Econometrics 6027
Lecture 5
Duration Models

Duration models estimate when an individual is likely to exit a state. The instantaneous likelihood of exit is the hazard function, which is a function of time. The survival function captures the probability of not exiting after a given time.

Duration analysis originates in health and engineering studies and is relevant in a variety of real-world situations:

- medical data that tracks how long patient illness lasts
- what determines the survival of firms in a market: why do they leave the market
- whether you are less likely to exit unemployment the longer you have been unemployed

In the rest of this lecture and the next, we will examine:

- some basic concepts
- nonparametric hazard estimation
- specific parametric forms, including the exponential and Weibull
- how to incorporate covariates, including through the proportional/accelerated hazard function
- how to address unobserved heterogeneity, including through a mixed proportional hazard function and Cox function
- censored duration data

1 Some basic concepts

Let A be the *duration of a process*, or the *time to exit from a state*. Let this continuous random variable be denoted T with an associated probability density function $f(t) = \Pr(T = t)$. The duration distribution function $F(t)$ represents the probability of exit from the state by time t, where

$$\begin{aligned} F(t) &= \Pr(T \leq t) \\ &= \int_{s=0}^{t} f(s)ds, \\ \text{which implies that } f(t) &= \frac{dF(t)}{dt} \\ &= \Pr(T = t) \end{aligned}$$

We are more commonly interested in the *probability of survival* $S(t)$ in a state to at least time t. This we get through the simple relation

$$\begin{aligned} S(t) &= Pr(T \geq t) \\ &= 1 - F(t) \\ &= 1 - Pr(T \leq t). \end{aligned}$$

The basic building block in duration modelling is the *exit rate* or *hazard function* at some time t, commonly denoted $h(t)$, which represents the *instantaneous exit rate from the state at time t*. The probability than an individual who has occupied the state until time t leaves the state in a short interval of length dt after t is:

$$\Pr(t \leq T \leq t + dt | T \geq t),$$

giving an average probability of exit per unit of time within the short interval dt of:

$$\frac{\Pr(t \leq T \leq t + dt | T \geq t)}{dt}$$

As we shorten the length of the interval over which this average probability is defined, we converge to the hazard rate $h(t)$. That is,

$$h(t) = \lim_{dt \to 0} \frac{\Pr(t \leq T \leq t + dt | T \geq t)}{dt}.$$

Let's look at the above definition more closely. Using the normal rules for conditional probabilities, we have that

$$\begin{aligned} h(t) &= \lim_{dt \to 0} \{ \frac{\Pr(t \leq T \leq t + dt | T \geq t)}{dt} \} \\ &= \lim_{dt \to 0} \frac{1}{dt} \{ \frac{\Pr(t \leq T \leq t + dt)}{\Pr(T \geq t)} \} \\ &= \lim_{dt \to 0} \frac{1}{dt} \{ \frac{\Pr(T \leq t + dt) - \Pr(T \leq t)}{\Pr(T \geq t)} \} \\ &= \lim_{dt \to 0} \frac{1}{S(t)} \{ \frac{\Pr(T \leq t + dt) - \Pr(T \leq t)}{dt} \} \\ &= \frac{1}{S(t)} \lim_{dt \to 0} \{ \frac{dF(t)}{dt} \} \end{aligned}$$

Therefore,

$$h(t) = \frac{f(t)}{S(t)}.$$

So the hazard rate is the ratio of the duration density to the survival function at time t. Let's explore the relationship between these functions a little more. Recall that $exp(w) = e^w$, where e is the inverse of the natural logarithm.

We have that

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ &= \frac{1}{S(t)} \frac{dF(t)}{dt} \\ &= \frac{1}{S(t)} [-\frac{dS(t)}{dt}] \\ &= -\frac{d \ln S(t)}{dt} \\ &= -\frac{d \ln[1 - F(t)]}{dt} \end{aligned}$$

Integrating the hazard function $h(t)$ across t gives:

$$\begin{aligned} \Lambda(t) &= \int_{s=0}^{t} h(s)ds \\ &= \int_{s=0}^{t} -\frac{d \ln[1 - F(t)]}{dt} ds \\ &= -\ln[1 - F(s)]|_0^t \\ &= -\ln[1 - F(t)] + \ln[1 - F(0)] \\ &= -\ln[1 - F(t)] \text{ since } F(0) = 0. \\ &= -\ln S(t) \end{aligned}$$

That is to say, the integrated hazard $\Lambda(t)$ is precisely the negative of the log survival function. By rearranging the above equation we have that

$$S(t) = \exp[-\int_{s=0}^{t} h(s)ds]$$

which leads to an expression for the density of t,

$$f(t) = \exp[-\int_{s=0}^{t} \lambda(s)ds]h(t).$$

Since $f(t) = S(t) * h(t)$.

So we can describe either the survival function or the duration density in terms of the hazard function $h(t)$.

If the probability of exit increases over time, then the hazard rate is increasing over time. This is called positive duration dependence and corresponds to a circumstance in which:

$$\frac{dh(t)}{dt} > 0.$$

Conversely, if the probability of exit decreases over time, then $h(t)$ decreases with t. This characteristic is termed negative duration dependence, and represents a situation in which for some $t = t^*$,

$$\frac{dh(t)}{dt} < 0.$$

For example, in the study of labour market transitions, one commonly cited view is that it becomes increasingly difficult to secure employment the longer one remains unemployed. In terms of instantaneous exit rate from the state of unemployment, this hypothesis would be supported by an empirical finding of negative duration dependence.

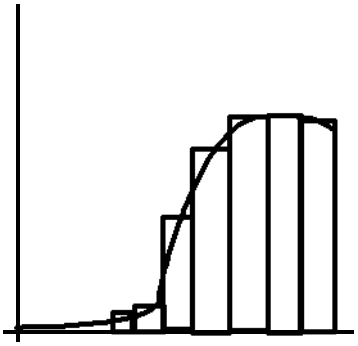
Clearly, the potential patterns of duration dependence depend on the form of $h(t)$. Perhaps the simplest hazard rate is one in which the instantaneous exit rate is constant over time, such that $h(t) = h_0$. However, $h(t)$ may neither be constant, nor indeed monotonic.

Given the importance of the hazard rate in modelling functions, it may be profitable to model the hazard directly rather than as a function of the duration distribution. To proceed further with empirical analysis, some functional specification for the hazard rate (or the survival function directly) is required.

2 Nonparametric Hazard Estimation

Say the curve in this graph is the true CDF, F(t), which we don't know:

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



We do not observe the CDF. We can use parametric or nonparametric means to estimate it. Nonparametric methods use the Empirical Distribution Function (EDF). The EDF counts, for each period, the number of times the duration is within that period, and divides by n. This count is shown as rectangles on the

graph. Mathematically, where N=total number in sample,

$$EDF = \hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_1 \leq t)^t$$

In words,

$$F(t) = \frac{\text{Number of exits}(t+dt) - \text{number of exits total}}{\text{Number in sample} - \text{number of exits}}$$

The survival function is:

$$\begin{aligned}\hat{S}(t) &= 1 - \frac{1}{N}\sum_{i=1}^{N} I(T \leq t)^t \\ &= \frac{N - \text{number of exits by t}}{N} \\ &= \frac{\text{total in sample} - \text{number of exits by t}}{\text{total in sample}}\end{aligned}$$

The PDF is:

$$\begin{aligned}\hat{f}(t) &= \frac{F(t+dt) - F(t)}{dt} \\ &= \frac{\frac{1}{N}\sum I(T \leq t+dt) - \frac{1}{N}\sum I(T_t \leq t)}{dt} \\ &= \frac{1}{N(dt)}(\text{number of exits by (t+dt) - number of exits by t}) \\ &= \frac{1}{N(dt)}(\text{number of exits between t and dt}) \\ &= \frac{\text{number of exits up until } (t+dt) - \text{number of exits up to } t}{\text{total sample}}\end{aligned}$$

The hazard function gives the probability of exit, given that exit hasn't yet happened:

$$\begin{aligned}\hat{h(t)} &= \frac{\hat{f(t)}}{\hat{s(t)}} \\ &= \frac{\text{number of exits up until } (t+dt) - \text{number of exits up to } (t)}{\text{total sample} - \text{number exits up to } (t)}\end{aligned}$$

2.1 Analysis of Nonparametric Hazard Estimation

- this Kaplan-Meier nonparametric estimator is good for estimating bumpy hazards, that is, where the CDF has a lot of turns that don't fit a given parametric or semiparametric model
- however, its hazard rate is not really a "rate" of exit

- it may have weak predictive power: it reflects the data, but doesn't generalize enough to predict the future
- duration here is estimated in discrete intervals, ie. one year, and though we can't make this go to zero as in a limit exercise, we can assume that it smooths towards a curve.

3 Some specific parametric forms

Parametric estimation of the hazard function introduces specific parameters to be estimated – such as the hazard rate level and curvature – and a given distribution for the errors.

3.1 The Exponential Distribution

The simplest case of constant duration dependence corresponds to a constant hazard rate $h(t) = h_0$. We know from duration theory that:

$$S(t) = \exp[-\int_{s=0}^{t} h(s)ds]$$

so: (proof 1):

$$\begin{aligned} S(t) &= \exp[-\int_{s=0}^{t} h_0 ds] \\ &= e^{-h_0 \int_{s=0}^{t} ds} \\ &= e^{-h_0 t} \end{aligned}$$

(proof 2):

From duration theory, we know that:

$$\begin{aligned} \frac{d \ln S(t)}{dt} &= -h_0 \\ \ln S(t) &= k - h_0 t \\ S(t) &= exp(k - h_0 t) \\ &= k * exp(-h_0 t) \text{ but at } s(0), k = 1 \text{ so } k = 1 \\ &= exp(-h_0 t) \\ &= e^{-h_0 t} \end{aligned}$$

The PDF is:

$$\begin{aligned} f(t) &= h(t) * s(t) \\ &= h_0 e^{-h_0 t} \end{aligned}$$

and the CDF is:

$$\begin{aligned} F(t) &= \frac{h(t) - f(t)}{h(t)} \\ &= \frac{h_0 - h_0 exp(-h_0 t)}{h_0} \\ &= 1 - e^{-h_0 t} \end{aligned}$$

We can estimate using a maximum likelihood function:

$$\hat{h_0} = \arg\max_{h_0} \sum_{i=1}^{n} \ln f(t_i)$$

3.1.1 Analysis of Exponential Distribution

- The exponential distribution (so-called because of the exponential form for the duration density and distribution functions) is clearly a restrictive specification, and appropriate only for certain economic or statistical applications.
- The major drawback inherent in this specification is that the conditional probability of exit is constant, which implies no positive or negative duration dependence. This characteristic can also be expressed in terms of a process with no memory.
- It can be argued that a model of labour market transition from unemployment founded on the reservation wage model may be modelled using the exponential distribution. That is to say, if wage offers from some (constant) wage offer distribution arrive at a constant rate over time, then the probability of labour market entry is also constant.

3.2 The Weibull Distribution

To introduce duration dependence (positive or negative) into a model of state transitions requires some more general functional specification for the hazard rate. A popular one-parameter generalisation of the exponential distribution is the Weibull distribution, for which

$$h(t) = h_0 \lambda (h_0 t)^{\lambda - 1}$$

Here, we have not only the h_0 parameter, but also another parameter λ which defines the shape of the curvature. h_0 and λ are positive by definition.

If $0 < \lambda < 1$, we have negative duration dependence since $\lambda - 1$ has to be negative with $0 < \lambda < 1$ so:

$$\frac{\partial h(t)}{\partial t} = h_0^2 \lambda (\lambda - 1) * (h_0 t)^{\lambda - 2} < 0$$

If $\lambda > 1$, we have positive duration dependence because

$$\frac{\partial h(t)}{\partial t} = h_0^2 \lambda(\lambda - 1) * (h_0 t)^{\lambda - 2} > 0.$$

Using the usual relationships between the hazard and survival functions, we can see that the Weibull hazard function corresponds to a survival function of the form:

$$s(t) = exp(-\int_0^t h(s)ds)$$

We can obtain the density through the definitions of the hazard and survival functions:

$$\begin{aligned}
h(t) * s(t) &= f(t) \\
f(t) &= h_0 \lambda (h_0 t)^{\lambda - 1} * exp(-\int_0^t h_0 \lambda (h_0 s)^{\lambda - 1} ds) \\
&= h_0^\lambda \lambda t^{\lambda - 1} exp(-h_0 \lambda h_0^{\lambda - 1} \int_0^t s^{\lambda - 1} ds) \\
&= h_0^\lambda \lambda t^{\lambda - 1} exp(-h_0^\lambda \lambda \int_0^t (\frac{s^\lambda}{\lambda}) ds) \\
&= h_0^\lambda \lambda t^{\lambda - 1} exp(-h_0^\lambda \lambda \frac{1}{\lambda} \int_0^t (s^\lambda) ds) \\
&= h_0^\lambda \lambda t^{\lambda - 1} exp(-h_0^\lambda s^\lambda |_0^t) \\
&= h_0^\lambda \lambda t^{\lambda - 1} exp(-h_0^\lambda t^\lambda) \\
&= h_0^\lambda \lambda t^{\lambda - 1} e^{-h_0^\lambda t^\lambda} \\
\text{since } s^\lambda |_0^t &= t^\lambda - 0^\lambda = t^\lambda
\end{aligned}$$

We can estimate using a maximum likelihood function:

$$\hat{h_0}, \hat{\lambda} = \arg\max_{h_0, \lambda} \sum_{i=1}^{u} \ln(h_0^\lambda) + \ln \lambda + (\lambda - 1) \ln t - h_0^\lambda t^\lambda$$

3.2.1 Analysis of the Weibull Distribution

- The Weibull model is more widely applicable than the simpler exponential distribution because we can use the parameter λ to define the sign and degree of duration dependence.
- The Weibull is nevertheless still restrictive, to the extent that the hazard rate is monotonic. That is to say, duration dependence is either entirely positive or entirely negative, and can never change sign over the range of state duration.

3.3 Log Logistic Distribution

The log-logistic distribution is an alternative one-parameter generalisation of the exponential distribution, and corresponds to a hazard rate of the form

$$h(t) = \frac{h_0 \lambda (h_0 t)^{\lambda - 1}}{1 + (h_0 t)^{\lambda}} \text{ for } h_0, \lambda > 0$$

The survival function corresponding to this distribution, although harder to derive, is

$$S(t) = \frac{1}{1 + (h_0 t)^{\lambda}}$$

The log logistic exhibits a different pattern of duration dependence to the Weibull; the hazard rate first increases with t before decreasing as t increases.