The Basic Components of Duration Analysis

1 Introduction

Assume we have N units $i = 1, 2 \ldots N$, each of which will experience some event. In general, we think that there is some latent failure time for unit i given by T_i^* and some latent censoring time C_i^*, and what we actually observe is $T_i = \min\{Y_i^*, C_i^*\}$.[1] One might think that all we need to do is to estimate an MLE model of $P(T_i)$ i.e. the probability that the observed duration for a unit is equal to some particular time t. We would do this by assigning some probability density, $f(t)$, to the probability of the event, then use MLE. But there are two problems with this:

1. The probability is conditional i.e. you cannot have an event at time t unless you have not had the event in previous periods.
2. Some observations are censored.

So, how do we estimate a duration model?

2 Continuous or Discrete Time Duration Processes

Before we get to how we estimate duration models, we need to recognize that duration data comes in two types: continuous and discrete time duration data. If we assume that transition events can occur at any particular instant in time, then we are dealing with what is called a *continuous time duration process*. In this setting, you can think of time as a continuum and events can occur at any non-negative point in time. This is only one way to think about duration processes, though. Another approach to duration processes is to think that events occur at discrete points in time – a *discrete time duration process*. There are two different reasons to think that events occur at discrete points in time. The first is that even if events can occur at any point in time, the way our data is recorded makes them look as if they occur at discrete intervals in time. In other words, our duration times are 'grouped' or 'banded' into discrete time intervals such as days, months, or years. For example, we might have data indicating that two governments collapsed after 100 days. In reality, it is almost certain that one of these governments collapsed before the other – maybe one collapsed at 9am on the 100th day and the other collapsed at 1pm on the 100th day. Instead of recording the data in time units that allow us to capture the different order in which these governments collapsed, we only know that they both died on the 100th day. In effect, all the data indicate is that both governments collapsed in the interval of time captured by the 100th day. In biostatistics, this type of situation is known as *interval censoring*. In general, this type of discrete duration data is known as discrete *grouped* duration data since observations are grouped in terms of the interval of time in which an event occurs. The second way to think about discrete time data is to believe that events really do occur at discrete points in time; in effect, the underlying event (transition)

[1]Note at this point that we are assuming that there is a latent failure time for each unit. In other words, if we observed a unit long enough, we would see it experience an event i.e. fail. In effect, almost all duration models that we will look at assume that all units will fail if we observe them long enough. This is something that we will come back to, particularly when we look at split-population models.

process is an *intrinsically discrete* one. The example of an intrinsically discrete time process given by Jenkins (2008) is that of fertility – if one were interested in the duration from puberty to first birth, it might make sense to measure time in terms of the number of menstrual cycles rather than in terms of the number of months or days. In practice, the same types of duration models are used for both grouped and intrinsically discrete duration processes. As a result, people tend to divide duration models into two main types: continuous and discrete.[2]

Types of Duration Data:

1. Continuous Time Duration Process
2. Discrete Time Duration Process
   - Grouped (Interval Censored) Duration Process
   - Intrinsically Discrete Duration Process

It is important to think about the type of duration process that you are dealing with and the type of duration data that you have. This will help you determine what kind of duration model is most appropriate for you. On the whole, intrinsically discrete duration processes are probably quite rare in political science. However, it is relatively common (always?) for duration data to be recorded in grouped form. In order to determine whether you should be using a continuous time or discrete time duration model, the key thing to think about is the length of the time intervals (hours, days, months, years etc.) relative to the typical duration spell length: the smaller the ratio of interval lengths to spell length, the more appropriate it is to use continuous time duration models. In other words, you might decide to use a continuous time duration model in situations where the interval length is quite short (days) but units last a reasonably long time (hundreds of days).[3] There is no rule of thumb here and you will have to think through these issues yourself.

In order to know how to estimate these different types of duration models, we need to know their basic mathematical components. This involves identifying things like the failure function, the survivor function, the hazard function, and the cumulative hazard function etc. We'll start by looking at continuous time duration processes.

3 Continuous Time Duration Models

3.1 Mathematical Components

In continuous time duration processes, the length of a spell for some unit (individual, government, country etc.) is a realization of a continuous random variable T with a cumulative distribution function $F(t)$ and a probability density function $f(t)$. $F(t)$ is known as the failure function.[4] In all of this, t is the elapsed time since a unit enters the study in the starting state at time 0.

[2] The mathematics behind the two different types of discrete time processes do differ slightly though.

[3] As Jenkins (2008) notes, if a large number of observations have the same duration times (something called 'tied data'), then you should probably take account of the banding of survival times by using a grouped discrete time duration model.

[4] Although the notes shown here are based on numerous sources, special mention should go to Jenkins (2008).

The failure function or cumulative probability of death (experiencing an event) is:

$$F(t) = \int_0^t f(u)du = P(T \leq t) \tag{1}$$

This is the probability of death on or before t (or probability that survival time T is less than or equal to some value t).

From this, we can derive what is known as the survivor function:

$$S(t) = 1 - F(t) = P(T > t) \tag{2}$$

This is the probability of surviving to time t.

The failure function $F(t)$ and the survivor function $S(t)$ are both probabilities. As a result, the survivor function lies between 0 and 1, and is a strictly decreasing function of time t. The survivor function equals 1 when $t = 0$ and decreases towards zero as t goes to infinity. Given that $S(t)$ is a probability, we know that:

$$0 \leq S(t) \leq 1 \tag{3}$$

$$S(0) = 1 \tag{4}$$

$$\lim_{t \to \infty} S(t) = 0 \tag{5}$$

$$\frac{\partial S}{\partial t} < 0 \tag{6}$$

For all the points that $F(t)$ is differentiable, then a probability density function $f(t)$ is defined, and can be expressed as:

$$f(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T \leq t + \Delta t)}{\Delta t} = \frac{\partial F(t)}{\partial t} = -\frac{\partial S(t)}{\partial t} \tag{7}$$

where Δt is a very, very small interval of time. As Eq. (7) indicates, $f(t)\Delta t$ indicates the probability that an observation will experience an event or die at exactly t i.e. the probability that it will experience an event in the very tiny interval of time $[t,\, t + \Delta t]$. $f(t)$ is sometimes known as the unconditional failure rate. $f(t)$ is non-negative and may be greater than one in value since, as Eq. (7) indicates, the density function does not summarize probabilities per se.

$$f(t) \geq 0 \tag{8}$$

Knowing something about failures through $f(t)$ and survivors through $S(t)$, we can now talk about *risk*. As we saw in the introduction, risk is simply the relationship between failure and survival. As an observation survives through time, it incurs a risk that at some point it will fail. This takes us to the idea of a hazard rate $h(t)$. The continuous time hazard rate, sometimes known as the conditional failure rate, is defined as

$$h(t) = \frac{f(t)}{S(t)} = \lim_{\triangle t \to 0} \frac{Pr(t \leq T \leq t + \triangle t | T \geq t)}{\triangle t} \tag{9}$$

What exactly is the hazard rate? The hazard rate tells us the rate of failure per time unit in the interval $[t,\ t+\triangle t)$ conditional on survival at or beyond time t. In effect, it gives the rate of failure conditional on survival; it's a measure of risk.

Many people interpret the hazard rate as being akin (but not the same as) to a conditional probability. In other words, they interpret the hazard rate as being the probability that an observation dies at time t conditional on the observation having survived to time t. What's the logic behind this? According to the rules of conditional probability, the probability that an observation dies at time t conditional on having survived to time t is given by:

$$Pr(t \leq T \leq t+\Delta t | T \geq t) = \frac{f(t)\Delta t}{S(t)} = h(t)\Delta t \tag{10}$$

If you compare Eq. (10) and Eq. (9), you will see that the hazard function h(t) is almost the same as this conditional probability but not quite. This is because $h(t)$ refers to the exact time t at which an event occurs and not the tiny interval Δt thereafter. Thus, you might find it useful to think of the hazard rate in a continuous time setting as being akin (but not the same as) to a conditional probability.[5] Given the properties of $F(t)$ and $S(t)$, we know that:

$$h(t) \geq 0 \tag{11}$$

The hazard rate can vary from 0 (meaning no risk at all) to infinity (meaning the certainty of failure at that instant). The hazard rate can theoretically take on many shapes over time - it might be constant, it might increase/decrease over time, it might go up and then down, and so on. We'll come back to this point later.

From the hazard rate we can also generate something that looks like a CDF of the hazard by integrating over time.

$$H(t) \equiv \Lambda(t) = \int_0^t h(t)dt \tag{12}$$

This is referred to as the integrated hazard and can also be written as:

$$H(t) = -\ln[S(t)] \tag{13}$$

The integrated hazard rate measures the total amount of risk that has accumulated up to time t. An interpretation of the integrated hazard rate is that it records the number of times that we would expect to observe failures over a given time period, if only the failure event were repeatable. For example, imagine that we observed some unit for two years. In the first year, let's assume that the unit faced a constant hazard rate of 4 per year in the first year and that the unit faced a constant hazard rate of 2 per year in the second year. At the end of the two year period of observation, the integrated hazard would be 6. This means that we would expect the unit to have died 6 times in this two year period if each time the unit died it could instantly resurrect and continue to be observed. This is called the count-data interpretation of the cumulative hazard.

As you can see from these equations, the hazard rate, survivor function, distribution, and density functions are all mathematically linked. If any of these are specified, then the others are fully determined. For example, if $h(t)$ was defined, one would be able to calculate $f(t)$, $S(t)$, $F(t)$

[5] We will see that the hazard rate is a conditional probability in the discrete time duration setup.

and $H(t)$. To demonstrate this, I use the following example from Jenkins (2008) which starts with the hazard rate $h(t)$.

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)} \tag{14}$$

$$= \frac{-\partial[1 - F(t)]/\partial t}{1 - F(t)} \tag{15}$$

$$= \frac{\partial{-\ln}[1 - F(t)]}{\partial t} \tag{16}$$

$$= \frac{\partial{-\ln}[S(t)]}{\partial t} \tag{17}$$

where we take account of the fact that $\partial \ln[g(x)]/\partial x = g'(x)/g(x)$ and $S(t) = 1 - F(t)$. If we integrate both sides over time, we have:

$$\int_0^t h(u)du = -\ln[1 - F(t)] \mid_0^t \tag{18}$$

Since $F(0) = 0$ and $\ln 1 = 0$, we have:

$$\ln[1 - F(t)] = \ln[S(t)] = -\int_0^t h(u)du \tag{19}$$

$$S(t) = \exp\left(-\int_0^t h(u)du\right) \tag{20}$$

$$S(t) = \exp[-H(t)] \tag{21}$$

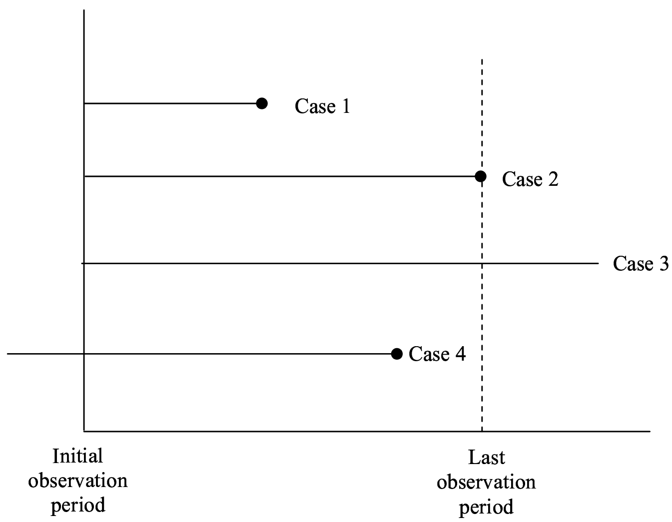
3.2 Estimation

With all this in hand, we can start to think about how to estimate a continuous time duration model. We will use maximum likelihood estimation. Let's start with the easy case where we have a random sample of observations that we observe from the beginning of their spell through to completion. In other words, let's start by ignoring issues such as sample selection, right-censoring, and left-truncation. In this situation, all we have to do is specify a probability density function, $f(t)$. We would then have the following sample likelihood function:

$$\begin{aligned} \mathcal{L}\{\beta|(t_1), \ldots, (t_n)\} &= \prod_{i=1}^{N} f(t_i) \\ &= \prod_{i=1}^{N} S(t_i)h(t_i) \\ &= \prod_{i=1}^{N} \mathcal{L}_i \end{aligned} \tag{22}$$

But what happens when we have censored or truncated observations? The basic issue with censoring and truncation is shown in Figure 1. Case 1 is observed until some time point, at which it then experiences an event. Case 2 is observed until the last observation period, at which it then experiences an event. Case 3 is observed up until the last observation period; however, at this point, case 3 is still 'surviving'. In other words, Case 3 is right-censored. Case 4 is left-truncated in that it enters the observation period already having amassed some 'history'; a portion of its duration is unobserved prior to the onset of the observation period.

Figure 1: Right-Censoring and Left-Truncation



3.2.1 Censoring

Let's think about censoring first. The important things to remember are (i) censoring is the removal of data for reasons other than the event of interest, (ii) censoring is defined by the researcher, (iii) censoring is typically assumed to be conditionally independent of both the event of interest and any covariates that are included i.e. exogenous, (iv) and that just because an observation is censored does not mean that it does not contain any information.

As we have seen, uncensored observations tell us both about the hazard of the event and the survival of individuals prior to the event; that is, the density of t_i for uncensored observations is $S(t_i)h(t_i)$. In other words, they tell us the exact time of failure. On the other hand, censored observations tell us only that the observation survived at least to time t^*, where t^* is the censoring point.[6] As a result, they only contribute information to the likelihood through their survival function, $S(t_i|X_i)$. In other words, censored observations contribute the probability that the observation

[6] In this example, we are assuming that the censoring point is the same for all units. However, this is not necessary. We can allow for different censoring points across different units.

survived at least up until the censoring point. Thus, the general likelihood for each observation is

$$\begin{aligned}\mathcal{L} &= \prod_{t_i \leq t^*} S(t_i)h(t_i) \prod_{t_i > t^*} S(t_i) \\ &= \prod_{t_i \leq t^*} f(t_i) \prod_{t_i > t^*} S(t_i) \end{aligned} \tag{23}$$

The likelihood function can be written to show exactly how the censored and uncensored observations are treated. We start by defining a censoring indicator, δ_i:

$$\delta_i = \begin{cases} 1 & \text{if } t_i \leq t^* \\ 0 & \text{if } t_i^* > t^* \end{cases}$$

When $\delta_i = 1$, then the observation is uncensored; when $\delta_i = 0$, then the observation is right-censored.

With this in hand, we can rewrite the likelihood dealing with censored observations as:

$$\mathcal{L} = \prod_{i=1}^{N} [f(t_i)]^{d_i} [S(t_i)]^{1-d_i} \tag{24}$$

and the log likelihood as:

$$\ln\mathcal{L} = \sum_{i=1}^{N} \{d_i \ln[f(t)] + (1-d_i)\ln[S(t)]\} \tag{25}$$

3.2.2 Truncation

In duration analysis, truncation is essentially a period over which the subject was not observed but is, a posteriori, known not to have failed. The main problem with left truncation is that had the case failed before our study started, we would not have observed the case at all. So, can we include a left-truncated case in a study of duration? Yes, but we have to take account of the fact that the case did not fail prior to our study starting. In effect, we start with our basic likelihood function (without censoring) *for individual* i:

$$\mathcal{L}_i\{\beta|(t_i)\} = S(t_i)h(t_i) \tag{26}$$

But then we must take account of the fact that the case had already survived up until time t_0, the start point of our study – this is just $S(t_{0i})$. And so, we rewrite the individual likelihood contribution of this unit as:

$$\mathcal{L}_i = \left[\frac{S(t_i)}{S(t_{0i})}\right] h(t_i) \tag{27}$$

because $\frac{S(t_i)}{S(t_{0i})}$ is the probability of surviving to t_i, given survival up to time t_{0i}. Note that Eq. (27) applies equally well to situations where there is no left truncation and the case enrolled at time $t_{0i} = 0$, since $S(0) = 1$. As a result, we can use Eq. (27) as a more general form of individual likelihood than what we had before. Thus, the likelihood for the sample (ignoring right-censoring for now) would be:

$$\mathcal{L} = \prod_{i=1}^{N} \left[\frac{S(t_i)}{S(t_{0i})}\right] h(t_i) \tag{28}$$

3.2.3 Censoring and Truncation

To take account of both censoring and truncation, we would have the following likelihood:

$$\begin{aligned} \mathcal{L} &= \prod_{i=1}^{N} \left[\frac{S(t_i)}{S(t_{0i})}\right]^{1-d_i} \left[\frac{f(t_i)}{S(t_{0i})}\right]^{d_i} \\ &= \prod_{i=1}^{N} \left[\frac{S(t_i)}{S(t_{0i})}\right] [h(t_i)]^{d_i} \end{aligned} \tag{29}$$

The log-likelihood function would be:

$$\ln\mathcal{L} = \sum_{i=1}^{N} \left\{ d_i \ln h(t_i) + \ln\left[\frac{S(t_i)}{S(t_{0i})}\right]\right\} \tag{30}$$

4 Discrete Time Duration Models – Intrinsically Discrete Data

4.1 Mathematical Components

In cases where the duration times are intrinsically discrete, the length of a spell for some unit (individual, government, country etc.) is a realization of a discrete random variable T with a probability mass function $f(t)$.

$$f(t) = \Pr(T = t_i) \tag{31}$$

where $i \in \{1, 2, 3, \ldots\}$. $f(t)$ is the unconditional probability of failing or experiencing an event at time t.

The survivor function – the probability of surviving beyond time t_i – is defined as:

$$S(t) = \Pr(T \geq t_i) = \sum_{j=i}^{\infty} f(t_j) \tag{32}$$

where j denotes a failure time. In other words, the probability of surviving up to and beyond t_i is simply the summation of the probabilities of failing at or after t_i.

The discrete time hazard rate, $h(t)$ is the conditional probability of having an event at t_i (conditional on having survived up to time t_i). Given the discrete nature of the time periods, this is equivalent to saying that the hazard rate is the conditional probability of having an event at t_i (conditional on having survived the period before t_i). This is given as:

$$\begin{aligned} h(t) &= \Pr(T = t_i | T \geq t_i) \\ &= \frac{f(t_i)}{S(t_{i-1})} \end{aligned} \tag{33}$$

Instead of thinking in terms of the conditional probability of failure, we might think in terms of the conditional probability of survival (conditional on survival up to some point in time).

$$1 - h(t) = \Pr(T > t_i | T \geq t_i) \tag{34}$$

From (33), we can see that $f(t) = h(t_i)S(t_{i-1})$. As we have seen, $h(t_i)$ is the probability of failing at time t_i conditional on having made it to t_i, and $S(t_{i-1})$ is the probability of surviving up through the period before t_i. It should be obvious that $S(t_{i-1})$ is simply the product of the probabilities of having survived beyond each of the discrete time points up through t_{i-1}. As a result, $f(t)$ – the unconditional probability of failure at t_i – can be written as:

$$\begin{aligned} f(t) = \Pr(T = t_i) = \Pr(T = t_i | T \geq t_i) \times \Pr(T > t_{i-1} | T \geq t_{i-1}) \times \ldots \\ \times \Pr(T > t_2 | T \geq t_2) \times \Pr(T > t_1 | T \geq t_1) \end{aligned} \tag{35}$$

where the subscripts 1 and 2 refer to the first and second time period. Thus, we can use Eq. (34) to reexpress Eq. (35) as:

$$\begin{aligned} f(t) &= h(t_i) \times (1 - h(t_{i-1})) \times \ldots \times (1 - h(t_2)) \times (1 - h(t_1)) \\ &= h(t_i) \prod_{i=1}^{t-1} (1 - h(t_i)) \end{aligned} \tag{36}$$

Thus, we can see that the unconditional probability of failing is simply the hazard probability multiplied by the product of the conditional survivor functions.

It is just as easy to see from (36), that the survivor function can be written as:

$$\begin{aligned} S(t) = \Pr(T \geq t_i) &= (1 - h(t)) \times (1 - h(t_{i-1})) \times \ldots \times (1 - h(t_2)) \times (1 - h(t_1)) \\ &= \prod_{i=1}^{t} (1 - h(t_i)) \end{aligned} \tag{37}$$

giving the intuition that the probability of surviving beyond t_i is simply equal to the conditional probability of surviving through t_i periods.

The discrete time failure function, which is the probability of failing or having an event before t_i is just:

$$\begin{aligned} F(t) = 1 - \Pr(T \geq t_i) = 1 - S(t) \\ = 1 - \prod_{i=1}^{t} (1 - h(t_i)) \end{aligned} \tag{38}$$

4.2 Estimation

So, how do we estimate this? Well, remember that the likelihood function for duration models with right censoring is:

$$\mathcal{L} = \prod_{i=1}^{n} \{f(t)\}^{y_{it}} \{S(t)\}^{1-y_{it}} \tag{39}$$

We can now substitute in our equations for $f(t)$ and $S(t)$ to get:[7]

$$
\begin{aligned}
\mathcal{L} &= \prod_{i}^{n} \{h(t_i)S(t_{i-1})\}^{y_{it}} \{S(t_i)\}^{1-y_{it}} \\
&= \prod_{i}^{n} \left[h(t_i) \prod_{i=1}^{t-1} (1 - h(t_i)) \right]^{y_{it}} \left[\prod_{i=1}^{t} (1 - h(t_i)) \right]^{1-y_{it}}
\end{aligned} \tag{41}
$$

Note that we are now using y_{it} rather than d_i to deal with right censoring issues. $y_{it} = 1$ indicates that the observation experiences an event or transition in period t and $y_{it} = 0$ indicates that the observation does not experience an event or transition in period t.[8]

Importantly, this likelihood function has the same basic structure as the likelihood function for a binary regression model in which y_{it} 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 discrete time period that a unit is at risk of experiencing an event or transition. An example of what the new data structure looks like is given in Table 1. The dependent variable would be Event Occurrence, but you can see how it is directly related to the duration variable, Time Elapsed. The only real

Table 1: Example of Discrete-Time Duration Data

Case ID	Event Occurrence	Year	Time Elapsed
1	0	1974	1
1	0	1975	2
⋮	⋮	⋮	⋮
1	0	1986	13
1	1	1987	14
5	1	1974	1
45	0	1974	1
45	0	1975	2
⋮	⋮	⋮	⋮
45	0	1992	19
45	0	1993	20

difference between this duration data and the continuous time duration data that we have seen before is that the duration data here is disaggregated into discrete time units. As you can see,

[7]It turns out that this can also be written as:

$$
\begin{aligned}
\mathcal{L} &= \prod_{i}^{n} \left[\left(\frac{h(t_i)}{1 - h(t_i)} \right) \prod_{i=1}^{t-1} (1 - h(t_i)) \right]^{y_{it}} \left[\prod_{i=1}^{t} (1 - h(t_i)) \right]^{1-y_{it}} \\
&= \prod_{i}^{n} \left[\left(\frac{h(t_i)}{1 - h(t_i)} \right)^{y_{it}} \prod_{i=1}^{t} (1 - h(t_i)) \right]
\end{aligned} \tag{40}
$$

[8]Whereas we had a right-censoring indicator in the continuous duration model, we can see that the dependent variable is already an implicit indicator of right censoring in the discrete time model.

discrete time duration data looks a lot like Binary dependent variable Time-Series Cross-Section (BTSCS) data. In fact, if you had BTSCS data, you would likely want to treat it as if it were discrete time duration data (Beck, Katz & Tucker 1998). In effect, we have binary data that are also event history data. Given the same basic structure as the likelihood function for a binary regression model, we are going to be able to estimate discrete time duration models using things like logit and probit. This makes our job a lot easier.

5 Discrete Time Duration Models – Grouped (Interval-Censored) Data

Notes to come.

References

Beck, Nathaniel, Jonathan Katz & Richard Tucker. 1998. "Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable." *American Journal of Political Science* pp. 1260–1288.

Jenkins, Stephen P. 2008. "Survival Analysis." Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester.