

# Intermediate Social Statistics

## Lecture 8: Duration Models

Raymond Duch  
Nuffield College, Oxford University  
[www.raymondduch.com](http://www.raymondduch.com)

March 8, 2008

# 1 Motivation

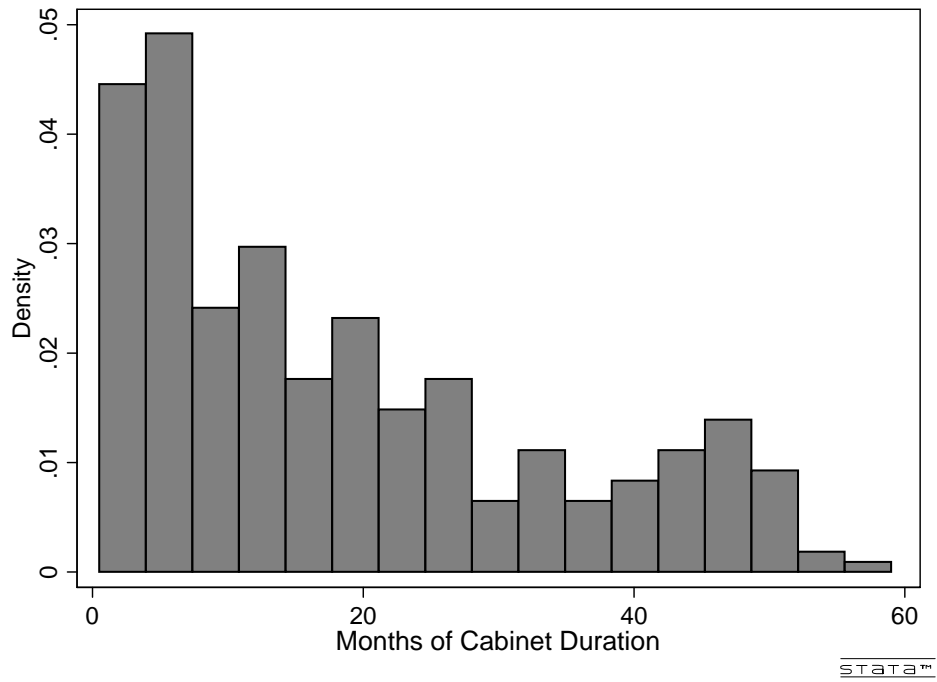
1. Why OLS might not be appropriate for estimating duration models.
2. Censuring
3. Time varying independent variables

# 2 Cabinet Duration as an Illustration

1. The puzzle is explaining why cabinets last longer in one context than another.

	DURAT	BELGIUM	SALIEN	OPPINFL	ROPINFL	VOLAT	RESPONSE
1.	3	yes	190.4412	high	4	93	split, majority winner
2.	7	yes	190.4412	high	4	93	split, majority winner
3.	20	yes	190.4412	high	4	93	winner only
4.	6	yes	190.4412	high	4	93	winner only
5.	7	yes	190.4412	high	4	93	winner only
6.	2	yes	261.2745	high	4	62	winner only
7.	17	yes	261.2745	high	4	62	winner only
8.	27	yes	261.2745	high	4	62	winner only
9.	49	yes	261.2745	high	4	62	even split
10.	4	yes	261.2745	high	4	62	winner only
11.	29	yes	261.2745	high	4	62	even split
12.	49	yes	365.4412	high	4	106	losers only
13.	6	yes	365.4412	high	4	106	losers only
14.	23	yes	365.4412	high	4	106	even split
15.	41	yes	365.4412	high	4	106	losers only
16.	10	yes	261.2745	high	4	71	even split
17.	12	yes	261.2745	high	4	71	split, majority loser
18.	2	yes	261.2745	high	4	71	even split
19.	33	yes	261.2745	high	4	71	split, majority loser
20.	1	yes	261.2745	high	4	71	even split
21.	16	yes	261.2745	high	4	71	even split
22.	2	yes	261.2745	high	4	71	even split
23.	9	yes	261.2745	high	4	71	split, majority winner
24.	3	yes	232.1079	high	4	97	even split
25.	5	yes	232.1079	high	4	97	split, majority winner

Figure 1: Cabinet Duration Data from King et al



### 3 The Probability Distribution for Analyzing Duration Data

The **Survival Function** is the probability that the duration of some episode (a cabinet for example) is at least  $t$ , and that the event by which the current episode comes to an end occurs later than  $t$ .

$$S(t) = Pr(T \geq t) = 1 - F(t) \tag{1}$$

which gives the probability that the episode (the cabinet) is of length (duration) at least  $t$ .

The **Hazard Rate** gives the rate at which episodes are completed at duration  $t$ , given that they lasted until  $t$ .

$$h(t) = \lim_{\Delta \rightarrow 0} \frac{Pr(t \geq T < t + \Delta | T \geq t)}{\Delta} = \lim_{\Delta \rightarrow 0} \frac{F(t + \Delta) - F(t)}{S(t)} = \frac{f(t)}{S(t)} \tag{2}$$

The issue then is what distribution should we use for  $F(\cdot)$

### 4 The Exponential Distribution

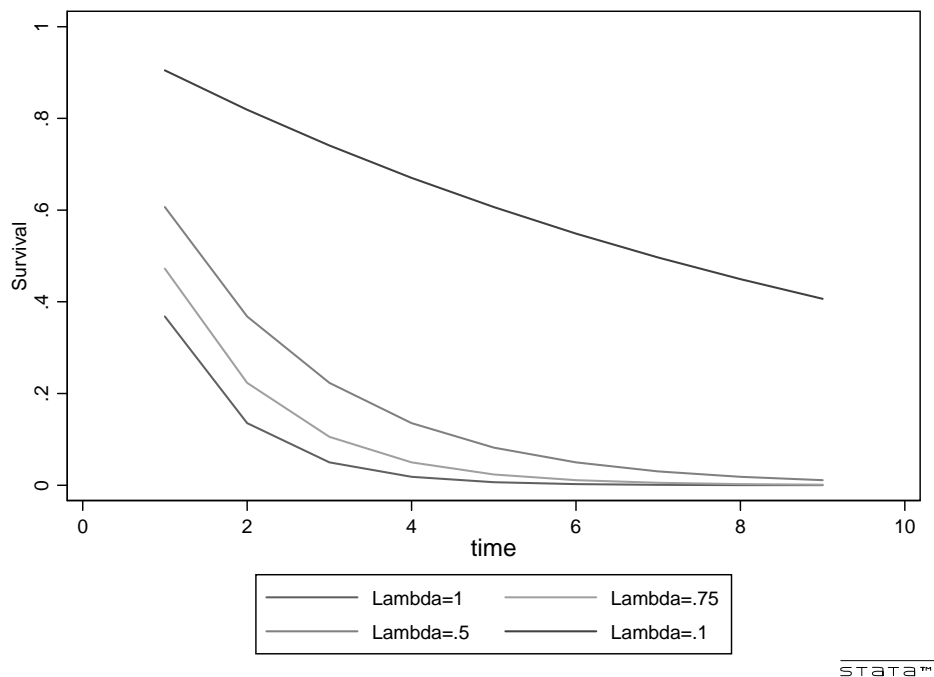
This is the simplest distribution for modeling duration data<sup>1</sup>:

The hazard function  $h(t)$  is simply  $\lambda$ , where  $\lambda = e^{-\lambda t}$  and the survival function  $S(t)$  is  $e^{-\lambda t}$ .

---

<sup>1</sup>I am using the accelerated failure time parameterization of this model rather than Proportional Hazards

Figure 2: Exponential Survival



```

. streg, distribution(exponential) time

      failure _d:  1 (meaning all fail)
analysis time _t:  DURAT

```

```

Iteration 0:  log likelihood = -460.73222
Iteration 1:  log likelihood = -460.73222

```

Exponential regression -- accelerated failure-time form

```

No. of subjects =          313                Number of obs   =          313
No. of failures =          313
Time at risk   =          5789
Log likelihood  = -460.73222                LR chi2(0)        =          0.00
                                                Prob > chi2       =          .

```

```

-----+-----
      _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      _cons |   2.917512   .0565233    51.62  0.000    2.806728    3.028295
-----+-----

```

```

.
end of do-file

```

The hazard rate is  $\lambda = e^{-2.917} = .054$ . The expected duration, expressed in months, of a typical cabinet government in this data set is then

$$E(Y_i) = \frac{1}{\lambda} = \frac{1}{e^{-2.917}} = 18.5$$

## 5 Adding a Systematic Component to the Exponential Distribution

The Hazard Function incorporating explanatory variables is

$$h(t) = \lambda_i = e^{-(\beta_0 + \beta_k \mathbf{x}_i)} \quad (3)$$

The Survivor Function then can be estimated as.

$$S(t) = \exp(-e^{\beta_0 + \beta_k \mathbf{x}_i} t) \quad (4)$$

```
. streg BELGIUM-SWEDEN, distribution(exponential) time
```

```
      failure _d: 1 (meaning all fail)
      analysis time _t: DURAT
Iteration 0:   log likelihood = -460.73222
Iteration 1:   log likelihood = -427.33351
Iteration 2:   log likelihood = -422.4328
Iteration 3:   log likelihood = -422.37544
Iteration 4:   log likelihood = -422.37542
```

```
Exponential regression -- accelerated failure-time form
```

```
No. of subjects =          313          Number of obs   =          313
No. of failures =          313
Time at risk    =          5789
Log likelihood  = -422.37542
LR chi2(14)     =          76.71
Prob > chi2     =          0.0000
```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
BELGIUM	-.6171175	.3054608	-2.02	0.043	-1.21581 - .0184254
CANADA	-.0073727	.3483152	-0.02	0.983	-.690058 .6753126
DENMARK	-.3548504	.3170019	-1.12	0.263	-.9761626 .2664619
FINLAND	-.7440277	.3017973	-2.47	0.014	-1.335539 -.152516
FRANCE	-1.832701	.3054608	-6.00	0.000	-2.431394 -1.234009
ICELAND	-.0637419	.3429971	-0.19	0.853	-.7360039 .6085202
IRELAND	.0442898	.3542459	0.13	0.901	-.6500194 .738599
ISRAEL	-.6717962	.3170019	-2.12	0.034	-1.293108 -.0504839
ITALY	-1.077935	.288468	-3.74	0.000	-1.643322 -.5125485
NETHER	-.1903214	.3429972	-0.55	0.579	-.8625835 .4819407
NORWAY	-.19697	.3298841	-0.60	0.550	-.8435309 .449591
PORTUG	-1.015143	.3985267	-2.55	0.011	-1.796241 -.2340449
SPAIN	-.0180909	.6262242	-0.03	0.977	-1.245468 1.209286
SWEDEN	-.19697	.3298841	-0.60	0.550	-.8435309 .449591
_cons	3.385387	.2425356	13.96	0.000	2.910026 3.860748

Having added a systematic component to the exponential hazard function we can now calculate the expected cabinet duration for each of the countries in our data set.

The UK case for example (which is captured by the intercept) is.

$$E(Y_i) = \frac{1}{\lambda} = \frac{1}{e^{-3.385}} = 29.5$$

And the Italian case is.

$$E(Y_i) = \frac{1}{\lambda} = \frac{1}{e^{-(3.385-1.08)}} = 10.05$$

	country	meantime	haz
1.	Belgium	15.93103	.0627706
2.	Canada	29.31249	.0341151
3.	Denmark	20.70833	.0482897
4.	Finland	14.03226	.0712644
5.	France	4.724138	.2116788
6.	Iceland	27.70588	.0360934
7.	Ireland	30.86665	.0323974
8.	Israel	15.08333	.0662983
9.	Italy	10.04878	.0995146
10.	Netherlands	24.41176	.0409639
11.	Norway	24.25	.0412371
12.	Portugal	10.7	.0934579
13.	Spain	29	.0344828
14.	Sweden	24.25	.0412371
15.	UK	29.52941	.0338646

Rather than simply distinguishing the hazard rates by country we can incorporate features of these countries as independent variables in the hazard function.

$$\lambda(t) = \exp(-[\beta_0 + \beta_1 \text{Country Attributes} + \beta_2 \text{Party Structure Attributes} + \beta_3 \text{Coalition Attributes}]) \quad (5)$$

## 6 Censoring

Observations of events are very often censored. Censoring occurs when the information about the duration in the origin state is incompletely recorded. *Left Censoring* occurs when the starting times of an episode are located before the beginning of the observational window. *Right Censoring* occurs because the observation is terminated at the right-side of the observation window.

In the cabinet duration example, right censoring is a consideration that needs to be incorporated into the model. There are two distinct types of observed cabinet durations.

1. The largest group of observed coalitions are those that break apart due to some critical events. These are *uncensored observations* for which we have exact time of failure. We will define  $d_i = 1$  for these uncensored observations.
2. The second group consists of those that come close to their constitutional inter-election period (CIEP). These are the *censored observations* and we only have information on the fact that they have survived at least to time  $T_i$ . We will define  $d_i = 0$  for these uncensored observations.

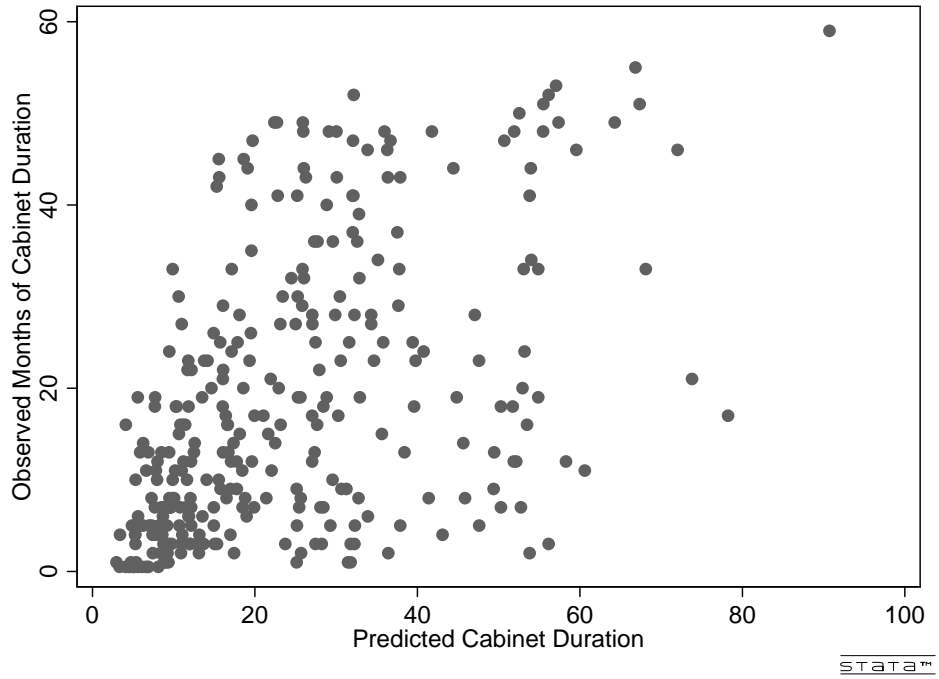
The complete censored likelihood is then a mixture of continuous (observed) and discrete (censored) observations.

$$\begin{aligned}
 \ln L &= \sum_{i=1}^N \left\{ d_i \ln \left[ e^{-\mathbf{X}_i \beta} e^{-(e^{-\mathbf{X}_i \beta} t)} \right] + (1 - d_i) \ln \left[ e^{-(e^{-\mathbf{X}_i \beta} t)} \right] \right\} \\
 &= \sum_{i=1}^N \left\{ d_i [(-\mathbf{X}_i \beta)(-e^{-\mathbf{X}_i \beta} t)] + (1 - d_i)(-e^{-\mathbf{X}_i \beta} t) \right\} \tag{6}
 \end{aligned}$$

where  $d_i$  takes on the value of one for uncensored observations, and zero otherwise. Hence notice the only unknown in the likelihood function is  $\beta$ .



Figure 3: Predicted and Observed Cabinet Durations from Censored Exponential Model



## 7 Weibull Hazard Models

Hazard Function (AFT)

$$h(t_i) = p\lambda_i t_i^{p-1} \quad (7)$$

where

$$\lambda_i = \exp(-p\mathbf{x}_i\beta)$$

Survivor Function (AFT)

$$S(t_i) = \exp(-\lambda_i t_i^p) \quad (8)$$

Note:  $p$  is the shape parameter in these models.

## 8 Readings for Duration Models

Allison, Paul D. 1986. *Event History Analysis* Newbury Park: Sage. Chapters 1-2.

Klein, John P. and Melvin L. Moeschberger. 1997. *Survival Analysis*. New York: Springer. Chapters 2-3, 8, 12.

Blossfeld, Hans-Peter, Alfred Hamerle and Karl Ulrich Mayer. 1989. *Event History Analysis: Statistical Theory and Applications in the Social Sciences*. Hillsdale, NJ: Erlbaum. Chapters 1-4.

King, Gary, James Alt, Michael Laver and Nancy Burns. 1990. "A unified model of cabinet dissolution in parliamentary democracies" *American Journal of Political Science*, 34: 847-71.

# Homework Questions

(Due Friday of Week 9, Hilary Term)

## Question 1

I have estimated an ordered multinomial logistic model to explore how characteristics of citizens impact their vote choice. I surveyed 1000 citizens before an election in which the Socialists, Conservatives, and Christian Democrats were running. The main explanatory variables are left/right, ideology and gender.

Data: surveys responses for 1000 respondents

Dependent Variable: coded 1 if voter voted for Socialists , 2 for Conservatives, 3 Christian Democrats

Independent variables:

ideology = scale of 1 to 10 with 1 being most left and 10 most right (mean = 5, sd = 2)

male = 1 identifies respondent was a male

Table 1: Results

	Coefficient
<hr/> <hr/>	
Socialist	
Ideology	-0.3
Male	0.3
Constant	0.88
<hr/> <hr/>	
Conservative	
Ideology	0.2
Male	0.02
Constant	-0.9
<hr/> <hr/>	

Baseline group is Christian Democrats

1. What is the stochastic component of this model?
2. What is the systematic component of this model?

3. What is the predicted probability of voting for each party for a male with ideology score 5?

## Question 2

I have replicated below the poisson regression results for the veto override example employed in class.

```
Poisson regression                               Number of obs   =          26
                                                LR chi2(4)      =          29.21
                                                Prob > chi2     =          0.0000
Log likelihood = -37.907938                    Pseudo R2       =          0.2781
```

nover	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
nveto	.0406958	.0090975	4.47	0.000	.0228651	.0585265
janpop	-.0296944	.0158098	-1.88	0.060	-.0606811	.0012923
presmaj	-1.167975	.6312272	-1.85	0.064	-2.405158	.0692071
pressmaj	.0084897	.4738788	0.02	0.986	-.9202958	.9372751
_cons	1.718194	.8871924	1.94	0.053	-.0206716	3.457059

1. What is the stochastic component of this model?
2. What is the systematic component of this model?
3. Here are the summary statistics for the independent variables employed in that example.

```
. summarize nover nveto janpop presmaj pressmaj
```

Variable	Obs	Mean	Std. Dev.	Min	Max
nover	26	1.730769	2.050516	0	8
nveto	26	15.61538	14.54119	0	70
janpop	26	58.23077	11.09705	36	74
presmaj	26	.3846154	.4961389	0	1
pressmaj	26	.5	.509902	0	1

What is the expected number of veto overrides for a President over the course of a Congressional term that had the following political characteristics:

4. the president's party controls neither the House nor the Senate.
5. the president's approval rating stands at 35 percent.
6. congress has exercised 35 vetoes.

What is the probability of 1 veto override in this context?

### Question 3

The dataset `conflict.dta` on the website is a subset of the Cross-National Time-Series Data Archive compiled by Arthur Banks. The dataset has the following variables:

**year** year

**country** country name

**popdens** population density

**defexpgdp** defense expenditure as a proportion of national expenditure

**phonespc** telephones per 100,000 people

**tvpc** televisions per 100,000 people

**gdppcfc** gdp/capita (factor costs)

**conflictevent** number of conflict events (riots, guerrilla wars, revolutions, assassinations, coups and government crises)

1. The variable `conflictevent` will be your dependent variable. Without running a regression, assess if the Poisson model will be sufficient or if you will need the negative binomial model to get efficient results.
2. We want to test three hypotheses.
  - H<sub>1</sub>** Structure in the way of gdp and population density are significant predictors of conflict
  - H<sub>2</sub>** Defense spending should decrease conflict because citizens then know that the government is well-armed.
  - H<sub>3</sub>** People really engage in conflict because there is nothing better to do so giving them phones and televisions will reduce the number of conflict events.

Run a Poisson model with `conflictevent` as the dependent variable and `defexpgdp`, `phonespc`, `tvbpc`, `gdppcfc` as the independent variables and test these three hypotheses above. Do you believe the Poisson model? Without running a negative binomial, how could you test to see whether your dependent variable is still overdispersed? Interpret your results in terms of percent changes holding all other variables constant.

3. Run a negative binomial regression similar to the one above. How do the results change and why do they change this way? What does the test of overdispersion in the NBR tell you?
4. Using the `prcounts` routine in `stata`, generate a graph that shows, on average, how close each model gets to the observed probability of the counts for the values 0-10.