EC327: Financial Econometrics, Spring 2008

Adapted from Stephen P. Jenkins, Notes on Survival Analysis, July 2005

Survival Analysis and Hazard Modelling

We consider 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 instance, the domain of marriage includes the states married, cohabiting, separated, divorced, and single. The domain of paid employment includes the states employed, self-employed, unemployed, inactive and retired.

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). The length of each "spell" 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 (firms or other analytical units), together with information that describes the characteristics of these individuals to be used as explanatory variables in multivariate models.

We consider the methods used to model transition data, and the relationship between transition patterns and characteristics. In general, there may be multiple states (with multi-state transitions) and repeat spells: an individual may be married more than once, for instance. 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

Some common assumptions:

1. 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*).

2. 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.

3. 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 multivariate models. Let us consider these features in turn.

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 pre-specified 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 a UI spell.

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).

Censoring and truncation of survival time data

Just as with standard cross-sectional data, survival time data may be considered as *censored* or *truncated*. 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:

1. *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 $T > t$.

2. *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.

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:

1. *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'.

2. *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 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 $T_i = \min[T_i, C_i]$. If right-censoring is not independent—if 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').

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. 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.

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 biomedical sciences. However discrete time data are relatively common in the social sciences. One should use models that reflect the nature of the data available.

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 as 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.

Why are distinctive statistical methods used?

We provide 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 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

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.)

If we regress $\log(T_i)$ on X_i (noting that survival times are all non-negative and distributions of survival times are typically skewed), we fit a linear relationship. But how should we handle censored cases? We could ignore them altogether (which might remove many observations from the sample, but proportionally more at higher T_i) or we could treat censored observations as if they were complete (again, under-reporting the prevalence of large values of T_i). Neither subsample will recover the appropriate estimate of the effect of X_i on T_i.

In the presence of time-varying covariates (multiple values of X_i per individual), how should we choose which is to be included in an OLS regression?

The arguments for structural modelling point out that economic models of job search, marital search, etc., are framed in terms of decisions about whether to do something (and observed transitions reflect that choice). That is, models are not formulated in terms of completed spell lengths. Perhaps, then, we should model transitions directly.

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 issue), 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 that seems rather arbitrary.

In addition, 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 are often used with other

sorts of data to explore relationships.) In particular, the dependent variable is mis-measured and censoring is not accounted for; and time-varying explanatory variables cannot be handled easily (current values may be misleading).

The hazard rate

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.

In 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 $S(t) = 1 - F(t)$; t is the elapsed time since entry to the state at time 0.

The failure function

$$Pr(T \leq t) = F(t)$$

implies the survivor function

$$Pr(T > t) = 1 - F(t) = S(t).$$

The pdf $f(t)$ is the slope of the cdf (failure) function, $\partial F(t)/\partial t$, or $-\partial S(t)/\partial t$. Both the failure function $F(t)$ and the survivor function $S(t)$ are probabilities and lie between zero and one. The survivor function is a monotone decreasing function of t, equal to 1 at the start of the spell and zero at infinity.

The continuous time hazard rate, $\theta(t)$, is defined as:

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

Thus $\theta(t)\Delta(t)$, for tiny t, 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. The only restriction on the hazard rate is that $\theta(t) \geq 0$.

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*. 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.

If time is intrinsically discrete, we define a discrete time hazard rate. For instance, a basketball team must avoid being eliminated in each round of a tournament in order to make it into the championship game. In the case in which survival times are instrinsically discrete, survival time T is now a discrete random variable with probabilities

$$f(j) = f_j = Pr(T = j)$$

where j indexes the set of positive integers, in terms of 'cycles' rather than intervals of equal length in calendar time. The discrete time survivor function for cycle j, showing the probability of survival for j cycles, is given by:

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

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:

$$h_j = Pr(T = j | T \geq j) = \frac{f(j)}{S(j-1)} = \Pi_{k=1}^{j}(1-h_k)$$

and the discrete time failure function is just one minus that last expression.

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?

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.

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.

To what extent can which economic theory provides suggestions for what the shape of the hazard rate function 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 $A(t)$.

Using a structural approach, 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,

$$\theta(t) = \xi(t)[1 - W(t)]$$

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.)

Using a reduced form approach which places fewer restrictions on the hazard function, we can write the hazard function more generally as

$$\theta(t) = \theta(X(t,s), t),$$

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: $\partial A / \partial t > 0$.

- 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 for Unemployment Insurance (UI) may lead to a benefit exhaustion effect, with the re-employment hazard (θ) rising as the time limit approaches.

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

The proportional hazards model

We consider the hazard rate to vary over individuals, assuming that their characteristics are captured in a vector of variables X, fixed over time, with a linear combination of those variables $\beta'X$ summarizing the individual effect on

hazard, now expressed as $\theta(t, X)$. The *proportional hazards* (PH) model satisfies a separability assumption:

$$\theta(t, X) = \theta_0(t) \exp(\beta' X) = \theta_0(t)\lambda$$

where $\theta_0(t)$, the baseline hazard function, depends on t but not X. It summarizes the pattern of 'duration dependence', assumed to be common to all persons. The parameter $\lambda = \exp(\beta' X)$ is a person-specific non-negative function of covariates which scales the baseline hazard function. The property of PH implies that *absolute* differences in X imply *proportionate* differences in the hazard at each t. With the assumption that the covariates are fixed (not time-varying), this gives rise to the *log relative hazard* model

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

If two individuals only differ on one characteristic, and that characteristic is an indicator

variable (the presence or absence of some feature), then the *hazard ratio* is

$$\frac{\theta(t, X_i)}{\theta(t, X_j)} = \exp(\beta_k)$$

which gives the proportionate change in hazard related to the change in that characteristic, ceteris paribus. Alternatively, we can write

$$\beta_k = \partial \log \theta(t, X)/\partial X_k$$

so that the regression coefficient summarizes the proportional effect on the hazard of absolute changes in the corresponding covariate. This effect does not vary with survival time. However, the PH model can be extended to deal with time-varying covariates.

Cox's proportional hazards 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. Hence the Cox model is sometimes referred to as a semi-parametric model. The model may be estimated in Stata using the command `stcox`. The handout presents several estimates employing the Cox PH model.

Parametric specifications

Although the Cox PH model avoids the need to explicitly model the baseline hazard function, you may be interested in that function. A parametric approach must then be followed in which some specific statistical distribution is chosen to represent the behavior of the hazard

function. These distributions include the exponential, Weilbull, log-logistic, lognormal, Gompertz and generalized Gamma distributions.

Parametric survival-time models may be estimated in Stata using the command `streg` with one of the `distribution()` options. The handout presents an example of estimating a survival model employing the parametric approach.