Microeconometria II

Duration (Failure Time) Models

Introduction

The use of duration (or survival) models is a relatively recent development in economics compared to its more extensive and longer use in the fields of engineering and biomedical research. Survival analysis, the term used within the biomedical tradition, is concerned with a group of individuals for whom a point event of some kind is defined. This point event is often referred to as a failure. The event of failure occurs after a length of time called the failure time and can occur at most once for any individual or phenomenon under scrutiny. Survival analysis can also apply to other phenomenon and may or may not be concerned with individuals or groups of individuals. Examples of failure times include the life-time of machine components in engineering and electronics, the time taken by subjects to complete a psychological experiment, the survival time of patients in a clinical trial, the time to failure of a business, the duration of an industrial strike, the time to the birth of a woman's first child, or the duration of unemployment experienced by an individual.

The purpose of this lecture is to explore the use of failure time models within economics but with a particular emphasis on what is known as the hazard function. The lecture formally introduces the concept of the hazard and its relationship to failure time. It will then introduce a couple of related parametric hazard models and a semi-parametric model based on individual level analysis. An alternative semi-parametric approach based on observational units known as spells at risk is also introduced. The concept of neglected heterogeneity (or 'frailty) and how it impacts the empirical hazard is also discussed in this lecture.

6.1 Defining Failure Times

There are three pre-conditions required to determine failure times precisely.

First, a time origin must be unambiguously defined. Although a precise definition is required for the time origin, this need not be represented by the same calendar time for each unit of observation. Most clinical trials, for instance, have staggered entry, and entry into unemployment, for instance, is also staggered by its nature. There may be cases, of course, where mass redundancies due to plant closure in a given area lead to the time origin having the same calendar date for a large proportion of the sample of unemployed in a given region.

Second, a scale for measuring the passage of time must be agreed. The scale is usually 'clock' or real time but this can vary depending on the application. For instance, in engineering applications, the operating time of a system to a fault, the mileage of a car to a breakdown, or the cumulative load to the collapse of a structure provide a set of alternative measurement scales. In terms of economic applications, however, the scale is always real time measured on person-specific clocks set to zero the moment the person or unit enters the state in question. The measurements may be in days, weeks, months or even years depending on the frequency available for such observational data

Third, the meaning of failure must be unambiguous, clearly understood and precisely defined. Therefore, the meaning and interpretation of the point event must be transparent. In medical work, failure could mean death from a specific cause (e.g., lung cancer) but what if death is from a source unrelated to cancer? In modelling an unemployment duration, we must be clear what constitutes a failure in this case. For instance, an individual may exit unemployment through undertaking training, gaining employment, retiring from the labour market, or dying. We have to be precise about which exits we wish to include in or exclude from our analysis in order to ensure that the empirical modelling is meaningful.

6.2 Censoring

A major problem with analysing survival data is that the data are generally (though not always) censored in one way or another. The common cause in economic applications is that the measurement is undertaken when the process of interest is still ongoing. Thus, if we obtain sample spells of unemployment for individuals drawn from surveys, these will include some individuals who are unemployed at the time of the survey. If the survey was undertaken at time period c_i, for these individuals, duration or survival is to time c_i but is not equal to it. The data are thus censored as those individuals who are unemployed at this point in time could continue in unemployment for considerably longer than c_i. Econometric estimation must take account of this form of censorship if it is present. The consequences of ignoring data censored in this way are analogous to the censorship problems encountered for linear regression analysis discussed in lecture five.

We can be more formal about the problem and introduce some notation. In the absence of censoring, the i^{th} individual in a sample of n has failure time denoted by T_i, a random variable. We assume that there is a period of observation (c_i) such that the observation on that individual ceases at c_i if failure has not occurred by then. The observations consist of $X_i = \min[T_i, c_i]$ together with the indicator variable $y_i = 1$ if $T_i \leq c_i$ (the set of uncensored observations) and $y_i = 0$ if $T_i > c_i$ (the set of censored observations). The c_i of individuals who are observed to fail (i.e., $y_i = 1$) are referred to as unrealised censoring times whilst the c_i of individuals who are observed not to fail (i.e., $y_i = 0$) are referred to as realised censoring times. Figure 6.1 in the lecture slides provides an illustration of differing duration times, a number of which are censored.

6.3 The Hazard Function

Econometricians use the term spell length to describe time occupancy or duration in a given state. The spell length is usually represented, as above, by a random variable, which is denoted by T. T is assumed to be a continuous random variable and we assume a large population of people enter a given state at a time defined at T = 0. As noted in section **6.1**, the calendar time of state entry need not be the same for all individuals. T represents the duration of an observational unit's stay in the state. The population is assumed to be homogeneous implying that every unit's duration of stay will be a realisation of a random variable from the same probability distribution. If we define the probability that a person, who has occupied a state for a time t, leaves it in the short interval of length Δt after t as:

$$\text{prob}[t \leq T \leq t + \Delta t \mid T \geq t] \qquad [6.1]$$

The conditioning event ($T \geq t$) in expression [6.1] is the event that the state is still occupied at time t. In other words, the conditioning event is that the individual has not left the state before time t. If we divide [6.1] by Δt we obtain the average probability of leaving per unit of time over a short interval after t. If we take this average over shorter and shorter intervals we can formally define:

$$\theta(t) = \lim_{\Delta t \to 0} \frac{\text{prob}[t \leq T < t + \Delta t | T \geq t]}{\Delta t} \qquad [6.2]$$

as the hazard function. It is the instantaneous rate of leaving per unit of time period at t. The interpretation of $\theta(t)\Delta t$ (sometimes expressed as $\theta(t)dt$) is the probability of exit from a given state in the short interval of time Δt after t, conditional on the state being occupied at time t.

It is also possible to specify the probability unconditionally (i.e., without the condition $T \geq t$). This is a very different concept from the hazard. For instance, in the context of mortality data, the hazard function gives the probability that a fifty-year old person will die, whereas the unconditional concept gives the probability that a person will die at fifty. In terms of relative frequencies $\theta(50)\Delta t$ gives the proportion of fifty-year olds who die within a small interval (Δt) of their 50th birthday. The unconditional concept gives the proportion of people (ever born) who die within a small interval (Δt) of their 50th birthday. These concepts are clearly different. It proves convenient below to express the hazard function and the unconditional probability of exit in terms of distribution and density functions of a continuous random variable T.

Define prob[$T < t$] = F(t) and note trivially that prob[$T \geq t$] = 1 – F(t). One minus the distribution function is an expression that recurs in applications involving duration or survival data. It is known as the survivor function since it gives the probability of survival to time t. In terms of frequencies, it gives the proportion of a given population who stay (or survive) in the state for at least t years. Because of its special significance in these applications, it has its own special notation:

$$1 - F(t) = \bar{F}(t)$$

Now recall from lecture two that the derivative of the cumulative distribution function is the density function, which we express here as:

$$\Delta F(t)/\Delta t = f(t)$$

Note also from basic statistics that the conditional probability may be written as: $f(x|y)=f(x,y) \div f(y)$. We can use this last result to write an expression for the conditional probability [6.1] as:

$$\text{prob}[t \leq T < t + \Delta t | T \geq t] = \frac{\text{prob}[t \leq T < t + \Delta t, T \geq t]}{\text{prob}[T \geq t]} \qquad [6.3]$$

The numerator of [6.3] is the joint probability that someone leaves a given state in the time interval specified <u>and</u> that the duration is given by $T \geq t$. There is another way of expressing this. In order to make things easier, define $t \leq T < t + \Delta t$ as event A, and $T \geq t$ as event B. Now, we can re-express [6.3] as:

$$\text{prob}(A|B) = \frac{\text{prob}(A \cap B)}{\text{prob}(B)} \quad \text{and note that prob}(B|A) = \frac{\text{prob}(A \cap B)}{\text{prob}(A)}$$

Therefore, we could also conveniently express $\text{prob}(A \cap B) = \text{prob}(B|A) \times \text{prob}(A)$. For the current application, this implies:

$$\text{prob}[t \leq T < t + \Delta t \text{ , } T \geq t\text{ }] = \text{prob}[T \geq t \mid \text{prob }(t \leq T < t + \Delta t)] \times \text{prob}[t \leq T < t + \Delta t\text{ }]$$

Now, $\text{prob}[T \geq t \mid \text{prob }(t \leq T < t + \Delta t)]$ must necessarily be equal to one given the definition of the events as there is an intersection of the two sets: $\{t \leq T \leq t + \Delta t\}$ and $\{T \geq t\}$ here. The latter is thus subsumed within the former. This means that:

$$\text{prob}[t \leq T < t + \Delta t \text{ , } T \geq t\text{ }] = \text{prob}[t \leq T < t + \Delta t\text{ }]$$

This allows us to re-write [6.3] as:

$$\text{prob}[t \leq T < t + \Delta t | T \geq t] = \frac{\text{prob}[t \leq T < t + \Delta t]}{\text{prob}[T \geq t]} \qquad [6.4]$$

The expression [6.4] can be re-cast in terms of the distribution function as:

$$\text{prob}[t \leq T < t + \Delta t | T \geq t] = \frac{F(t + \Delta t) - F(t)}{1 - F(t)} \qquad [6.5]$$

If we divide expression [6.5] through by Δt and take the limit as Δt approaches zero we obtain:

$$\theta(t) = \lim_{t \to 0} \frac{\text{prob}[t \leq T < t + \Delta t | T \geq t]}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \frac{1}{1 - F(t)} \qquad [6.6]$$

The final part of expression [6.6], excluding the reciprocal of the survivor function, is the derivative of the distribution function with respect to t, which, of course, is the density function. It should be noted that $F(t+\Delta t) - F(t) = \Delta F(t)$ and recall $\Delta F(t)/\Delta t = f(t)$. This allows us to write [6.6] as:

$$\theta(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{\bar{F}(t)} \qquad [6.7]$$

In the limit, of course, the 'Δ' notation is replaced in the expression for the derivative by 'd'. The expression [6.7] provides a more compact formula for the hazard function. This expression also allows us to note the difference between the unconditional exit probability which is the area under the probability density function of T from t to $t + \Delta t$

[i.e., $f(t)\Delta t$], and the conditional probability. The former is different from the conditional exit probability (except at the extreme case t=0 where $\bar{F}=1$), which is expressed as $\frac{f(t)\Delta t}{\bar{F}(t)}$. Note also that expression [6.7] appears similar to the expression for the lower truncated or conditional probability density function encountered in lecture five (see expression [5.1] in lecture five). For this reason, [6.7] is sometimes erroneously referred to as a conditional probability density function. It is not since it is a function of t defined over the whole non-negative time axis and not just a truncated part of it. Prior to proceeding further, it is worth noting that $\theta(t)$ can be expressed in an alternative way.

The log of the survivor function can be written as:

$\log[1 - F(t)]$

What is $\frac{d\log[1-F(t)]}{dt}$? If we code $1 - F(t) = z$, this can be obtained using the chain rule as:

$\frac{d\log[1-F(t)]}{dt} = \frac{d\log[z]}{dz}\frac{dz}{dt} = \frac{1}{z}[-f(t)] = \frac{-f(t)}{1-F(t)} = -\theta(t)$. This can be then be re-expressed as:

$$\theta(t) = -d\log[1-F(t)]/dt = -d\log[\bar{F}(t)]/dt \quad \text{or} \quad d\log[1-F(t)]/dt = -\theta(t)$$

Finally, one other concept related to the above expression, which has importance in this literature, is the concept of the integrated hazard (or cumulative hazard). This is defined as:

$$\Lambda(t) = \int_0^t \theta(s)ds = -\log(1 - F(T)) = -\log(\bar{F}(t)) \qquad [6.8]$$

Regardless of what distribution the underlying random variable T follows, $\Lambda(t)$ is unit exponentially distributed and $\log(\Lambda(t))$ is extreme-value distributed.

6.4 Modelling the Hazard Rate Using Parametric Models

Plotting hazard functions can provide some useful insights into how the exit probability is behaving with respect to state duration (i.e., time). The nature of the relationship between the hazard rate and the duration of occupancy is known as duration dependence. For instance, it is possible to have:

(a) $d\theta/dt = 0$ suggesting the hazard rate is constant and the instantaneous rate of exit is invariant to spell duration.

(b) $d\theta/dt > 0$ suggesting positive duration dependence where the instantaneous rate of exit increases with spell duration (e.g., employment/ job durations may provide an example of this – the longer you have been in a given job the greater the likelihood of a quit).

(c) $d\theta/dt < 0$ suggesting negative duration dependence where the instantaneous rate of exit decreases with spell duration (e.g., in labour economics, the exit rate from unemployment is seen to decline with duration due to scarring effects).

There are a number of special distributions that are useful for modelling duration data. The use of specific distributions to model time to failure assumes what is called a parametric approach. It is worth noting that the parametric distributions embody hazards that are restrictive in terms of their relationship with duration. We first examine two popular and related distributions originally used for failure time in other disciplines but also fairly extensively used in economic applications.

6.5.1 The Exponential Distribution

The cumulative distribution function for the exponential is given by:

$F(t) = 1 - \exp[-\theta t]$ for $t \geq 0$ and θ is a positive parameter.

The corresponding density function is given by:

$$\frac{dF(t)}{dt} = \theta \exp[-\theta t] = f(t)$$

The hazard function is thus expressed as:

$$\theta(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{\bar{F}(t)} = \frac{\theta \exp[-\theta t]}{\exp[-\theta t]} = \theta \qquad [6.9]$$

The hazard is a constant and thus independent of time. This distribution has been used to model the time until failure of electronic components primarily because of the memory-less property inherent in the distribution. It is completely described by one parameter (θ). Each unique value of θ determines a different exponential distribution thus implying the existence of a family of exponential distributions. The distribution is skewed to the right and the values of the random variable (T) can vary from 0 to $+\infty$. However, given that the distribution has only one adjustable parameter, estimation methods based on it are very sensitive to even modest departures in the tail of the distribution. This fact, in conjunction with the constancy of the hazard with respect to time, has encouraged applied economists to look at alternative distributions for modelling failure time.

6.5.2 The Weibull Distribution

In contrast to the hazard function for the exponential distribution which is invariant to spell duration, the Weibull distribution is monotonically increasing or decreasing in duration depending on certain parameter values. This is clearly less restrictive than the exponential. The survivor function for the Weibull can be written as:

$$\bar{F}(t) = \exp[-(\lambda t)^{\alpha}] \qquad [6.10]$$

and the cumulative distribution function is given by

$$F(t) = 1 - \exp[-(\lambda t)^{\alpha}] \qquad [6.11]$$

The derivative of expression [6.11] with respect to t yields the density function which, in this case, is:

$$f(t) = \alpha\lambda^{\alpha}t^{\alpha-1}\exp[-(\lambda t)^{\alpha}] \qquad [6.12]$$

and the hazard function can be expressed as:

$$f(t)/\bar{F}(t) = \theta(t) = \alpha\lambda^{\alpha}t^{\alpha-1} \qquad [6.13]$$

This is the Weibull family of distributions and it has been fairly extensively used in applied duration (or failure time) studies. The distribution was first developed by a Swedish industrial engineer (Ernst Waloddi Weibull) who was concerned, among other things, with the strength of steel bars and cotton fibres. The popularity of this family of distributions is attributable to the simplicity of the expressions [6.10] to [6.13]. In contrast to the exponential case, the hazard is not constant and can either rise or fall with duration. The λ parameter is known as the scale parameter and α is known as the shape parameter. The behaviour of the hazard depends on α (the shape parameter). The following should be noted from expression [6.13]:

If $\alpha > 1$ this implies $d\theta/dt > 0$ and suggests an increasing hazard rate with respect to time and positive duration dependence.

If $\alpha = 1$ this implies $d\theta/dt = 0$ and suggests a constant hazard rate with respect to time and no duration dependence.

if $\alpha < 1$ this implies $d\theta/dt < 0$ and suggests a decreasing hazard rate with respect to time and negative duration dependence.

Figure 6.2 in the lecture slides provides a graphical illustration of these three cases.

It is clear from the above that use of the exponential distribution could be problematic if the duration data are characterized by either positive or negative duration dependence. This is one reason why the Weibull is more popular among economists. However, its limitation is that it only allows for increasing or decreasing hazards and not combinations of both. In other words, the hazard is monotonic in time (t). It is entirely possible that the data are not consistent with such monotonicity. Note that it is sometimes conventional in the applied econometrics literature to express the Weibull hazard as:

$$f(t)/\bar{F}(t) = \theta(t) = \alpha\lambda t^{\alpha-1} \qquad [6.13']$$

This allows for the easy introduction of covariates through the scale measure λ, which acts as a link function, as will subsequently be explored.

6.6 Maximum Likelihood Estimation

The parameters α and λ in these types of duration models can be estimated by maximum likelihood procedures. The likelihood function bears a very strong resemblance to that encountered for the censored tobit likelihood function outlined in lecture five. The likelihood function (ϑ) in this case could be defined as:

$$\vartheta = \prod_{\text{Uncensored}} f(t \mid \alpha,\lambda) \prod_{\text{Censored}} \overline{F}(t \mid \alpha,\lambda) \qquad [6.15]$$

where f(·) is the density function. The likelihood functions can be expressed in logarithmic form and the usual algorithms employed for the estimation of its parameters. The inverse of the information matrix can be used to compute the asymptotic variance-covariance matrix for the parameter estimates.

6.7 Exogenous Variables and Duration Analysis

A limitation of the duration models outlined so far is that there has been no role for external factors in determining the duration or the hazard. There are measured differences in individuals or firms that may influence their survival chances in any given application. Regressors can be introduced into the duration models outlined above to capture such differences. These regressors are generally called 'covariates' in this literature and can either be time-invariant or time-varying. For instance, gender and race are time invariant covariates but the age of an individual (or company) or the state of the local economy (i.e., the local unemployment rate) are time-varying covariates. The use of time-varying covariates, however, poses problems for some duration models. The specification of the likelihood function becomes considerably more complex and is not discussed further here. However, a simple alternative approach is suggested in section **6.11** below that reduces the complexity of the problem.

The introduction of time-invariant covariates into the hazard specification is relatively straightforward. For instance, recall that the hazard function for the Weibull model can be expressed as:

$$f(t)/\overline{F}(t) = \theta(t) = \alpha\lambda t^{\alpha-1} \qquad [6.13']$$

In this case the covariates are introduced as a function of λ with $\lambda = \exp[\beta'\mathbf{x}_i]$ where the $\mathbf{x}_i$ includes a constant term and a set of regressors assumed not to change from T=0 to the failure time T=t. These are known as fixed or time-invariant covariates. In using [6.13'] rather than [6.13], the implicit assumption is that the shape parameter is implicitly absorbed into the $\boldsymbol{\beta}$ vector.

This model is sometimes cast as an 'accelerated failure time' (AFT) model depending on the application. The effect of the covariates in accelerated time models is to re-scale time. In other words, a covariate either accelerates or decelerates the time to failure. This is in contrast to a hazard model where the role of the covariate is to change the hazard rate. The advantage of the AFT model is that it has a linear regression model interpretation if α=1 (i.e., the exponential case) of the following form:

$$\ln(T_i) = -\beta'\mathbf{x}_i + v_i \qquad [6.16]$$

where v_i is a random error term that follows some continuous distribution and T_i is time. The distribution of the random term does not involve either $\mathbf{x}$ or β and the regression model is thus homoscedastic. On the other hand, if $\alpha \neq 1$ (i.e., the Weibull case), then the regression model is non-linear and interpreting the estimated effects for the covariate on time to failure is more difficult. The Weibull specification (with the exponential as the special case when $\alpha=1$) is the only member that belongs to both the AFT and the proportional hazards family of models (see below).

6.8 Proportional Hazards Model

If we introduce a function containing a set of covariates, then the hazard is generally re-expressed as:

$$\theta(\mathbf{x};t) = k_1(\mathbf{x})k_2(t) \qquad [6.17]$$

where k_1 and k_2 are the same functions for all individuals. The baseline hazard is common to all units in the population and does not vary across individual units. Individual hazards differ proportionately based only on the realizations of the covariates. The model is called a proportional hazard model because for any two individual units with regressor realizations $\mathbf{x}_1$ and $\mathbf{x}_2$, the hazards for the two are in the same ratio $k_1(\mathbf{x}_1)/k_1(\mathbf{x}_2)$ for all t. The proportionate effect of $\mathbf{x}$ on the hazard is the same for all dates. Thus, if being over fifty years of age lowers the probability of exit from unemployment on the first day by two percent, it lowers the probability of exit on the hundredth day by the same amount. This is a relatively restrictive form and it should be noted that there is no obvious reason why hazards should be proportional in this way for data drawn from economic applications. In certain settings the assumption of a proportional hazard is empirically testable.

6.9 The Weibull Proportional Hazard Model

This became one of the more popular parametric proportional hazard models in the applied econometrics field when economists first started to use such a modelling approach primarily in the research area of unemployment duration. It provides estimates for the baseline hazard and the covariate vector. The estimates for the baseline hazard provide information on the nature of duration dependence and so this may be important from a policy perspective. Under the assumption of a Weibull specification, the baseline hazard is expressed as:

$$k_2(t) = \alpha t^{\alpha-1}$$

We now need to assume some functional form for $k_1(\mathbf{x}_i)$ and generally the most tractable used is:

$$k_1(\mathbf{x}_i) = \exp(\beta'\mathbf{x}_i)$$

Lancaster (1979) modelled unemployment durations using this type of specification for a sample of the unemployed stock drawn at a date during 1973 and interviewed approximately five weeks later. The Weibull proportional hazard is given by:

$$\theta_i(\mathbf{x};t) = \exp(\beta'\mathbf{x}_i)\ \alpha t^{\alpha-1} \qquad [6.18]$$

The choice of the exponential functional form for the covariates ($k_1(\mathbf{x})$) in the link function is one of the most commonly used in this literature and this is primarily because it renders the hazard a log-linear function of the covariates. This then facilitates a cleaner interpretation of the covariate effects.

6.10 The Cox Proportional Hazards Model

Given a proportional hazard of the generic form depicted in [6.17] above, Cox (1972) suggested a partial maximum likelihood method for its estimation. It should be stressed the model's interpretation as a proportional hazard one is only valid when the covariates included are time-invariant, which we will assume to be the case here. The key difference between the Weibull and the Cox proportional hazard model is how the baseline hazard is treated. The functional form for the baseline hazard in the Cox model is left unspecified and so is (implicitly) a flexible form. As a consequence, it is known as a semi-parametric proportional hazards model because it only makes a parametric assumption in regard to how the explanatory variables (or covariates) impact the hazard (i.e., an exponential distribution) but, in contrast to the Weibull model, makes no distributional assumption about the nature of the baseline hazard function itself. The Cox proportional hazards model may be expressed as:

$$\theta_i(\mathbf{x};t) = \exp(\beta'\mathbf{x}_i)\ \lambda_0(t) \qquad [6.19]$$

In this case, the interaction between time (t) and the explanatory variables contained in the link function containing the covariates is multiplicative in nature. The baseline hazard could be interpreted as an intercept term that varies with time. Thus, the unknown baseline hazard parameters potentially have an infinite dimension.

Cox (1972) demonstrated that the parameter vector β could be estimated without specifying any form for the baseline hazard λ_0. In order to see how this was done, rank order the N individual durations (t_i) from lowest to highest (i.e., $t_1 < t_2 < t_3 \ldots\ldots < t_N$). Then, given that one failure is to occur, the conditional probability that the shortest duration observation 1 ended at t_1 given that any of the N observations could have ended at this time is given in general terms by:

$$\frac{\theta(t_1; \beta'\mathbf{x}_1)}{\sum_{i=1}^{N} \theta(t_i; \beta'\mathbf{x}_i)} \qquad [6.20]$$

In other words, this represents the conditional probability that observation 1 is chosen from the risk set comprised of all possible units that could fail at this time.

The above form can be re-expressed more explicitly in this case as:

$$\frac{\lambda_0(t_i)\exp(\beta'\mathbf{x}_i)}{\lambda_0(t_i)\sum_{i=1}^{N} \exp(\beta'\mathbf{x}_i)} \qquad [6.21]$$

Again, the popular exponential form is used for the covariates in the link function. Given the baseline hazard is the same across all individuals given the proportional hazards assumption, the above collapses to:

$$\frac{\exp(\beta'\mathbf{x}_i)}{\sum_{i=1}^{N}\exp(\beta'\mathbf{x}_i)} \qquad [6.21']$$

This expression can then be used in the formulation of a partial likelihood function for estimation purposes. There is some modification that is needed to deal with censorship but this is not developed further here. The term in [6.21'] represents the contribution of the i^{th} unit to this likelihood function. The practical implication of the approach is that the baseline hazard cancels out. In a sense, the parameters of the baseline hazard are treated as a set of nuisance parameters (of potentially an infinite dimension) that needs to be eliminated prior to estimation.

The estimation of the Cox proportional hazard model as described in [6.21'] exploits a partial likelihood function rather than the conventional likelihood function described earlier for the Weibull and related models. The reason it is called partial is because only a partial set of parameters are estimated using this approach. The parameters estimated in this case are the parameters corresponding to the covariates of interest and not those for the baseline hazard as the conditioning approach adopted sweeps out the baseline hazard function given the proportional hazard assumption. The use of a partial likelihood function approach in this case converts a semi-parametric inference problem regarding the unknown baseline function and the unknown $\boldsymbol{\beta}$ parameter vector into a parametric inference problem involving only the unknown $\boldsymbol{\beta}$ parameter vector. The Cox proportional hazard does not contain an intercept term, as this is not identified given the baseline hazard has been conditioned out of the partial likelihood function.

The advantage of this approach is that it provides maximum likelihood estimates of the $\boldsymbol{\beta}$ vector of covariates without requiring specification of a baseline hazard. Therefore, it is viewed as considerably more flexible than a fully parametric hazard model. The partial likelihood is known as a limited information likelihood function. Its status as a valid likelihood function, however, has been questioned in the statistics literature raising questions as to whether it actually yields consistent and efficient estimates. The general consensus is that it does provide consistent estimates though there is some modest efficiency loss when compared to using say a fully parametric Weibull model. The downside to using a fully parametric hazard model, however, is that if any part of the model is mis-specified through use of an inappropriate distributional assumption, the estimation procedure yields inconsistent estimates. A semi-parametric model, like the Cox proportional hazard model, provides some degree of protection against this type of mis-specification if its provenance resides in the baseline hazard.

The Cox model has some inherent disadvantages in the sense that we may be interested in knowing the nature of duration dependence in a particular application and since the baseline hazard is not specified, this is difficult (though not actually impossible) to determine. Secondly, it is cumbersome to adjust the partial likelihood function for the hazard if the data are characterised by many ties in failure times. As noted above, the point of departure in developing the Cox model was to rank order the individual durations. In a continuous setting, the type assumed when using the Cox model, ties in failure times would be extremely unlikely. However, such ties can be fairly common in economics data where a large number of the units in the risk set may exit unemployment in the same week or a large number of strikes terminate on the same day. This follows from the discrete (as distinct from continuous) nature of most economic data. The larger the number of ties in failure time relative to the risk set, the greater the problem. Fortunately, a number of methods are now routinely used to adjust for such ties (e.g., the Breslow approximation or the Efron approximation). Third, the assumption of a proportional hazard is a fairly restrictive assumption in its own right. As noted earlier, this built-in assumption assumes the effect of a covariate on the hazard rate is constant over the span of the failure time analysis. If this assumption is violated, then regression techniques that exploit a proportional hazard form may yield inconsistent and inefficient estimates.

It is possible to investigate graphically the proportional hazard assumption within the Cox model and then statistically test for its presence. The various approaches adopted use a variety of residual measures. One of the most common in this literature is a quantity known as the Schoenfeld residual. The Schoenfeld residual is computed as the difference between the actual covariate value for the individual unit that failed at time t_i and the expected value of the covariate for the risk set at time t_i. The latter is a weighted average of the covariate, where the weights are determined by each individual unit's likelihood of failing at time t_i. These residuals can only be computed for the set of units that actually fail and exit the state. In order for the assumption of a proportional hazard to be satisfied, these residuals are required to be independent of time. Thus, plots of the Schoenfeld residuals against the time to failure for all non-censored units provide visual evidence of whether the residuals for particular covariates and time are independent.

A formal test for the proportional hazard assumption can be conducted using the Schoenfeld residuals. The test exploits an auxiliary regression and is analogous to testing whether the estimated slope from a regression of the (scaled) Schoenfeld residuals on time using only the failed units is statistically different from zero or not. If the slope is zero then the proportional hazard assumption is satisfied. This is ultimately a test of independence between the Schoenfeld residuals and time and can be undertaken covariate by covariate. In addition, an overall (or global) test for all the covariates can be computed to inform on the validity of the proportional hazard assumption for a particular model.

6.11 Neglected Heterogeneity

The problem of heterogeneity can be viewed as the result of an incomplete specification. The inclusion of individual specific covariates is designed to incorporate observation specific effects. If the model specification is incomplete, and if systematic individual differences in the distribution remain after the observed or measured factors have been controlled for, then inferences based on a mis-specified model may be incorrect. This problem is akin to misspecification in the linear regression model. For instance, the ability bias noted in wage equations yields biased OLS estimates in a linear regression framework. Duration analysis can be extended to handle neglected heterogeneity but the mathematics required is challenging. However, Lancaster (1979) proposed an alternative specification for the hazard function as:

$$\mu_i(\mathbf{x};t) = \nu_i\theta_i(\mathbf{x};t) \qquad [6.22]$$

where ν_i is an unobservable random variable independently and identically distributed as Gamma$(1,\sigma^2)$. This random variable may be regarded as a proxy for all unobservable exogenous variables. This is the mechanism through which unobservables are captured and is analogous to the inclusion of an error term in a regression model.

After some extensive (and tedious) algebra, which we do not pursue here, the hazard function can be written (suppressing i subscripts) as:

$$\theta^*(\mathbf{x};t) = \theta(t)[1 - F^*(\mathbf{x};t)]^{\sigma 2} \qquad [6.23]$$

where $F^*(\mathbf{x};t)$ is the cumulative distribution function conditional on v, the unobservables.

This approach was a relatively common one adopted to model neglected heterogeneity in the past. Because $[1 - F^*(x;t)]^{\sigma 2}$ is a decreasing function of t, [6.23] demonstrates that heterogeneity introduces a tendency for a decreasing hazard rate. Estimation of this model is undertaken using maximum likelihood techniques and involves the estimation of an additional parameter σ^2. It should be noted that if $\sigma^2 = 0$ (i.e., the variance of v_i is zero), then there is no heterogeneity present in the data and the model collapses to $\theta^*(t) = \theta(t)$ or the standard Weibull model in this particular case.

In the Lancaster study of unemployment duration, the maximum likelihood estimate of α rises from 0.7 to 0.9 confirming that the negative duration dependence in the Weibull model is more attributable to neglected heterogeneity than to a pure duration dependence effects. The intuition for this is obvious. More mobile and employable individuals are more likely to be the first to leave unemployment leaving the least mobile and less employable behind thus creating the illusion of a stronger negative duration dependence in the data than is actually the case. It can be mathematically shown that a failure to control for such unobserved heterogeneity biases the estimated hazards towards negative duration dependence. A corresponding failure to control for observables has similar effects but the direction of the bias cannot be known *a priori*.

The highly parametric approach to addressing the neglected heterogeneity problem has had its critics. In particular, it has been argued that it tends to over-parameterize the survival distribution leading to serious errors in inference, and estimates can vary dramatically depending on the functional form specified and the mixing distribution used for the neglected heterogeneity. In addition, the choice of the distribution for heterogeneity (for instance, the Gamma distribution) is not motivated by regard to any obvious economic consideration but more by mathematical convenience. Alternative distributions have been suggested other than the Gamma distribution and the inverse Gaussian, which is right-skewed with a heavy tail, is one that has become more popular and commonly used in the economics literature.

The biomedical literature uses a different term to describe the concept of neglected heterogeneity known as 'frailty'. This follows from the fact that an individual unit might possess unobservables that render the unit more prone to exit the living state (i.e., dying). These are interpreted as 'frail' observations, hence the term used in this strand of the failure time literature. Strictly speaking, this is generally known as 'unshared frailty' as it refers to unknown but individual-specific or unit-specific factors that affect the likelihood of failure. The case of 'shared frailty' is where clusters of individuals or units share some common unknown factor that makes all individuals or units within that cluster more likely to fail.

6.12 Discrete-Time Duration Models

6.12.1 The Kaplan–Meier Product Limit Estimator

Economists are generally interested in modelling hazard functions, so it is always convenient to graph the empirical hazard function associated with a particular application in the first instance. A very simple non-parametric procedure, known as the Kaplan–Meier hazard function, is generally used for this purpose. The Kaplan–Meier hazard formula is defined as:

$$\hat{\theta}(T_k) = \frac{h_k}{n_k} \qquad [6.24]$$

where k is the number of distinct survival times; T_k is the risk set at the k^{th} survival time; n_k is the size of the risk set at time k; and h_k is the number of observation spells completed at time k.

This is a non-parametric empirical calculation that imposes no restrictions on the data in the way the parametric (i.e., the Exponential or the Weibull) distributions encountered above do. The computation of this measure can be illustrated using the data reported in the following table.

Table 6.1: Employer Change for 200 Employees

Year	Number Changing Employer	Risk Set
1	11	200
2	25	189
3	10	164
4	13	154
5	12	141

Given the data above, the hazard rates for the five spells at risk could be computed using expression [6.24]. The risk set at the start of the time is 200. Therefore, n_1= 200. In the first year, 11 individuals leave their employer implying h_1 = 11.

Therefore: $\hat{\theta}(T_1) = \frac{11}{200} = 0.055$.

For the following years, we then have:

$\hat{\theta}(T_2) = \frac{25}{189} = 0.132$; $\hat{\theta}(T_3) = \frac{10}{164} = 0.061$;

$\hat{\theta}(T_4) = \dfrac{13}{154} = 0.084$; $\hat{\theta}(T_5) = \dfrac{12}{141} = 0.085$;

The interpretation of the first hazard rate estimate is that there is a 5.5% chance of exiting the state in the first year. The interpretation of the second hazard estimate is that conditional on surviving to the second year, there is a 13.2% chance of exiting the state. These are clearly useful empirical estimates. However, the approach does not allow for the introduction of any covariates. The most important thing to note is that in the Kaplan–Meier procedure the unit of observation for the empirical analysis is no longer the individual but the spell at risk of the event occurring.

6.12.2 Using a Logit Model for Discrete Duration Modelling

Duration or failure time analysis is usually situated in a continuous framework, which has been the approach adopted in the lecture so far. There are sound reasons for the prominence of models rooted in continuous time. First, in most economic models there is no natural time unit within which an individual makes a decision and takes an action. Second, even if there were natural time units there is no guarantee that it would correspond to the monthly, quarterly or annual data normally available to applied economists. Thirdly, inferences about an underlying stochastic process based on interval or point sampled data may be misleading if the assumption of discrete time invoked is incorrect. Fourthly, continuous models are invariant to the time units used. This is not the case for discrete models. Other reasons are also invariably cited such as the fact that continuous time is simpler mathematically and more elegant with some also arguing that it is much more efficient to undertake theoretical thinking within a continuous time context.

More recently, however, setting the analysis within a discrete-time context has become popular in duration analysis as it allows for a degree of simplification in the econometric modelling demands. Jenkins (1995) suggested a useful approach to the estimation of discrete-time duration models using a binary logistic regression model. The approach requires re-organization of the data away from the individual as the unit of observation to the spell at risk of the exit event occurring (as done with the Kaplan-Meier example above). Each individual unit contributes multiple observations to an expanded likelihood function. In the first period an individual unit either stays or exits the state occupied. Those that survive into the second period either remain or exit in that period and so forth into subsequent periods with the numbers surviving, and hence contributing to the likelihood function, diminishing as individuals leave the state in question. The dependent variable in this type of setting could be denoted $y_{i,t}$ and defined as 1 if the i^{th} individual exits the state at time period t, and zero otherwise.

The re-organization of the data allows construction of an unbalanced panel dataset where a maximum of N individuals over T discrete time periods are observed. This could be developed a little further by introducing covariates and separate intercepts for each failure period to capture the baseline hazard. The logistic is a computationally straight-forward transformation that can be used and possesses the added advantage of providing a non-proportional hazard function. This may be formulated as:

$$\text{prob}(y_{i,t}=1) = \frac{\exp[\boldsymbol{\beta}'\mathbf{x}_{i,t} + \boldsymbol{\gamma}'\mathbf{D}_{i,t}]}{1+\exp[\boldsymbol{\beta}'\mathbf{x}_{i,t} + \boldsymbol{\gamma}'\mathbf{D}_{i,t}]} \qquad [6.25]$$

The dependent variable represents the probability of individual i exiting in an interval around period t conditional on having survived to period t, where $\mathbf{x}_{i,t}$ is a vector of covariates, which may or may not vary over time. The vector $\mathbf{D}_{i,t}$ denotes the baseline hazard, which can be specified by a set of dummy variables for each failure time period. This provides another advantage to the discrete-time approach in that the estimates for the baseline hazard are obtained directly as part of the estimation procedure. This econometric approach has the merit of simplicity and also allows an easier introduction of time-varying covariates than do the parametric or semi-parametric continuous-time duration models. It is possible to formulate a discrete failure time model that embodies the restrictive proportional hazards form. For example, the complementary log-log form is an example of such a model, but this is not explored further here.

The Kaplan-Meier hazard function introduced earlier provides a basis for explaining the nature of the 'trick' used to convert the data from an individual to a spell at risk level. We can use an even simpler example than the earlier one. Assume 10 employees are observed for a maximum of three years and have the following turnover pattern within a certain firm:

Table 6.2: Employer Change for 10 Employees

Year	Number Changing Employer	Risk Set
1	3	10
2	1	7
3	2	6
Greater than 3	4	

These data could be converted into separate observations for each year for which each person is observed. Thus, those who changed employer in the first year contributed one-person year each, those who changed employer in the second year contributed two-person years each, and those that changed in the third year contributed the maximum three-person years each. The final four cases that are still present after the third year comprise the set of censored observations. The total sample size can be obtained by summing the risk set. In this case a sample size of 23 is obtained. We now construct a dummy variable y_{it} =1 if the individual exited the state and zero otherwise. This variable would be constructed as follows from the above:

y = (0,0,0,0,0,0,0,1,1,1), (0,0,0,0,0,0,1), (0,0,0,0,1,1)

The first set of binary numbers in conventional brackets refers to the person-specific contributions for year one, the second bracketed set is the year two cases, and the third bracketed set is for the year three cases. If we construct three dummies D_1 = 1 for year one, D_2=1 for year two and D_3 = 1 for year three, we could specify the following linear probability model (LPM) without a constant term as follows:

$$y_{i,t} = \alpha_1 D_{i,1} + \alpha_2 D_{i,2} + \alpha_3 D_{i,3} + u_{i,t}$$

We get the following OLS estimates:

$$y_{i,t} = 0.3D_{i,1} + 0.14D_{i,2} + 0.33D_{i,3} + \hat{u}_{i,t}$$

These estimates are the Kaplan-Meier hazard rates for year one, two and three respectively. Although we are well aware of the limitations of the LPM in this type of context, it provides a useful way of obtaining the Kaplan-Meier hazard rate point estimates. Recall from lecture two that the LPM yields unbiased and consistent estimates. This could be developed a little further by introducing covariates and using a more appropriate model given the discrete nature of the data. This model could be the logit model described in [6.25], for instance.

There are clearly a number of advantages to this approach. Economic data may be reported at discrete intervals (e.g., days, weeks or months) and this approach commends itself for use in such circumstances, though it is acknowledged that the decisions of agents may not be made in such discrete intervals. The approach allows the introduction of time-varying covariates and thus reduces a need to develop application-specific likelihood functions for continuous time models that may be extremely complicated in structure, and for which convergence to global maxima may be difficult. The approach also allows scope for the development of a very flexible non-parametric baseline hazard that is ultimately determined by the data rather than a specific distributional assumption (e.g., the Weibull).

This discrete approach, though computationally straightforward, is not free from criticism. The first relates to the inflation of the sample size through the re-organization of the data. Though a legitimate concern, the estimates obtained remain maximum likelihood and retain the asymptotic properties of such an estimator. The second relates to the treatment of the repeated observations as if they were independent of each other. This is clearly not the case given the data re-organization leads to the potential for correlation across observations in the newly constructed panel. Correlation across observations is likely to introduce some degree of inefficiency in the estimates and a potential downward bias in the sampling variance. A potential solution to this problem is the introduction of an error term of the type introduced to model neglected heterogeneity in duration models or random effects in panel models. Finally, the censoring problem is ignored in estimation. The implication of this does depend on the scale of the censoring problem and this may or may or may not be a trivial issue as it is application dependent.