Introduction to Duration Models

1 What is Duration Analysis?

Duration analysis is sometimes referred to as survival analysis or event history analysis. Duration data can be thought as being generated by what is called a 'failure time process'. A failure time process consists of units – individuals, governments, countries, and so on – that are observed at some starting point in time. These units are in some state – the individual is healthy, the government is in power, a country is at peace, and so on – and are then observed over time. At any given point in time, these units are 'at risk' of experiencing some event, where an 'event' essentially represents a change or transition to another state – the individual dies, the government falls from power, a country is at war, and so on. After the event is experienced, the unit is either no longer observed or it is at risk of experiencing another kind of event. In some circumstances, units are not observed experiencing an event; that is, no transition is made from one state to another while the unit is being observed – the individual remains healthy, the government remains in power, the country remains at peace, and so on. As we will see, we call these cases 'censored' since we do not observe the subsequent history of the unit after the last observation point. This process is called a 'failure time process' because (a) units are observed at an initial point in time, (b) the unit survives for some length of time (or spell), and (c) then the unit 'fails' or is 'censored'.

2 Some Intuition

2.1 The Basics of Duration Analysis

In duration analysis, we are typically interested in some event occuring.[1] For example, we might be interested in one of the following events:

- Government collapses
- War starts
- Cabinet reshuffle
- Democracy emerges
- Policy is adopted
- Job loss
- and so on

[1] These notes on the intuition behind duration analysis are based on notes from Brad Jones, August 2005, ICPSR MLE-2.

In general, we think of events as being probabilistic. And so we might ask:

- What are the chances that the government collapses?
- How likely is it for war to start?
- Will there be a cabinet reshuffle?
- What are the chances that democracy emerges
- and so on

Given this setup, we might want to incorporate time into these questions. If we do so, then 'chance' essentially becomes synonymous with 'risk'. In other words, instead of saying 'What are the chances that the government collapses?', we would say, 'What is the risk that the government collapses?' Below are some more examples of what I mean.

- Duration of Government
  - Event: Government collapses
  - Timing: Days in office
  - Risk: 'Given a government has stayed in office for 340 days, what is the risk that it will now lose office?'
- Duration of peace
  - Event: War onset
  - Timing: Years of peace
  - Risk: 'Given that a country has been at peace for 14 years, what is the likelihood or risk that it will be at war next year?'

As these examples illustrate,

Timing implies Risk: Given that something has not yet happened, what are the chances it will happen subsequently?

But, what exactly is risk? Risk is captured in the following formula:

$$\text{Risk} = \frac{\text{Chance that Something Happens}}{\text{Chance that it hasn't Happened Yet}}$$

In other words, 'risk' is a ratio – a relationship between the chances that something *can* happen relative to the chances that it has not happened yet. A slightly different way to write this ratio, which will be helpful as you will see, is:

$$\text{Risk} = \frac{\text{P(Failure)}}{\text{P(Survival)}}$$

With this in hand, I should provide some definitions.

- Failure: The unconditional probability that an event will occur.
- Survival: The probability that 'up until now' the event has not yet occurred.
- Risk: The conditional failure rate – given that the event has not yet occurred, what are the chances that it will occur?

And so the risk ratio can be written as:

$$R = \frac{\mathrm{F}}{\mathrm{S}}$$

Conventionally in duration analysis, the risk ratio is called the HAZARD RATIO. As we will see, the hazard ratio is at the center of duration analysis.

Typically, scholars will follow something like the procedure outlined below to conduct their duration analysis:

1. Develop a theory linking various factors (independent variables) to some event. For example, scholars might develop a theory linking factors such as the number of parties in government to government collapse.
2. Observe some 'sample' over time. For example, scholars might look at different governments in a country and/or across countries.
3. Record whether some event of interest occurs over time. Does a government collapse?
4. Collect data on the independent variables
5. Model the 'event' or 'time until the event' as a function of the independent variables.

2.2 The Basics of Duration Data

So, what does basic duration data look like? In Figure 1, I show data from the Constitutional Change and Parliamentary Democracy (CCPD) project on the duration of governments in Belgium from 1946 to 1998.[2] The data indicate the name of the prime minister, the parties in the government, the date the government started, the date the government ended, the duration of the government, whether the government was a minimal winning coalition or not, and whether the government ended (event). The EVENT variable represents the transition from one state (one government) to another (a different government). The premise of duration analysis is to model both the duration of time spent in the initial state and the transition to a subsequent state. Units not experiencing an event by the last observation period are called 'right-censored'. As you can see in Figure 1, the observation for government 33 in Belgium is right-censored – we know when it started (June 23, 1995) but this

[2]The Constitutional Change and Parliamentary Democracy (CCPD) project can be found at `http://www.pol.umu.se/ccpd/CCPD/index.asp`.

Figure 1: Duration of Governments in Belgium

	cabinetcode	countryname	cabinet_name	cabinet_parties	date_in	date_out	duration	mwc	event
1	201	belgium	Spaak	PSB/BSP	460313	460320	7	0	1
2	202	belgium	Van Acker III	PSB/BSP, LP/PL, PCB/KPB	460331	460709	100	1	1
3	203	belgium	Huysmans	PSB/BSP, LP/PL, PCB/KPB	460803	470313	222	1	1
4	204	belgium	Spaak II	CVP/PSC, PSB/BSP	470320	490627	830	1	1
5	205	belgium	Eyskens	CVP/PSC, LP/PL	490811	500318	219	1	1
6	206	belgium	Duvieusart	CVP/PSC	500608	500811	64	0	1
7	207	belgium	Pholien	CVP/PSC	500816	520109	511	0	1
8	208	belgium	Van Houtte	CVP/PSC	520115	540412	818	0	1
9	209	belgium	Van Acker IV	PSB/BSP, LP/PL	540422	580602	1502	1	1
10	210	belgium	Eyskens II	CVP/PSC	580623	581104	134	0	1
11	211	belgium	Eyskens III	CVP/PSC, LP/PL	581106	610327	872	1	1
12	212	belgium	Lefevre	CVP/PSC, PSB/BSP	610425	650524	1490	1	1
13	213	belgium	Harmel	CVP/PSC, PSB/BSP	650727	660211	199	1	1
14	214	belgium	Van den Boeynants I	CVP/PSC, LP/PL	660319	680207	690	1	1
15	215	belgium	Eyskens IV	CVP, PSC, PSB/BSP	680617	711108	1239	0	1
16	216	belgium	Eyskens V	CVP, PSC, PSB/BSP	720121	721123	307	0	1
17	217	belgium	Leburton	PSB/BSP, CVP, PSC, VLD, PRL	730126	740119	358	0	1
18	218	belgium	Tindemans	CVP, PSC, VLD, PRL	740425	740611	47	0	1
19	219	belgium	Tindemans II	CVP, PSC, VLD, PRL, RW	740612	770304	996	1	1
20	220	belgium	Tindemans III	CVP, PSC, VLD, PRL	770306	770418	43	1	1
21	221	belgium	Tindemans IV	CVP, PSC, PSB/BSP, VU, FDF	770603	781011	495	0	1
22	222	belgium	Van den Boeynants II	CVP, PSC, PSB/BSP, VU, FDF	781020	781218	59	0	1
23	223	belgium	Martens I	CVP, PSC, PS, SP, FDF	790403	800116	288	0	1
24	224	belgium	Martens II	CVP, PSC, PS, SP	800123	800409	77	0	1
25	225	belgium	Martens III	CVP, PSC, PS, SP, VLD, PRL	800518	801007	142	0	1
26	226	belgium	Martens IV	CVP, PSC, PS, SP	801022	810402	162	1	1
27	227	belgium	M Eyskens	CVP, PSC, PS, SP	810406	810921	168	0	1
28	228	belgium	Martens V	CVP, PSC, VLD, PRL	811217	851014	1397	1	1
29	229	belgium	Martens VI	CVP, PSC, VLD, PRL	851128	871214	746	1	1
30	230	belgium	Martens VII	CVP, PSC, PS, SP, VU	880509	910929	1238	0	1
31	231	belgium	Martens VIII	CVP, PSC, PS, SP	910929	911125	57	0	1
32	232	belgium	Dehaenen I	CVP, PSC, PS, SP	920307	950521	1170	1	1
33	233	belgium	Dehaene II	CVP, PSC, PS, SP	950623	.	.	1	0

government was still in power when the CCPD project stopped observing governments in Belgium (December 31, 1998) and, as a result, we do not know when it ended.[3]

The data shown in Figure 1 are relatively straightforward and are what we call 'single-spell' data because each government is only at risk of collapse once - when the government collapses, it leaves the data set. Later on, we will look at 'multiple-spell' data in which units are at risk of an event multiple times. We will also look at data in which units are at risk of multiple (different) events, possibly multiple times.

2.3 Why not OLS?

As Figure 1 illustrates, we have a variable indicating whether the government is a minimal winning coalition (MWC) or not. There are various arguments in political science suggesting that minimal winning coalition governments are likely to last longer than other types of governments. Suppose

[3]In some applications of duration analysis, we have what is called 'left-truncation'. Left truncation occurs when we do not observe the time-of-origin for an observation. In other words, an observation enters the study already in process. In our example of government duration in Belgium, left-truncation would involve something like the following – when we started observing Belgian government's in January 1945, there was already a government in place but for some reason we did not know when that government first came to office.

Figure 2: Duration of Governments in Belgium

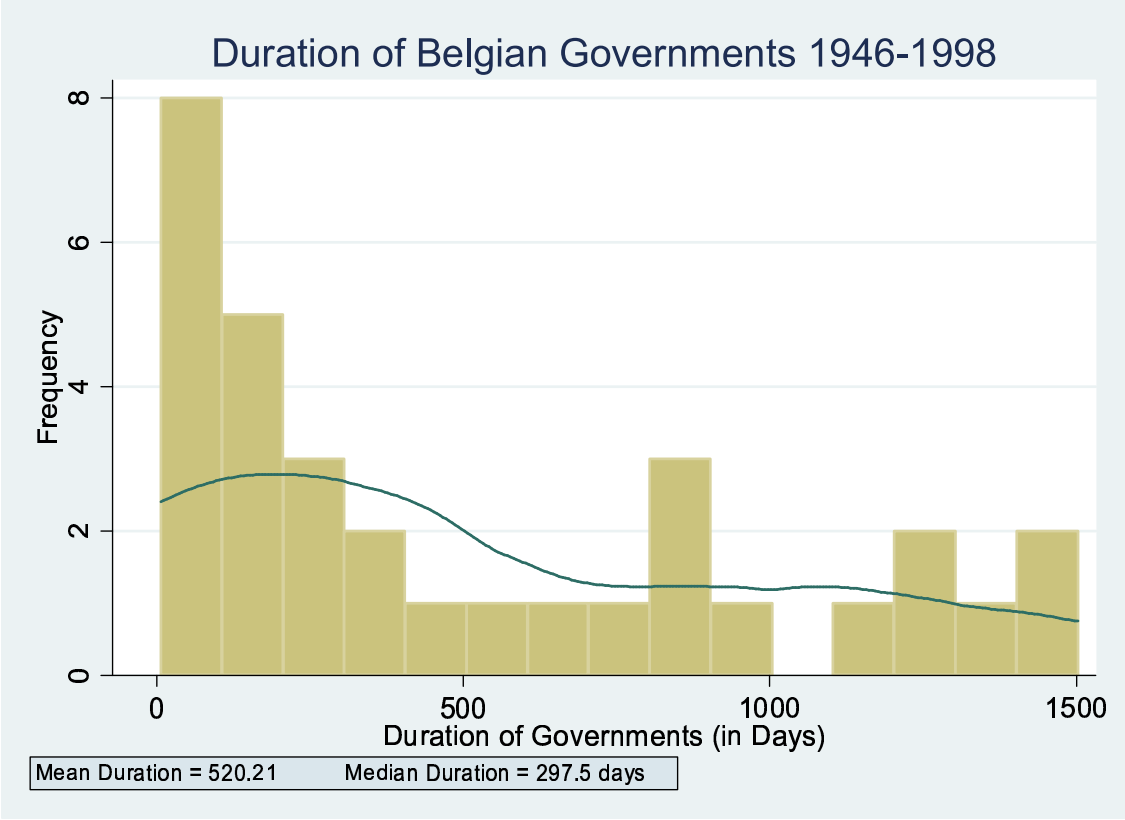
```
. regress  duration mwc;

      Source |       SS       df       MS              Number of obs =      32
-------------+------------------------------           F(  1,    30) =    4.72
       Model |  1008392.83     1  1008392.83           Prob > F      =  0.0379
    Residual |  6410742.64    30  213691.421           R-squared     =  0.1359
-------------+------------------------------           Adj R-squared =  0.1071
       Total |  7419135.47    31  239326.951           Root MSE      =  462.27

------------------------------------------------------------------------------
    duration |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mwc |   355.7294   163.7565     2.17   0.038      21.2941    690.1647
       _cons |   353.4706   112.1164     3.15   0.004     124.4984    582.4428
------------------------------------------------------------------------------
```

we wanted to test to see whether this is indeed the case using data from Belgium. What should our modeling strategy be? Well, you might think to use ordinary least squares (OLS) and use the DURATION variable as the dependent variable and the MWC variable as the independent variable. You would then interpret the results in the standard way. For example, the positive and significant coefficient on MWC tells us that governments that are minimal winning coalitions have a significantly longer duration (survival time) in office than governments that are not minimal winning coalitions. Specifically, the expected duration (survival time) of a government that is a minimal winning coalition has historically been 355.7 days longer than a government that is not a minimal winning coalition. This indicates that governments that are minimal winning coalitions have historically had a higher 'survival' rate, thereby implying that they have historically had a lower 'risk' or 'hazard rate'.

Although we obviously could use OLS to test our hypothesis, we almost certainly shouldn't. As I'll now demonstrate, there are a number of problems with using OLS when we have duration data. These problems nearly always necessitate moving towards some kind of modeling strategy that is specifically designed for duration data. So, what are the problems with OLS?

1. **Normality Assumption**
   OLS assumes that the duration times (conditional on the independent variables) are normally distributed. This assumption is nearly always unrealistic in the context of duration data where data often exhibit asymmetry, particularly if some observations have very long durations (right skewed). Among other things, this means that the median will describe the central tendency better than the mean. If you look at Figure 3, this is exactly the case with the data on the duration of governments in Belgium. One common fix to this problem is to transform the dependent variable by taking the natural log and then applying OLS. For example, you would have $\ln y_i = \beta x_i + \epsilon_i$. Although this mitigates the skewness problem, it does not solve other serious problems that we will now discuss.

Figure 3: Duration of Governments in Belgium



2. **Negative Predicted Values**
   OLS may return negative predicted values even though this is impossible – survival times must be positive.

3. **Censoring**
   OLS does not easily distinguish between 'censored' and 'uncensored' observations. Dropping censored observations could lead to sample selection problems.[4]

4. **Time Varying Covariates**
   OLS cannot easily accommodate independent variables that change value over time (TVCs).

These problems are important and suggest we should move towards a modeling strategy that is specifically designed for duration data. But what strategy would that be?

[4]Typically, right-censored observations can be dealt with using something like a tobit model.

2.4 Parametric, Semi-Parametric, and Non-Parametric Approaches

There are essentially three main approaches: parametric, semi-parametric, and non-parametric models.

2.4.1 Parametric Models

The main problem with OLS has to do with the assumption that the disturbances are normally distributed.[5] Substituting in a more reasonable distributional assumption for the disturbances leads to the adoption of parametric duration models such as the exponential, weibull, log-logistic, and so on. One potential problem with these models is that we are still having to make an assumption about the distribution of the disturbances and, hence, the distribution of the survival times. As a result, we might want to look for a modeling strategy that does not require any assumption about the distribution of the survival times. How can we do this?

2.4.2 Semiparametric Models

The key to removing the distributional assumption is that, because events occur at given points in time, they can be ordered. We can then conduct our analysis using the ordering of the survival times exclusively. Let me demonstrate this point. In Table 1, I present failure times for five units where x is some independent variable.

Table 1: Sample Duration Data

time	x
1	3
5	2
9	4
20	9
22	10

Suppose we now ask the following question: 'What is the probability of failure after the exposure to the risk of failure for one unit of time?' This reduces the analysis to a basic binary outcome analysis. See Table 2. We might then conduct a logit or probit analysis where OUTCOME is the dependent variable.

```
logit outcome x
```

Obviously, this would be an extremely inefficient use of the data. The advantage of this approach, though, is that there is no need to make an assumption about the distribution of survival (failure) times. Note, though, that there is nothing magical about the first failure time. We could have

[5] Much of this material is based on Cleves, Gould, & Gutierrez (2004).

Table 2: Sample Duration Data

time	x	outcome
1	3	1
5	2	0
9	4	0
20	9	0
22	10	0

chosen to analyze the second failure time, which in this case is at TIME = 5. This would have meant asking, 'What's the probability of failure, given exposure to 5 units of time?' To conduct such an analysis, we would have to drop the first observation since this observation was not exposed to failure for 5 units of time, recode the OUTCOME variable, and then run our logit or probit model.

```
drop outcome
generate outcome = cond(time==5,1,0) if time>=5
logit outcome x if time>=5
```

We could repeat this procedure for all of the different failure times. You might wonder whether it is possible to combine all of these different analyses and constrain the appropriate regression coefficients, such as the coefficient on x, to be the same. The answer is that this is possible. Such an approach leads to semiparametric duration analysis and, specifically, to the Cox model if a conditional logit model is fit for each analysis.[6] Note that time plays no role in this analysis other than to indicate the ordering of the observations. For example, no account is taken of the fact that the second observation in Table 2 fails in period 5 whereas the first observation failed in period 1; all that matters is that the second observation failed after the first one. The approach to duration analysis that I have just indicated goes under the name of semiparametric analysis. The reason is that this approach is non-parametric when it comes to time, but it is still parametric in the sense that we are still parameterizing the effect of X. In other words, there is still a parametric component (logit, probit, etc.) to the analysis.

2.4.3 Nonparametric Models

An entirely non-parametric approach does away with assumptions about how each unit's observed X value determines the probability of failure. Nonparametric methods that are used elsewhere in the social sciences such as lowess often have difficulty with duration data because they cannot deal adequately with censoring and other issues. That said, when there are no independent variables, or the independent variables are qualitative in nature, then we can use certain nonparametric methods to estimate things like the probability of survival past a certain point in time. Nonparametric models do not make assumptions about (a) the distribution of failure times or (b) how independent variables change survival experiences.

[6]The only difference between the conditional logit model and the logit model we have used here is that the conditional logit model conditions on the fact that 'outcome==1' for one and only one observation within each separate analysis.

2.4.4 Connecting the Approaches

The semiparametric approach that we just looked at basically conducted a series of analyses:

P(failure after exposure for (exactly) 1 unit of time)

P(failure after exposure for (exactly) 5 unit of time)

P(failure after exposure for (exactly) 9 unit of time)

P(failure after exposure for (exactly) 20 unit of time)

P(failure after exposure for (exactly) 22 unit of time)

But why not try to make this approach more efficient by also adding other analyses:

P(failure after exposure for (exactly) 1.1 unit of time)

P(failure after exposure for (exactly) 1.2 unit of time)

P(failure after exposure for (exactly) 1.3 unit of time)

...

The problem is that to conduct these analyses, we would need to make an assumption about the distribution of failure times. Basically, this is what parametric models do. As Cleves, Gould, and Guttierrez (2004, 6) note, 'Semiparametric analysis is nothing more than a combination of separate binary-outcome analyses, one per failure time, while parametric analysis is a combination of several analyses at *all* possible failure times.' Parametric analyses will be more efficient (standard errors will be smaller) so long as the distributional assumptions that are made are appropriate.

When no covariates are present, semiparameric methods such as the Cox model will produce estimates of relevant quantities such as the probability of survival past a certain point in time that are identical to the nonparametric estimates. If the covariates are qualitative in nature, parametric and semiparametric models should yield more efficient tests and comparisons of the groups than nonparametric methods, and these tests should agree. Should the tests disagree, then this is a signal that some of the assumptions made by the parametric and semiparametric models are incorrect.

References

Cleves, Mario A., William W. Gould & Roberto G. Gutierrez. 2004. *An Introduction to Survival Analysis Using STATA*. Texas: STATA Corporation.