interval hazards will depend on the application in question, and how wide the intervals are.

There is also a connection between the expression $\mathcal{L}_3$ and the expression (9.23) describing the likelihood in the 'multinomial' case. Recall that this was:

$$\mathcal{L} = S(j)\left[\frac{h\left(j\right)}{1-h\left(j\right)}\right]^{\delta^A+\delta^B}\left[\frac{h_A\left(j\right)}{h\left(j\right)}\right]^{\delta^A}\left[\frac{h_B\left(j\right)}{h\left(j\right)}\right]^{\delta^B}. \tag{9.81}$$

If intervals are short, or interval-hazards are relatively small then, the more likely it is that $h_A\left(j\right) \approx \overline{\theta}_{Aj}$, $h_B\left(j\right) \approx \overline{\theta}_{Bj}$, In this case, the 'multinomial model' will provide estimates that are similar to the interval-censoring model assuming a constant hazard within intervals (and also assuming that each uses the same specification for duration dependence in the discrete hazard).

9.3.4 Destination-specific proportional hazards with a common baseline hazard function

This assumption underlies the derivation of competing risks models presented by, for example, Hamerle and Tutz (1989). Let us consider the case with two potential destinations A and B. The authors assume that the destination-specific continuous time hazards have the following proportional hazards forms:

$$\begin{aligned} \theta_A(t) &= \theta_0(t)\lambda_A, \text{ where } \lambda_A = \exp(\beta_A' X) && (9.82) \\ \theta_B(t) &= \theta_0(t)\lambda_B, \text{ where } \lambda_B = \exp(\beta_B' X) && (9.83) \end{aligned}$$

where baseline hazard function $\theta_0(t)$ is common to each destination-specific hazard. This is an example of a 'proportional intensity' model. These models have the property that, conditional on exit being made at a particular time, the probability that it is to one specific state rather than any other one does not depend on survival time (Lancaster, 1990, pp. 103–4).

To aid our subsequent derivations, also define the integrated baseline hazard function

$$H(t) = \int_0^t \theta_0(t)dt \tag{9.84}$$

which means that the continuous time hazards can be re-written in terms of the derivative of the integrated hazard function:

$$\begin{aligned} \theta_A(t) &= \lambda_A H'(t) && (9.85) \\ \theta_B(t) &= \lambda_B H'(t). && (9.86) \end{aligned}$$

The survivor function can also be written in terms of the integrated hazard:

$$\begin{aligned} S(t) &= \exp\left[-\int_0^t \theta(t)dt\right] \\ &= \exp[-(\lambda_A + \lambda_B)H(t)]. && (9.87) \end{aligned}$$

The discrete time (interval) hazard rate for the jth interval is

$$\begin{aligned} h(j) &= 1 - \left[\frac{S(a_j)}{S(a_j - 1)}\right] && (9.88) \\ &= 1 - \exp[-(\lambda_A + \lambda_B)(H(a_j) - H(a_j - 1))] && (9.89) \\ &= 1 - \exp[-(\widetilde{\lambda}_{Aj} + \widetilde{\lambda}_{Bj})] && (9.90) \end{aligned}$$

where

$$\begin{aligned} \widetilde{\lambda}_{Aj} &= \exp(\gamma_j + \beta_A' X) = \exp(\gamma_j)\lambda_A && (9.91) \\ \widetilde{\lambda}_{Bj} &= \exp(\gamma_j + \beta_B' X) = \exp(\gamma_j)\lambda_B && (9.92) \end{aligned}$$

and the $\gamma_j \equiv \log[H(a_j)] - H(a_j - 1)]$ are interval-specific parameters. Similarly, we may define destination-specific interval hazard rates

$$h_A(j) = 1 - \exp(-\widetilde{\lambda}_{Aj}) \quad (9.93)$$

$$h_B(j) = 1 - \exp(-\widetilde{\lambda}_{Bj}) \quad (9.94)$$

and survivor functions:

$$S_A(t) = \exp[-\lambda_A I(t)] \quad (9.95)$$

$$S_B(t) = \exp[-\lambda_B I(t)] \quad (9.96)$$

Using arguments similar to those developed earlier for the other cases, the likelihood contribution for an individual with exit to destination A in interval j is

$$\begin{aligned} \mathcal{L}^A &= \int_{a_j-1}^{a_j} \int_u^{a_j} \theta_A(u) S_A(u) \theta_B(u) S_B(u) dv du \\ &\quad + \int_{a_j-1}^{a_j} \int_{a_j}^{\infty} f_A(u) f_B(v) dv du. \end{aligned} \quad (9.97)$$

$\mathcal{L}^A(\text{second term}) = h_A(j) S_A(j-1) S_B(j)$, but what is $\mathcal{L}^A(\text{first term})$? We can rewrite it in terms of λ_A, λ_B, and the integrated baseline hazard function:

$$\mathcal{L}^A(\text{first term}) = \int_{a_j-1}^{a_j} \int_u^{a_j} \lambda_A H'(u) \exp[-\lambda_A H(u)] \lambda_B H'(v) \exp[-\lambda_A H(u)] dv du. \quad (9.98)$$

Using derivations similar to those for Case 3, one can show that this expression can be written as:

$$\mathcal{L}^A = S(j-1) h(j) \left(\frac{\widetilde{\lambda}_{Aj}}{\widetilde{\lambda}_{Aj} + \widetilde{\lambda}_{Bj}} \right) = S(j-1) h(j) \left(\frac{\theta_{Aj}}{\theta_{Aj} + \theta_{Bj}} \right). \quad (9.99)$$

The overall likelihood contribution is given by

$$\begin{aligned} \mathcal{L}_4 &= S(j) \left(\frac{h(j)}{1 - h(j)} \right)^{\delta^A + \delta^B} \left(\frac{\theta_{Aj}}{\theta_{Aj} + \theta_{Bj}} \right)^{\delta^A} \left(\frac{\theta_{Bj}}{\theta_{Aj} + \theta_{Bj}} \right)^{\delta^B} \\ &= S(j-1) \left[1 - h(j) \right]^{1 - \delta^A - \delta^B} \left[h(j) \right]^{\delta^A + \delta^B} \left(\frac{\theta_{Aj}}{\theta_j} \right)^{\delta^A} \left(\frac{\theta_{Bj}}{\theta_j} \right)^{\delta^B} \end{aligned} \quad (9.100)$$

This corresponds to the expression derived by Hamerle and Tutz (1989, pp. 85–7). Clearly the likelihood contribution has a very similar shape that derived for the case where hazards were assumed to be constant within each interval. (Indeed Model 3 can be derived from Model 4. Suppose that $\theta_0(t)$ in

the Proportional Intensities model is time-invariant, in which case $\gamma_j = \gamma$, for all j. Now let the intercept terms in λ_A and λ_B be interval-specific, in which case Model 3 is what results.)

In sum, the proportional intensity model adds some generality (weaker assumptions about the shape of the baseline hazard within intervals) at the cost of forcing a common baseline hazard function for the different latent risks. Analysts would typically like to allow the shapes of the baseline hazards to differ for the various destination types, and to test for equality rather than imposing it from the very start. The PCE model allows each of the baseline hazards to vary in a flexible manner with survival time, and may provide information about whether it is appropriate to make the proportionate intensities assumption.

9.3.5 The log of the integrated hazard changes at a constant rate over the interval

This assumption about the linearity of the within-interval log of the baseline hazards was made by Han and Hausman (1990) and Sueyoshi (1992). We shall not consider it as it has been used relatively rarely and is, in any case, rather complicated to explain. As Sueyoshi (1992, Appendix B) explains, the assumption implies that the hazard increases within each interval. The assumption of a constant density within each interval (case 2 above) also implies that hazards rise within each interval, and contrasts with the assumption of a constant hazard made in the last subsection.

9.4 Extensions

9.4.1 Left-truncated data

The derivations so far has assumed that analyst has access to a random sample of spells. The models can all be easily adapted to the case in which the interval-censored survival time data are subject to left truncation, also known as 'delayed entry'. Suppose that the data for a given subject are truncated at a date within the i^{th} interval, where $i < j$. As we have seen in earlier chapters, to derive the correct likelihood contributions in this case, one needs to condition on survival up to the truncation date. This means dividing the likelihood contribution expression for the random sample of spells case (as considered earlier) by $S(i)$. Now, each of the likelihood expressions for interval-censored data considered earlier ($\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3, \mathcal{L}_4$) is of the form $\mathcal{L} = S(j)Z$, and so the likelihood expression for the left truncation case is simply $\mathcal{L} = S(j)Z/S(i)$. But, given the relationship between the survivor function and the interval hazard, there is a convenient cancelling result: $S(j)/S(i) = \prod_{k=i}^{j}(1 - h_k)$. This has a convenient implication for empirical researchers. Software programs for maximizing $\log \mathcal{L}$ are typically applied to data sets organised so that there is one row for each interval that each subject is at risk of experiencing a transition. In the random sample of spells case, a subject with j contributes j data rows and the log-likelihood

contribution is $\log[S(j)Z] = \sum_{k=1}^{j} \log(1-h_k) + \log(Z)$. With left-truncated data, a subject contributes $j-i$ data rows and the log-likelihood contribution is $\log[S(j)Z/S(i)] = \sum_{k=i+1}^{j} \log(1-h_k) + \log(Z)$. Thus, ICR models for interval-censored and left-truncated survival data can be easily-estimated using the same programs as for random samples of spells, applied to data sets in which data rows corresponding to the intervals prior to the truncation point have been excluded: see Chapter 6.

9.4.2 Correlated risks

The models may also be extended to allow for correlated rather than independent risks.[2] Suppose that each destination-specific hazard rate now depends on individual-specific unobserved heterogeneity in addition to, but independent of, observed heterogeneity (X). Specifically, the hazard for transtions to destination A is rewritten to be a function of $\beta_A'X+\mu_A$ (rather than of $\beta_A'X$ as before), and X no longer includes an intercept term. Similarly, the hazard of transition to destination A is rewritten to be a function of $\beta_B'X + \mu_B$. The latent durations T_A, T_B are now assumed to be independent conditional on the unobserved heterogeneity components. But, by allowing μ_A and μ_B to be correlated, the unconditional latent durations may be correlated.

Perhaps the most straightforward means of estimating the extended model is to suppose that μ_A and μ_B have a discrete distribution with M points of support, and these mass points are estimated together with the probabilities π_m for $m = 1,\dots M$, of the different combinations of μ_A and μ_B. That is, letting the likelihood contribution for each combination of mass points be given by $\mathcal{L}_m(\vartheta\,|\mu_A,\mu_B)$, where regression parameters be represented by the vector ϑ, the overall likelihood contribution for each person is now

$$\mathcal{L}^* = \sum_{m=1}^{M} \pi_m \mathcal{L}_m(\vartheta\,|\mu_A,\mu_B) \text{ with } \sum_{m=1}^{M} \pi_m = 1. \tag{9.101}$$

This is similar to the discussion of discrete mixing distributions to characterize unobserved heterogeneity (see the previous chapter). The difference is that here the distribution is multivariate – the points of support refer to a joint distribution rather than a marginal distribution. Rather than using a discrete mixture distribution, one could assume a continuous distribution, e.g. supposing that the heterogeneity distribution is multivariate normal (cf. Schneider and Uhlendorff, 2004).

[2] See e.g. Dolton and van der Klaauw (1999), Røed and Zhang (2002a), Han and Hausman (1990), Sueyoshi (1992), and Van den Berg and Lindeboom (1998). Conditions for identification of proportional hazards models with regressors for dependent competing risks are given by Abbring and van den Berg (2003), Han and Hausman (1990), and Heckman and Honoré (1989).

9.5 Conclusions and additional issues

Our analysis of competing risks models has shown that, if we assume independence in competing risks, then estimation can be straightforward for several important types of data generation process. Extensions to allow for correlated risks have been introduced in the literature, but have not been used much in applied work yet.

If we have continuous time data, the independent competing risks model can be estimated using standard single-risk models, as discussed elsewhere in this book. Some complications arise with discrete-time data. If the data are genuinely discrete, one cannot use a single-risk model approach, but the 'multinominal logit' model is easy to estimate nonetheless.

With interval-censored data, if one is prepared to assume that transitions can only occur at the interval boundaries, then the modelling can be undertaken straightforwardly, in a manner analogous to that for continuous time data. But in other situations, one cannot avoid modelling the destination-specific models jointly, with different specifications arising depending on the assumptions made about the shape of the hazard (or density) function within each interval. Since alternative assumptions are possible (and each is intrinsically non-testable), this raises the question of whether the estimates derived are sensitive to the choice made. There can be no definitive answer to this question – it depends on the context. Nonetheless the situations in which transitions occur only at the interval boundaries are perhaps relatively rare. Regarding the other alternatives, we may note that Han and Hausman (1990) and Dolton and van der Klaauw (1995) suggested that their estimates were relatively robust to alternative choices. By constrast, Sueyoshi (1992) emphasised that robustness can depend on the width of the intervals, particularly when the models include time-varying covariates.

Finally, a cautionary note about the interpretation of coefficient estimates from destination-specific hazard regressions. In the standard single-risk and given the parameterizations used earlier, if the coefficient estimate associated with a particular explanatory variable X is positive, then there is a straightforward interpretation. For example, with a proportional hazards model, larger values of X imply a larger hazard rate, and shorter survival times. In competing risks, interpretations of coefficients are not always so straightforward. For example, one may be interested in more than simply how a covariate impacts on the destination-specific latent hazard. One may be also interested in estimating the (unconditional) probability of exiting to a particular destination A (say), or the (conditional) probability of exit to A conditional on exiting at a particular survival time, or expected spell length in the state conditional on exit to A.

It turns out that all these quantities depend on all the parameters in the competing risk model, not only those in the destination-specific model of exit to A, as Thomas (1996) explains. He also shows that if each hazard takes the proportional hazard form, then an increase in X will increase the conditional probability of exit to A if its estimated coefficient in the equation for the hazard of exit via A is larger than the corresponding coefficients in the hazards for all other risks. The same issues arise in the intrinsically discrete 'multinomial

logit' competing risks model. For multinomial logit models, it is well-known that increases in a variable with a positive coefficient in the equation for one outcome need not lead to an increase in the probability of that outcome, as the probability of another outcome may increase by even more. Simulating the predicted values of interest for different values of the explanatory variables is one way of addressing the issues raised in this paragraph.

The estimation of unconditional probabilities of exit to a particular destination is known in the biostatistics literature as the estimation of cumulative incidence. See e.g. Gooley et al. (1999). A Stata program **stcompet** is provided by Coviello and Boggess (2004).

Chapter 10

Additional topics

Nothing on this at present – something for the future.

Potential topics might include:

- Estimation using repeated spells and multi-state transition models, and correlated competing risks (both are essentially an extension to the chapter on unobserved heterogeneity). For a helpful introduction to the topic, see Allison (1984). For an extensive, but more advanced, discussion based extensively on the application of mixture models, see Van den Berg (2001).
- Simulation of survival time data

References

Abbring, J.H. and van den Berg, G.J.(2003), 'The identifiability of the mixed proportional hazards competing risks model', *Journal of the Royal Statistical Society (A)*, 65, 701–710.

Abbring, J.H. and van den Berg, G.J.(2003), 'The unobserved heterogeneity distribution in duration analysis', unpublished paper, Tinbergen Institute, Amsterdam.

Allison, P. (1982), 'Discrete time methods for the analysis of event histories', pp. 61–98, in S. Leinhardt (ed) *Sociological Methodology 1982*, Jossey-Bass, San Francisco.

Allison, P. (1984), *Event History Analysis*, Sage, Newbury Park CA.

Allison, P. (1995), *Survival Analysis Using the SAS® System: A Practical Guide*, SAS Institute, Gary NC.

Arulampulam, W. and Stewart, M. (1995), 'The determinants of individual unemployment durations in an era of high unemployment', *Economic Journal* 105, 321–332.

Atkinson, A.B., Gomulka, J., Micklewright, J., and Rau, N. (1984), 'Unemployment benefit, duration and incentives in Britain', *Journal of Public Economics* 23, 3–26.

Baker, M., and Molino, A. (2000), 'Duration dependence and non-parametric heterogeneity: a Monte-Carlo study', *Journal of Econometrics* 96, 357–393.

Bane, M. and Ellwood, D. (1986), 'Slipping in and out of poverty', *Journal of Human Resources* 21, 1–23.

Blank, R.M. (1989), 'Analyzing the length of welfare spells', *Journal of Public Economics* 39, 245–273.

Bergström, R. and Edin, P.-A. (1992), 'Time aggregation and the distributional shape of unemployment duration', *Journal of Applied Econometrics* 7, 5–30.

Blossfeld, H.-P., Hamerle, A., and Mayer, K. (1989), *Event History Analysis*, Lawrence Urbaum Associates, Hillsdale NJ.

Blossfeld, H.-P., and Rohwer, G. (2002) *Techniques of Event History Modeling: New Approaches to Causal Analysis*, second edition, Lawrence Urbaum Associates, Mahwah NJ.

Box–Steffensmeier, J.M. and Jones, B.S. (2004), *Event History Modeling: A Guide for Social Scientists*, Cambridge University Press, Cambridge. 2004

Cleves, M., Gould, W.W., and Gutierrez, R. (2002), *An Introduction to Survival Analysis Using Stata*, Stata Press, College Station TX.

Coviello, V. and Boggess, M. (2004), 'Cumulative incidence estimation in the presence of competing risks', *The Stata Journal*, 4, 103–112.

Cox, D. and Oakes, D. (1984), *Analysis of Survival Data*, Chapman Hall, London.

Dolton, P. and van der Klaauw, W. (1995), 'Leaving teaching in the UK: a duration analysis', *Economic Journal* 105, 431–435.

Dolton, P. and van der Klaauw, W. (1999), 'The turnover of teachers: a competing risks explanation', *Review of Economics and Statistics* 81, 543–552.

Elandt-Johnson, R.C. and Johnson, N.L. (1980, reprinted 1999), *Survival Methods and Data Analysis*, John Wiley, New York.

Ermisch, J.F. and Ogawa, N. (1994), 'Age at motherhood in Japan', *Journal of Population Economics* 7, 393–420.

Gooley, T.A., Leisenring, W., Crowley, J., and Storer, B.E. (1999), 'Estimation of failure probabilities in the presence of competing risks: new representations of old estimators, *Statistics in Medicine*, 18, 695–706.

Gould, W., Pitblado, J. and Sribney, W. (2003), *Maximum Likelihood Estimation with Stata*, second edition, Stata Press, College Station TX.

Gourieroux, C. (2000), *Econometrics of Qualitative Dependent Variables*, Cambridge University Press, Cambridge.

Greene, W. (2003), *Econometric Analysis*, 5th edition, Prentice-Hall International, Englewood Cliffs NJ.

Guo, G. (1993), 'Event-history analysis for left-truncated data', pp. 217–243 in P.V. Marsden (ed.), *Sociological Methodology 1993*, Blackwell, Cambridge MA.

Hachen, D.S. Jr. (1988), 'The competing risks model. A method for analyzing processes with multiple types of events', *Sociological Methods and Research* 17, 21–54.

Ham, J. and Rae, A. (1987) 'Unemployment insurance and male unemployment duration in Canada', *Journal of Labor Economics* 5, 325–353.

Hamerle, A. and Tutz, G. (1988), *Diskrete Modelle zur Analyse von Verweildauer und Lebenzeiten*, Campus, Frankfurt and New York.

Han, A. and Hausman, J. (1990), 'Flexible parametric estimation of duration and competing risks models', *Journal of Applied Econometrics* 5, 1–28.

Heckman, J.J. and Honoré, B.E. (1989), 'The identifiability of the competing risks model', *Biometrika*, 76, 325–330.

Heckman, J.J. and Singer, B. (1984), 'A method for minimising the impact of distributional assumptions in econometric models for duration data', *Econometrica* 52, 271–320.

Holford, T.R. (1980), 'The analysis of rates and survivorship using log-linear models', *Biometrics* 36, 299–305.

Hosmer, D.W. and Lemeshow, S. (1999), *Applied Survival Analysis*, Wiley, New York.

Hoynes, H. and MaCurdy, T. (1994), 'Has the decline in welfare shortened welfare spells?', *American Economic Review Papers and Proceedings* 84, 43–48.

Jenkins, S.P. (1995), 'Easy ways to estimate discrete time duration models', *Oxford Bulletin of Economics and Statistics* 57, 129–138.

Jenkins, S.P. and García-Serrano, C. (2004), 'The relationship between unemployment benefits and re-employment probabilities: evidence from Spain', *Oxford Bulletin of Economics and Statistics* 66, 239–260. Longer version available at http://www.iser.essex.ac.uk/pubs/workpaps/wp2000-17.php

Jenkins, S.P. and Rigg, J.A. (2001), *The Dynamics of Poverty in Britain*, Department for Work and Pensions Research Report No. 157, Corporate Document Services, Leeds. http://www.dwp.gov.uk/asd/asd5/rrep157.asp

Kalbfleisch, J.D. and Prentice, R.L. (1980), *The Statistical Analysis of Failure Time Data*, John Wiley, New York.

Katz, L.F. and Meyer, B.D. (1990). 'The impact of the potential duration of unemployment benefits on the duration of unemployment', *Journal of Public Economics*, 41, 45–72.

Kiefer, N. (1985), 'Econometric analysis of duration data', *Journal of Econometrics* 28, 1–169.

Kiefer, N. (1988), 'Economic duration data and hazard functions', *Journal of Economic Literature* 26, 646–679.

Kiefer, N. (1990), 'Econometric methods for grouped duration data', pp. 97–117 in J. Hartog, G. Ridder, and J. Theeuwes (eds), *Panel Data and Labor Market Studies*, North-Holland, Amsterdam.

Klein, J.P. and Moeschberger, M.L. (1997), *Survival Analysis. Techniques for Censored & Truncated Data*, Springer-Verlag, New York.

Lancaster, T. (1990), *The Econometric Analysis of Transition Data*. Cambridge University Press, Cambridge.

Lancaster, T. (1979), 'Econometric methods for the duration of unemployment', *Econometrica* 47, 939–956.

Meyer, B. (1990), 'Unemployment insurance and unemployment spells', *Econometrica* 58, 757–782.

Narandranathan, W. and Stewart, M. (1991), 'Simple methods for testing the proportionality of cause-specific hazards in competing risks models', *Oxford Bulletin of Economics and Statistics* 53, 331–340.

Narandranathan, W. and Stewart, M. (1993), 'Modelling the probability of leaving unemployment: competing risk models with flexible base-line hazards', *Applied Statistics* 42, 63–83.

Narandranathan, W. and Stewart, M. (1995), 'How does the benefit effect vary as unemployment spells lengthen?', *Journal of Applied Econometrics* 8, 361–381.

Nickell, S. (1979), 'Estimating the probability of leaving unemployment', *Econometrica* 47, 1249–1266.

Nickell, S. and Lancaster, T. (1980), 'The analysis of re-employment probabilities of the unemployed', *Journal of the Royal Statistical Society Series A*, 143, 141–165.

Ondrich, J. and Rhody, S. (1999), 'Multiple spells in the Prentice-Gloeckler-Meyer likelihood with unobserved heterogeneity', *Economics Letters* 63, 139–144.

O'Neill, J., Bassi, L. and Wolf, D. (1987), 'The duration of welfare spells', *Review of Economics and Statistics* 69, 241–248.

Petersen, T. (1991), 'Time-aggregation bias in continuous-time hazard-rate models', pp. 263–290 in P.V. Marsden (ed.), *Sociological Methodology 1991*, Blackwell, Cambridge MA.

Petersen, T. and Koput, K.W. (1992), 'Time-aggregation hazard-rate models with covariates', *Sociological Methods and Research* 21, 25–51.

Rodríguez, G. (1999?), 'Survival analysis', http://data.princeton.edu/wws509/notes/c7.pdf.

Prentice, R.L. and Gloeckler, L.A. (1978), 'Regression analysis of grouped survival data with application to breast cancer data', *Biometrics* 34, 57–67.

Schneider, H. and Uhlendorff, A. (2004), 'The transition from welfare to work and the role of potential labor income', Working Paper 1420, IZA, Bonn.

Shaw, A., Walker, R., Ashworth, K., Jenkins, S., and Middleton, S. (1996), *Moving Off Income Support: Barriers and Bridges*, DSS Research Report No. 53, HMSO, London. HV245.M6

Singer, J.D. and Willett, J.B. (1993), 'It's about time: using discrete-time survival analysis to study duration and the timing of events', *Journal of Educational Statistics* 18, 155–195.

Singer, J.D. and Willett, J.B. (1995), 'It's déjà vu all over again: using multiple-spell discrete-time survival analysis', *Journal of Educational Statistics* 20, 41–67.

Singer, J.D. and Willett, J.B. (2003), *Applied Longitudinal Data Analysis. Modeling Change and Event Occurence*, Oxford University Press, Oxford.

StataCorp (2003), *Stata Statistical Software: Release 8*, Stata Corporation, College Station TX.

Sueyoshi, G.T. (1992), 'Semiparametric proportional hazards estimation of competing risks models with time-varying covariates', *Journal of Econometrics* 51, 25–58.

Sueyoshi, G.T. (1995), 'A class of binary response models for grouped duration data', *Journal of Applied Econometrics* 10, 411–431.

Therneau, T.M. and Grambsch, P.M. (2000), *Modeling Survival Data: Extending the Cox Model*, Springer, New York.

Thomas, J. (1996), 'On the interpretation of covariate estimates in independent competing-risks models', *Oxford Bulletin of Economics and Statistics* 48, 27–39.

Tuma, N.B. and Hannan, M.T., (1984), *Social Dynamics: Models and Methods*, Academic Press, Orlando FL.

Van den Berg, G.J. (2001), 'Duration models: specification, identification and multiple durations', in *Handbook of Econometrics Volume 5* (eds.) JJ. Heckman and E. Leamer, North-Holland, Amsterdam.

Van den Berg, G.J. Lindeboom, M., and Ridder, G. (1994), 'Attrition in longitudinal panel data and the empirical analysis of labor market behavior', *Journal of Applied Econometrics* 9, 421–435.

Van den Berg, G.J. and Lindeboom, M. (1994), 'Attrition in panel survey data and the estimation of multi-state labor market models', *Journal of Human Resources* 33, 458–478.

Wooldridge, J.M. (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge MA.

Yamaguchi, K. (1991), *Event History Analysis*, Sage, Newbury Park CA.

Appendix

Nothing here at present. Future versions may contain appendices about e.g. Maximum Likelihood, and more about Partial Likelihood, or more about the details of statistical testing in general (Wald and likelihood ratio), and residual analysis.