

INTERMEDIATE SOCIAL STATISTICS CLASSES

WEEK 8: MODELS FOR COUNT DATA

Mr. Dave Armstrong(david.armstrong@politics.ox.ac.uk)

Dr. Michelle Jackson (michelle.jackson@politics.ox.ac.uk)

Objectives: Running and interpreting models for count data, including poisson regression models and negative binomial models.

STATA Commands: histogram, summarize, poisson, listcoef, prtab, prvalue, poisgof, nbreg, prcounts, graph.

Dataset: Scott Long's data set on the careers of 308 Ph.D. biochemists:
(`publications.dta`)

http://www.politics.ox.ac.uk/teaching/res_meths/reading_lists/intermstats/materials/datasets_0506.asp

This week, we're going to discuss models for counts and perhaps models for duration. Count data pose two particular problems:

- 1) They are non-negative integers, so negative predictions are meaningless
- 2) The errors are likely to be heteroskedastic.

We have a couple of different options for dealing with these types of data. The Poisson model and a generalization of the Poisson called the Negative Binomial Regression model.

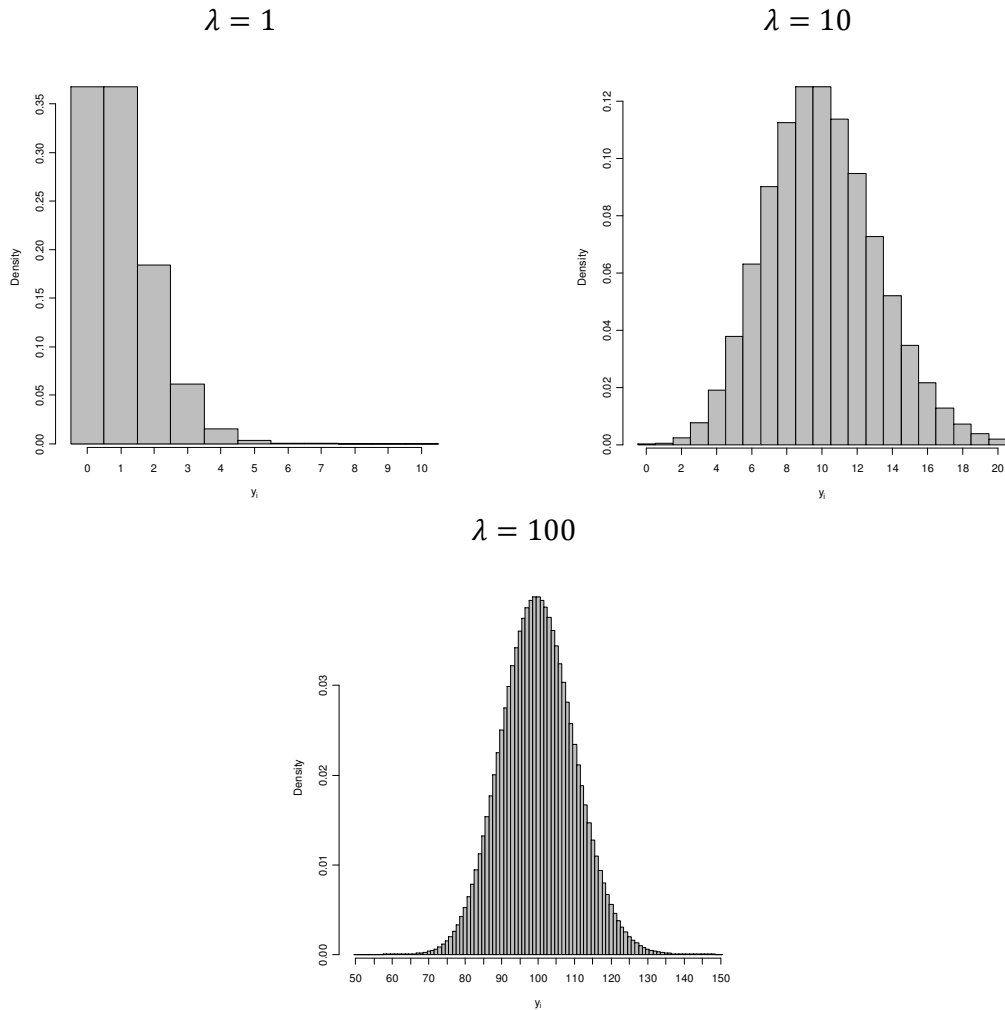
1. POISSON REGRESSION

Models with event count data as the dependent variable are not uncommon in the social sciences and include studies of varied topics such as number of legislative vetoes, party switching, police arrests and capital punishment. Count variables indicate how many times something happened. They are often treated as though they are continuous and a linear regression model is applied. However, the use of linear models for count outcomes can result in inefficient, inconsistent and biased estimates. Hence, in this class we will examine the two most common alternative models for count data: the Poisson regression model (PRM) and the negative binomial regression model (NBRM).

The Poisson regression is the easiest and most common model for count data. With this model, the probability of a count is determined by a Poisson distribution, where the mean of the distribution is a function of the independent variable(s). Let λ be the rate of occurrence or the expected number of times an event will occur over a given period. Let y_i be a random variable indicating the number of times an event did occur. The relationship between the expected count λ and the probability of observing any observed count is specified by the Poisson distribution:

$$\Pr(y_i|\lambda) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \text{ for } y_i = 0,1,2, \dots$$

where λ is the sole parameter defining the distribution, subject to the condition that $\lambda > 0$. The model has the defining characteristic that the conditional mean of the outcome is equal to the conditional variance. This means that Poisson distributions that have large means also have large variances and those with small means have small variances. Let's take a look at a few Poisson Distributions:



This is a relatively strong assumption to make about the underlying random process that you are modeling. In practice, data are often *overdispersed* meaning the conditional variance exceeds the conditional mean or *underdispersed* meaning the conditional variance is smaller than the conditional mean. We're not going to talk about the case where variables are underdispersed as this is not a particularly common problem in the social sciences. We will talk about the problem of overdispersion, however. The negative binomial model essentially adds another parameter to the estimation: α , the dispersion parameter. Though there are a number of ways to motivate this model, Long(1997) suggests thinking about it in terms of an added random component (similar to the error in a linear model). So, we could think about what this does to the conditional mean of the Poisson distribution. Remember, in Poisson, the conditional mean is $\lambda_i = \exp(x_i\beta)$, in the negative binomial, the conditional mean is given by

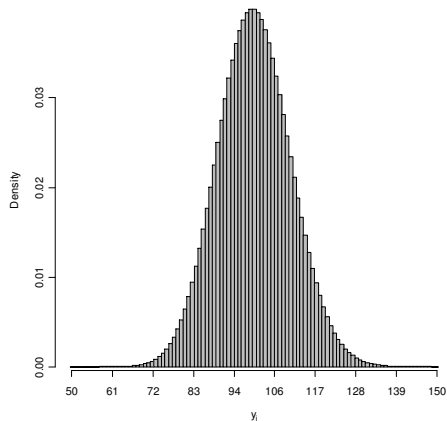
$$\begin{aligned}
\widetilde{\lambda}_i &= \exp(x_i\beta + \varepsilon_i) \\
&= \exp(x_i\beta) \exp(\varepsilon_i) \\
&= \lambda_i \exp(\varepsilon_i) \\
&= \lambda_i \delta_i
\end{aligned}$$

If we assume $E(\varepsilon_i) = 0$, then $E(\delta_i) = 1$. This proves that given this assumption, the mean structure of the negative binomial is the same as the mean structure of the Poisson. You'll notice that above, we suggested that α was the term we added to the model. However, we see δ in the above equation. Since both λ_i and δ_i are unknown, we can't estimate the model. We basically need to average our results over all possible values of δ . To do this, we have to assume a distribution for δ . Other smarter people than us have figured out that assuming the dispersion parameter follows a gamma distribution is effective and tractable. The important properties of the gamma distribution for our purposes are 1) that its values are always positive and 2) its expected value is one. So, this distribution conforms to the assumptions we mentioned above. The parameter we actually estimate is ν_i , which is sometimes referred to as $\nu_i = \alpha_i^{-1}$. This is a parameter of the gamma distribution and we know that $var(\delta_i) = \frac{1}{\nu_i} = \alpha_i$. Putting all of this together, we can calculate the probability of a particular count as:

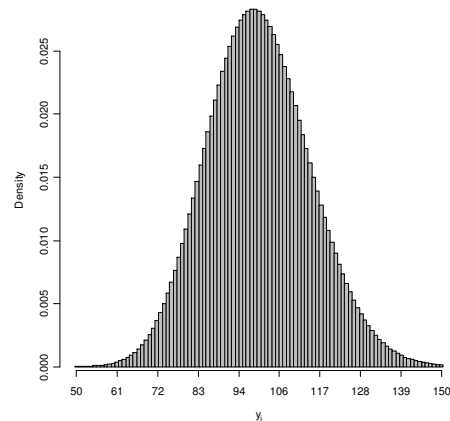
$$\Pr(y|\mathbf{x}) = \frac{\Gamma(y + \alpha^{-1})}{y! \Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \lambda} \right)^{\alpha^{-1}} \left(\frac{\lambda}{\alpha^{-1} + \lambda} \right)^y,$$

where $\lambda = \exp(x\beta)$. We can substitute estimated quantities in for the unknowns to calculate $\Pr(y|\mathbf{x})$ for our sample. The estimation of α allows the Poisson variance to be bigger, changing it from λ to $\lambda + \alpha\lambda^2$. Notice we're adding to the Poisson variance a squared term, which must be positive multiplied by a dispersion parameter which must also be positive. Thus, the negative binomial will allow for variance greater than λ , but not less. What does this mean in terms of variance for the Poisson:

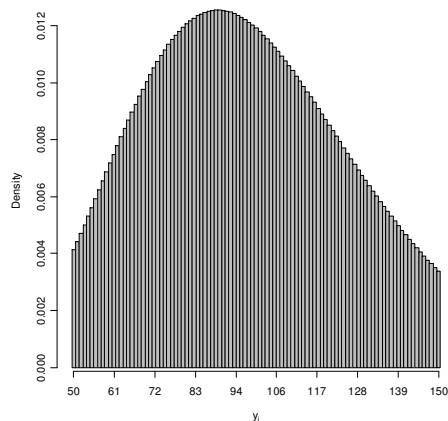
$$\mu = 1, \alpha = 0$$



$$\mu = 10, \alpha = 0.01$$



$$\mu = 100, \alpha = 0.1$$



You can see how these three distributions with the same means have very different variances. Since the PRM is a special case nested in the NBRM, we can do a likelihood ratio test to see whether one model is better than the other. We will show below how you can do this.

Using Scott Long's dataset `publications.dta`, we are going to examine what determines the number of publications of academics. We estimate a simple model where the dependent variable is the number of publications of individual researchers three years after a completed PhD degree `pub3`. We hypothesise that this count is positively associated with number of publications one year prior to the PhD degree `pub1`, the number of citations of the researcher's mentor, `mcit3`, and whether or not the researcher is a member of a university faculty, `faculty`.

First, we examine our dependent variable to inspect whether we can use a linear regression model:

```
.histogram pub3
```

We can see that the data are strongly skewed to the right with a large number of 0s. Consequently, OLS would be inappropriate. We can also use the `summarize` command to get more information about the distribution (using the `detail` option)

```
.summarize pub3, detail
```

We can see that the variance of `pub3` is nearly 5 times larger than the mean. Does this mean that we should use a negative binomial model? Well, if we wanted to predict our variable with just a constant, then yes. However, if we have covariates, that will take care of the overdispersion somewhat because instead of $\lambda_i = \lambda$ (where every observation has the same mean, for example if we only had a constant in the model), when we have covariates we get $\lambda_i = \exp(X_i\beta)$, so each observation (potentially) has its own mean. Cameron and Triverdi (1998) suggest that if the variance is more than twice the mean, covariates are unlikely to model out the overdispersion; this means we would need to explicitly model the overdispersion with a negative binomial model. The distribution of `pub3` is thus displaying signs of overdispersion. But before we look at the alternatives to Poisson distributions, we will run a Poisson regression model, using the `poisson` command:

```
.poisson pub3 faculty mcit3 pub1, nolog
```

Remember, this is a non-linear model, so the coefficients are not interpretable directly, since it is not λ_i that is related linearly to the independent variables, rather it is $\log(\lambda_i)$ that is linearly related to the X's. A common method of interpreting these coefficients is the *factor change* in the rate: for a unit change in the independent variable x_k , the expected count changes by a factor of $\exp(\beta_k)$, holding all other variables constant. Remember from lecture:

$$\frac{E(y|\mathbf{x}, x_k + \delta)}{E(y|\mathbf{x}, x_k)} = e^{\beta_k \delta}$$

The command `listcoef` will generate this for the $\delta = 1$

```
.listcoef
```

Sometimes it is easier to use the percent change – that is, by what percentage will counts change if a variable changes one unit. We can calculate this as follows:

$$100 \times \frac{E(y|\mathbf{x}, x_k + \delta) - E(y|\mathbf{x}, x_k)}{E(y|\mathbf{x}, x_k)} = 100 \times [\exp(\beta_k \times \delta) - 1]$$

```
.listcoef, percent help
```

This output shows that being a faculty member increases the expected number of articles by 36 percent, holding all other variables constant. For every additional article published in the year before the PhD degree, a researcher's predicted mean productivity increases by 14 percent, holding all other variables constant. On the other hand, the mentor's citations make little difference.

The `prtab` command constructs a table of predicted values (events) for all combinations of categorical variables listed. For example we may want to look at the predicted number of publications given for all values of `pub1` and `faculty`:

```
.prtab pub1 faculty, rest(median)
```

The `prvalue` lists predicted values for a given set of values for the independent variables. For example, we can estimate the predicted counts for a faculty researcher with 1 publication prior to his/her PhD:

```
. prvalue, x(faculty=1 pub1=1) rest(median)
```

Looking down the first column of numbers, the first number is identified as the “Rate”, which is the predicted number of publications. The rest of the numbers in that column refer to the probability of seeing various counts given the values provided for the covariates. The highest probability number of publications is 2 with 0.24. Notice, however, from the confidence intervals that it is unlikely that we could statistically differentiate between any of the first three values.

If we want to evaluate the model fit of a Poisson model, *Stata* reports a Pseudo R^2 and “LR chi2,” the model chi-squared statistic. Both of these compare the log-likelihood of the estimated model to the log-likelihood of the null model. The pseudo R^2 measure is calculated as: $R^2 = 1 - \frac{\ln L_m}{\ln L_0}$, where $\ln L_m$ is the log-likelihood of the estimated model and $\ln L_0$ is the log-likelihood of the null model. The LR χ^2 statistic is calculated as $-2(\ln L_0 - \ln L_m)$. As with the other categorical models we looked at, this is akin to the F-test that gets reported with a linear regression – it tests:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

To obtain the deviance statistic, or likelihood ratio chi-squared statistic comparing the model at hand with the saturated model, you have to use the separate command `poisgof`.

```
.poisgof
```

This is calculated with the following formula:

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - \exp(x_i\beta))^2}{\exp(x_i\beta)}$$

In words, this is calculating the squared distance of the observed value from the predicted value and dividing by the predicted value. The degrees of freedom for this test are $n-k-1$. You’ll notice here that big predicted values can be relatively further away from their observed values than smaller predictions and still contribute the same amount to the χ^2 statistic. A p-value close to zero suggests overdispersion. Hence, as already suspected, the lack of model fit suggests that the data are overdispersed. Estimates of a Poisson regression model for overdispersed data are unbiased, but overconfident with standard errors biased downward (Cameron and Trivedi 1998; Long 1997).

2. NEGATIVE BINOMIAL REGRESSION

Negative binomial regression models are used to estimate models of the number of occurrences (counts) of an event when the event has *extra-Poisson variation*; that is, overdispersion. The NBRM addresses the problem of overdispersion in the PRM model by adding a parameter α that reflects unobserved heterogeneity among observations. The two models have the same mean structure. That is, if the assumptions of the NBRM are correct, the expected rate for a given level of independent variables will be (approximately) the same in both models. However, the standard errors in the Poisson regression will be biased downwards, resulting in spuriously small p-values.

Let us therefore fit the same model, but this time we estimate a NBRM, using the `nbreg` command:

```
.nbreg pub3 faculty mcit3 pub1, nolog
```

The likelihood ratio test at the bottom of the analysis is a test of the overdispersion parameter alpha. This examines the null hypothesis of $\alpha = 0$. When the overdispersion parameter is zero, the negative binomial distribution is equivalent to a Poisson distribution. Stata's α is the estimate of the dispersion parameter; it is estimated to be .34 and is significant (non zero). Twice the difference in log-likelihoods between this model and the (nested) Poisson model gives a likelihood ratio statistic of 91.52. We thus have overwhelming evidence of overdispersion. You might notice that the statistic is called χ^2 instead of just χ^2 - this is to take account of the fact that α must be non-negative, so we shouldn't congratulate ourselves for finding values above zero. Thus, this gives an appropriate test of the parameter of interest.

As the mean structure of the NBRM is identical to that for the PRM, the same methods of interpretation can be used. We can obtain the factor and percent change coefficients by using the `listcoef` command:

```
.listcoef, percent help
```

As you can see the results are similar, but not identical, to the results of the PRM. The NBRM output shows that being member of a faculty increases the expected number of articles by 26 percent, holding all other variables constant (compared with 36 percent in PRM). For every additional article published in the year before the PhD degree, a researcher's predicted mean productivity increases by 18 percent, holding all other variables constant (compared with 14 in PRM).

We can use the `prcounts` command to compare the results from the PRM and the NBRM. First, we need to re-estimate the PRM of pub3 on faculty, mcit3, and pub1 using `poisson`.

```
.poisson pub3 faculty mcit3 pub1, nolog
```

Then we can compute predictions (rates and probabilities) with `prcounts`.

```
.prcounts prm, plot max(9)
```

Now we re-estimate the NBRM of pub3 on faculty, mcit3, and pub1 with `nbreg`

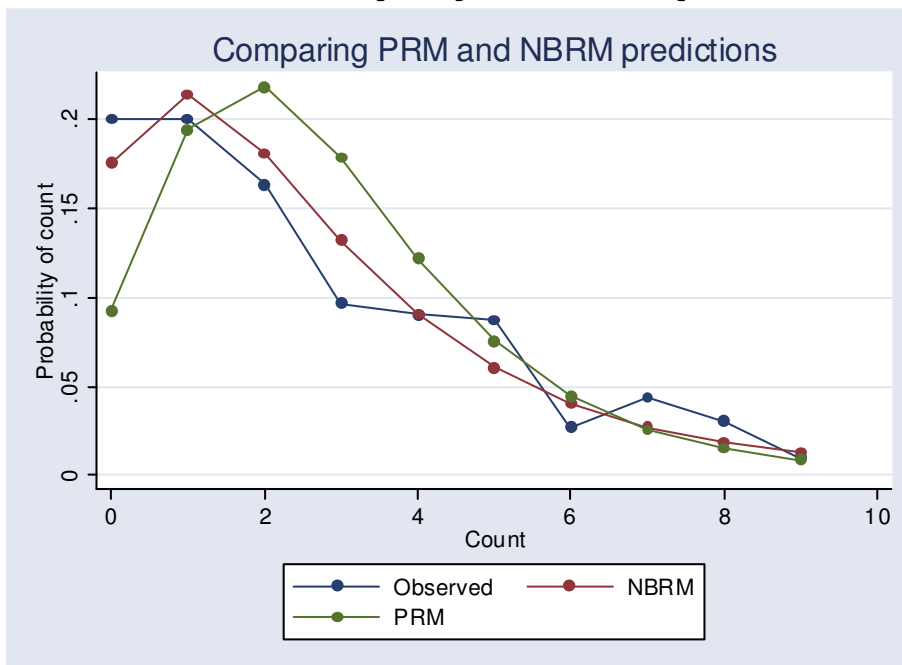
```
.nbreg pub3 faculty mcit3 pub1, nolog
```

And then we computed predictions (rates and probabilities) with `-prcounts-`

```
.prcounts nbr, plot max(9)
```

Plot predicted probabilities with `graph`:

```
.graph twoway connected nbrobeq nbrpreq prmpreq nbrval, connect(direct///  
direct direct) title("Comparing PRM and NBRM predictions")
```



As you can see, the predictions of NBRM fit the observed data better than the PRM.

EXERCISE

Estimate a model with a scientist's number of citations over 3-year period ending the third year of post-PhD period (`cit3`) as the dependent variable. Use gender (`female`), the mentor's number of citations (`mcit3`), number of publications in the year prior to the PhD (`pub1`) and the prestige of the first job (`jobclass`) as explanatory variables.

Which statistical model would you use to estimate this model? Why?

Which of the independent variables has the greatest effect on the number of predicted citations?

What is the probability that a female researcher employed at a good university (`jobclass=2`) with no publications before her PhD will have *no* citations over the 3-year period following her PhD?

3. ZERO-INFLATED AND HURDLE MODELS

There are extensions to the count models called “Zero-Inflated” and hurdle poisson and negative binomial models. These treat zeros different from non-zero counts. The hurdle models are in essence a combination of a binary model that estimates the difference between zero and non-zero and then a Poisson or NB model for the positive counts. The zero inflated models are a bit different. They assume that some of the zeros are zero because there was essentially no theoretical possibility of that observation having a positive count. Other zeros are zero because they just had the right concatenation of independent variables that generate zero, but could have had a positive count. We won’t go into these models in detail, but you can investigate the commands `zip` and `zinb` in Stata if you’re interested.

USEFUL REFERENCES

Cameron, A. C. and P. K. Trivedi (1998) *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.

Long, J. Scott (1997) *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage Publications.

Long, J. S., and J. Freese (2006) *Regression Models for Categorical Dependent Variables Using Stata*. Stata Press.

INTERMEDIATE SOCIAL STATISTICS CLASSES

WEEK 8: MODELS FOR DURATION DATA

Mr. Dave Armstrong(david.armstrong@politics.ox.ac.uk)

Dr. Michelle Jackson(michelle.jackson@politics.ox.ac.uk)

Objectives: Running and interpreting models for duration data – survival analysis.

STATA Commands: `stset`, `streg`.

Dataset: King et al's dataset on cabinet duration: (`king.dta`)

Available at: <http://www.quantoid.net/Oxford.php>

This week, we're going to discuss models for duration – referred to by different fields as survival models, duration models, event history analysis and time-to-failure models. As with other types of data we've considered, there is nothing computationally that prevents us from estimating an OLS on some quantity here. However, as with the other types of data we've considered, OLS will quite possibly give us *wrong answers*. That is, the results could be biased, inconsistent and inefficient. Duration data suffer from a number of similar problems to binary and count data:

- 3) They are non-negative integers, so negative predictions are meaningless.
- 4) The series of points is conditioned on previous points – that is, you can only fail at time t if you have not failed by time $t-1$.
- 5) Some observations do not fail (experience the event) by the time of investigation, so they are *censored*.

We will talk about a broad class of models called duration or survival models and specifically about a couple of operationalizations of these models.

To give intellectual credit where it is due, these are a simplification and subset of Chris Zorn's Spring School class from June 18-20, 2007.

4. REVIEW OF DURATION MODELS: THEORY

There was some confusion yesterday about how we define these different terms. Here is a brief description of the quantities in which we might be interested.

Hazard function (also called hazard rate, or the hazard):

$$h(t) = \frac{f(t)}{S(t)} = \lim_{\Delta t \rightarrow 0} \frac{\Pr(t \leq T_i < t + \Delta t | T_i \geq t)}{\Delta t} = \Pr(T_i = t | T_i \geq t)$$

In words, the hazard is just the probability of an event happening this period given that it hasn't happened already.

Density (Probability Density Function): $f(t) = \Pr(T_i = t)$
 In words, the density is the probability that an event happens in this period.

Cumulative Distribution Function:

$$F(t) = \int_0^t f(t)dt = \Pr(T_i \leq t)$$

In words, the cumulative density is the probability that an event has happened on or before time t .

Survival Function: $S(t) = \Pr(T_i > t) = 1 - F(t)$

In words, the survival function is the probability that an event has not happened yet (i.e., the probability that the observation has “survived” at least this long).

Cumulative Hazard (integrated hazard):

$$H(t) = \int_0^t h(t)dt = -\ln [S(t)]^1$$

In words, the cumulative hazard is basically the hazard averaged over all of the values of t from 0 to t .

Notice that we haven’t said anything about distributions yet. These quantities are defined by their theoretical properties and the definitions above are not attached to any particular distribution. For now, we will be most interested in the hazard function, $h(t)$, and the survival function, $S(t)$. In essence, we want to know – what’s the probability of an event happening at time t given that it has not happened by $t-1$? This quantity is known as the *hazard*:

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

Bigger hazards mean lower probabilities of surviving and vice versa.

It is probably worth talking about censoring now. Censoring is leaving the sample for some reason other than the event of interest. We are assuming that all observations have the event of interest if time were to go to ∞ , however they may not have had the event happen up until the end of the investigation. These observations tell us that observation i has

¹ Since we know that $S(t) = 1 - \int_0^t f(t)dt$, then we know: $f(t) = \frac{-\partial S(t)}{\partial t}$, so we can re-express the hazard rate as:

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{-\partial S(t)}{\partial t}}{S(t)} = \frac{-\partial \ln S(t)}{\partial t}$$

We can then think about the cumulative hazard:

$$H(t) = \int_0^t h(t)dt = \int_0^t \frac{-\partial \ln S(t)}{\partial t} dt = -\ln [S(t)]$$

survived until at least time t . We'll talk about how to deal with censoring when we talk about the models below.

Finally, there are a number of different parameterizations of the models we'll talk about today. The Proportional Hazard (PH) parameterization estimates the hazard, so positive coefficients mean increases in the hazard, and consequently, decreases in survival time. The Accelerated Failure Time (AFT) parameterization estimates the log of survival time as a function of covariates and parameters attached to those covariates. Here, positive coefficients mean longer survival times, and consequently, smaller hazards. Ray used the AFT parameterization in class. We'll give you the details of both, but examples here will also use the AFT model.

5. ESTIMATING SIMPLE DURATION MODELS

To get estimates of the quantities mentioned above we don't necessarily need to assume survival times take any particular distribution. However, for our purposes, if we want to add covariates, we will have to make distributional assumptions, as we do below. Using the `king.dta` we can estimate a simple model for cabinet duration where every country has the same hazard function. `DURAT` is the variable indicating the duration of the cabinet. Estimates of the survival function are generated as follows:

$$\hat{S}(t_k) = \prod_{t \leq t_k} \frac{n_t - d_t}{n_t}$$

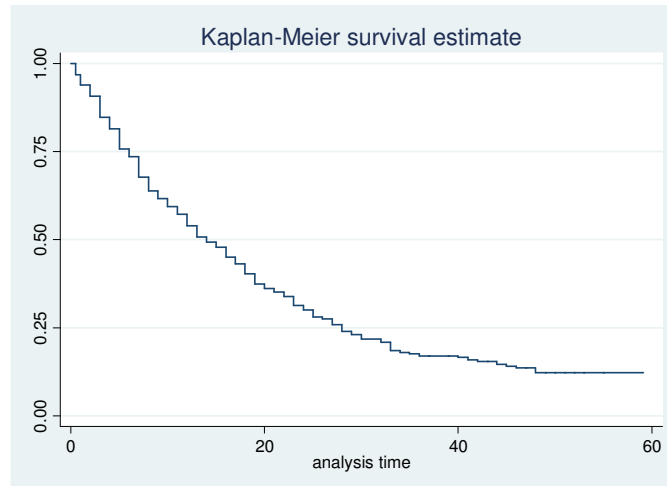
where n_t is the number of events at risk at time t and d_t is the number of observations that failed at time t . We can see this easily in Stata. First, we have to tell Stata that we have duration data and what the duration variable is.

```
. stset duration, failure(CIEP12)
```

Notice here, that the `failure(CIEP12)` option tells Stata that observations with `CIEP12=1` are right-censored. Next, we can ask Stata to show us the estimate of the survival function.

```
. sts graph
```

```
      failure _d:  CIEP12
analysis time _t:  DURAT
```



These estimates come directly from the data and as such are a bit “coarse”. What we will do in the upcoming sections is smooth this curve out by estimating parametric survival models.

6. PARAMETRIC MODELS: EXPONENTIAL

In the parametric model, we will estimate a parameter of some distribution. We are going to talk about two parametric models today. The first is the exponential model – this assumes that survival times follow an exponential distribution. Why the exponential? Well, two reasons, really. First, the exponential distribution takes values $0 \leq x < \infty$. This distribution is also relatively simple. We have to assume that events are independent. The quantities above can be estimated as follows:

<u>PH</u>	<u>AFT</u>
$\lambda = \exp(xb)$	$\lambda = \exp(-xb)$
$h(t) = \lambda$	$h(t) = \lambda$
$S(t) = \exp(-\lambda t)$	$S(t) = \exp(-\lambda t)$
$E(Y_i) = \frac{1}{\lambda}$	$E(Y_i) = \frac{1}{\lambda}$

Let’s ask Stata to generate one of these models for cabinet duration where every observation has the same hazard rate – that is, there is nothing systematic about the way in which different observations fail (we’ll show this probably isn’t realistic later, but for demonstration purposes we’ll look at this).

```

. streg, time dist(exp)

      failure _d:  CIEP12
      analysis time _t:  DURAT

Iteration 0:  log likelihood = -496.39175
Iteration 1:  log likelihood = -496.39175

Exponential regression -- accelerated failure-time form

No. of subjects =          313          Number of obs =          313
No. of failures =          270
Time at risk   =          5789

Log likelihood = -496.39175          LR chi2(0) =          -0.00
                                          Prob > chi2 =          .

-----+-----
      _t |      Coef.   Std. Err.      z    P>|z|    [95% Conf. Interval]
-----+-----
      _cons |  3.065293   .0608581   50.37   0.000   2.946013   3.184572
-----+-----

```

The `time` option above gives the AFT model parameterization. Replacing `time` with `nohr` will give the PH parameterization of the model. The only difference in these two with the exponential distribution is that the coefficients all switch sign. There are more complicated differences in the Weibull model we'll talk about below.

Remember that the hazard for this model is just λ .

$$h(t) = \lambda = \exp(-x\beta)$$

From the output above, we know that $\lambda = \exp(-3.065293) = 0.0466$. We can use this to calculate $S(t)$, the survival function. The survival function requires that we have the time variable. What we want to know is – what does the survival function look like over increasing values of time. To do this, we need to generate a variable called `time` that goes from 0 to the highest observed time in the dataset (59). We can do this in Stata as follows:

```

gen time = _n-1
quietly sum time
replace time= time*59/312

```

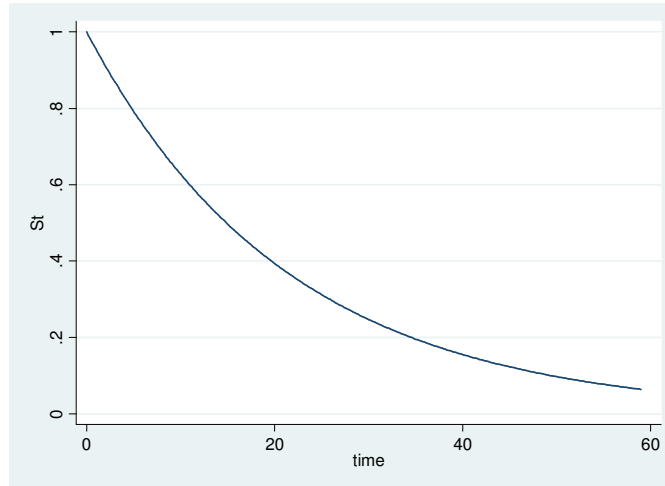
Stata stores the row number in a local macro called `_n`, which starts at 1 and ends with the total number of rows. We can use this to generate a new variable, but since we want it to start at 0 and not 1, we have to subtract one from all of the values. The second command just summarizes this new variable so we can see what the maximum of `time` is (312). Since we want it to go to 59 instead of 312, we need to multiply our current time variable by $\frac{59}{312}$, which is what the third line does.

```

. gen St = exp(-0.0466*time)

. twoway line St time

```



Notice the difference – this curve is smooth because we’ve used a parametric model. The general pattern seems to be about the same. We can then calculate the mean time to failure (dissolution) as $E(Y_i) = \frac{1}{0.0466} = 21.46$. We can then think about adding covariates (a systematic component). First, let’s add dummy variables representing different countries:

```
. streg BELGIUM-SWEDEN, time dist(exp) nolog

      failure _d:  CIEP12
      analysis time _t:  DURAT

Exponential regression -- accelerated failure-time form

No. of subjects =          313          Number of obs =          313
No. of failures =          270
Time at risk   =          5789
Log likelihood  = -443.41678          LR chi2(14)   =          105.95
                                          Prob > chi2   =           0.0000
```

	_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
BELGIUM		-.7369617	.3419402	-2.16	0.031	-1.407152	-.0667713
CANADA		.0990567	.4096732	0.24	0.809	-.7038881	.9020015
DENMARK		-.5361032	.3498251	-1.53	0.125	-1.221748	.1495415
FINLAND		-.9795021	.3320488	-2.95	0.003	-1.630306	-.3286984
FRANCE		-2.100966	.3337751	-6.29	0.000	-2.755153	-1.446778
ICELAND		.1986222	.4206222	0.47	0.637	-.6257822	1.023027
IRELAND		-.154981	.3851644	-0.40	0.687	-.9098895	.5999274
ISRAEL		-.853049	.3498251	-2.44	0.015	-1.538694	-.1674043
ITALY		-1.346199	.3182975	-4.23	0.000	-1.970051	-.7223478
NETHER		-.0232675	.4096732	-0.06	0.955	-.8262123	.7796773
NORWAY		-.2420906	.3733394	-0.65	0.517	-.9739294	.4897482
PORTUG		-1.283407	.4206222	-3.05	0.002	-2.107811	-.4590026
SPAIN		.1191102	.7595545	0.16	0.875	-1.369589	1.60781
SWEDEN		-.1085592	.3851644	-0.28	0.778	-.8634676	.6463492
_cons		3.653651	.2773501	13.17	0.000	3.110055	4.197247

Now, we can undergo an exercise similar to the one above, where we could calculate $\lambda_i = \exp(-x_i b)$.

$$\lambda_{UK} = \exp(-3.3653651) = 0.026$$

$$\lambda_{BEL} = \exp(-(3.3653651 - 0.7369617)) = 0.054$$

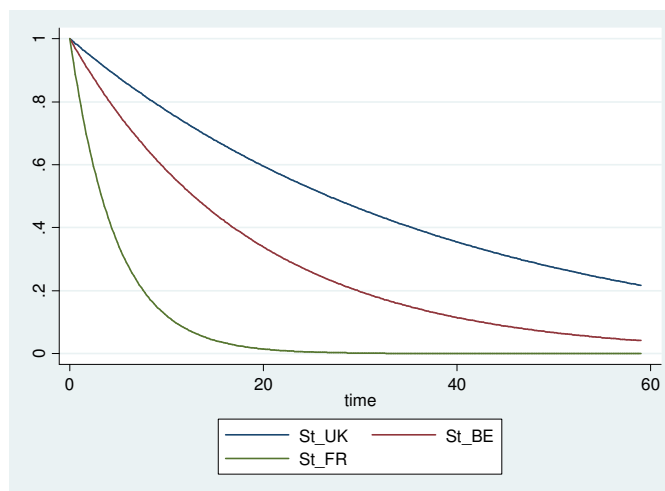
We could also calculate the mean time to failure for each of these places by taking the reciprocal of λ . For the UK it is 38.62 and for Belgium it is 18.48. This means that on average cabinets last about 20 months longer in the UK than in Belgium.

Since the dependent variable is the log of time, this lends itself to a relatively simple interpretation of the coefficients. For every one-unit increase in X_k , there is a corresponding change of $100 \times [1 - \exp(\hat{\beta}_k)]$ percent in the expected survival time. Given that the only coefficients in the model now are dummy variables indicating country, we can say that being in Belgium produces an expected change of $100 \times (1 - 0.479) = 52\%$ in the survival time.

We can also have Stata express the coefficients in *hazard ratios* (this is different from hazard rate), which is the default for the `streg` command. These are the exponentiated coefficients from the model using the `nohr` option above. This is interpreted the same way as an odds ratio. For instance, the hazard ratio for Belgium is 2.08 (which you might be able to recognize as $\exp(-\hat{\beta}_k)$), which means the hazard for Belgium will be 2.08 times as big as the hazard rate for the reference category (the UK). This tells us that cabinets will dissolve faster in the Belgium (a higher hazard) than they will in the UK.

If you're interested in a graphical presentation, we could graph any of the functions mentioned above (survival function, hazard or cumulative hazard) for some or all of the countries above. In general, we do this like predicted probabilities, where we set the levels of the covariates at interesting values and examine either the mean time or the survival function.

```
. gen St_UK = exp(-exp(-3.653651)*time)
. gen St_BE = exp(-exp(-(3.653651-0.7369617))*time)
. gen St_FR = exp(-exp(-(3.653651-2.100966))*time)
. twoway line St_UK St_BE St_FR time
```



Let's add the variable representing number of formation attempts (FORMAT) to the model to see what happens.

```
. streg BELGIUM-SWEDEN FORMAT, dist(exp) time nolog
Exponential regression -- accelerated failure-time form
No. of subjects =          313          Number of obs =          313
No. of failures =          270
Time at risk    =          5789
Log likelihood  = -442.91739          LR chi2(15) =          106.95
                                          Prob > chi2 =           0.0000
```

_____	_____	_____	_____	_____	_____	_____
_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
BELGIUM	-.666748	.3495743	-1.91	0.056	-1.351901	.0184051
CANADA	.0983726	.4096738	0.24	0.810	-.7045734	.9013185
DENMARK	-.4884176	.353478	-1.38	0.167	-1.181222	.2043865
FINLAND	-.9328556	.335331	-2.78	0.005	-1.590092	-.275619
FRANCE	-2.004348	.3485557	-5.75	0.000	-2.687505	-1.321191
ICELAND	.2401057	.4228238	0.57	0.570	-.5886137	1.068825
IRELAND	-.1421875	.3853769	-0.37	0.712	-.8975123	.6131373
ISRAEL	-.8294879	.3506653	-2.37	0.018	-1.516779	-.1421966
ITALY	-1.303726	.3210861	-4.06	0.000	-1.933044	-.6744091
NETHER	.0974275	.4278615	0.23	0.820	-.7411657	.9360207
NORWAY	-.2364608	.3734377	-0.63	0.527	-.9683852	.4954635
PORTUG	-1.275734	.4206951	-3.03	0.002	-2.100282	-.4511872
SPAIN	.118426	.7595548	0.16	0.876	-1.370274	1.607126
SWEDEN	-.1053112	.3851791	-0.27	0.785	-.8602483	.6496259
FORMAT	-.0479149	.04712	-1.02	0.309	-.1402685	.0444387
_cons	3.70225	.2814405	13.15	0.000	3.150637	4.253863

The coefficient on FORMAT is -0.048, which we can interpret as – for every one extra formation attempt, survival time will decrease by $100 \times [1 - \exp(-0.048)] = 4.7\%$. We could also consider how the survival time changes in any given country for a move across the different values of FORMAT (1-8). Let's look at Sweden. We can calculate λ as above:

$$\lambda_{F=1} = \exp(-3.702 + 0.105 + 0.048 \times 1) = 0.028$$

$$\lambda_{F=8} = \exp(-3.702 + 0.105 + 0.048 \times 8) = 0.040$$

We can use these to calculate the mean survival time $\frac{1}{\lambda}$ for FORMAT=1, it is 34.78 and for FORMAT=8, it is 24.87, so for Sweden, formation attempts have an effect of decreasing survival time by 10 months. Before we make a lot of this prediction, let's consider whether this is a good idea. Does it make sense for us to be making predictions about 8 formation attempts in Sweden? How can we tell whether it makes sense? Look at the data! Ask Stata to do the following:

```
. table country, contents(min FORMAT max FORMAT)
```

country	min (FORMAT)	max (FORMAT)
belgium	1	8
canada	1	1
denmark	1	8
finland	1	5
france	1	8
iceland	1	5
ireland	1	2
israel	1	5
italy	1	4
netherlands	1	8
norway	1	3
portugal	1	3
spain	1	1
sweden	1	3
uk	1	2

In Sweden, the maximum number of formation attempts is three, so it probably doesn't make sense to make inferences about eight formation attempts when that is well outside of the observed data. There are, then, two other things we could do – change the prediction for Sweden to account for the observed range of the data or use a different country, say France, for instance.

$$\lambda_{F=1} = \exp(-3.702 + 2.004 + 0.048 \times 1) = 0.192$$

$$\lambda_{F=8} = \exp(-3.702 + 2.004 + 0.048 \times 8) = 0.269$$

We can use these to calculate the mean survival time $\frac{1}{\lambda}$ for $FORMAT=1$, it is 5.20 and for $FORMAT=8$, it is 3.71. Thus, in France, because its baseline hazard is relatively high, there isn't much else that formation attempts can do to make cabinets last less long in the absolute sense. However, look at the ratio of the two different predictions for Sweden and France:

$$\frac{E(Y_i|SW, FORMAT = 1)}{E(Y_i|SW, FORMAT = 8)} = \frac{34.78}{24.87} = 1.4$$

$$\frac{E(Y_i|FR, FORMAT = 1)}{E(Y_i|FR, FORMAT = 8)} = \frac{5.20}{3.714} = 1.4$$

The ratio of the two predictions is the same even though they differ in absolute value. Just like in Logit or Probit, this effect will be different for different countries because of the non-linear nature of the model.

7. PARAMETRIC MODELS: WEIBULL

You'll notice that in the exponential model, the hazard, $h(t) = \lambda$, is constant over time. This is an assumption of the model that we can relax and test. The Weibull model allows hazards to either increase or decrease over time. Using the Weibull distribution adds another parameter p to the model. The quantities above can be estimated as follows:

<u>PH</u>	<u>AFT</u>
$\lambda = \exp(xb)$	$\lambda = \exp(-pxb)$
$h(t) = p\lambda t^{p-1}$	$h(t) = p\lambda t^{p-1}$
$S(t) = \exp(-\lambda t^p)$	$S(t) = \exp(-\lambda t^p)$
$E(Y_i) = \exp\left(\frac{-1}{p}xb\right)\Gamma\left(1 + \frac{1}{p}\right)$	$E(Y_i) = \exp(xb)\Gamma\left(1 + \frac{1}{p}\right)$

Notice, that when $p=1$, $h(t) = 1 \times \lambda \times t^0 = \lambda$, so the exponential model is nested in the Weibull model and we can test the exponential assumption by testing $H_0: p = 1$. If $p > 1$, then hazards are rising over time and if $p < 1$, hazards are decreasing over time. We can use the same sorts of interpretation that we used above with the exponential model. The only difference is that we have a different mean. We can ask Stata to estimate a weibull model by changing the distribution in the `streg` command.

```
. streg BELGIUM-SWEDEN, dist(weibull) time nolog

      failure _d:  CIEP12
analysis time _t:  DURAT

Weibull regression -- accelerated failure-time form

No. of subjects =          313                Number of obs   =          313
No. of failures =           270                LR chi2(14)       =          116.57
Time at risk    =          5789                Prob > chi2       =           0.0000

Log likelihood  = -438.03998

-----+-----
      _t |          Coef.   Std. Err.      z    P>|z|    [95% Conf. Interval]
-----+-----
BELGIUM | - .6808617   .2888412   -2.36  0.018   -1.24698   -.1147433
CANADA  |  .0943015   .3454958    0.27  0.785   -.5828578   .7714608
DENMARK | -.516148    .2950682   -1.75  0.080   -1.094471   .0621751
FINLAND | -.9233396   .2804854   -3.29  0.001   -1.473081   -.3735984
FRANCE  | -2.028053   .2822396   -7.19  0.000   -2.581233   -1.474874
ICELAND | .1577223    .3549043    0.44  0.657   -.5378773   .853322
IRELAND | -.1404914   .3248432   -0.43  0.665   -.7771724   .4961896
ISRAEL  | -.8152453   .2952206   -2.76  0.006   -1.393867   -.2366237
ITALY   | -1.281932   .2690655   -4.76  0.000   -1.809291   -.7545732
NETHER  | -.0352752   .3455065   -0.10  0.919   -.7124554   .6419051
NORWAY  | -.2423165   .3148985   -0.77  0.442   -.8595062   .3748731
PORTUG  | -1.222062   .3551474   -3.44  0.001   -1.918138   -.5259855
SPAIN   | .0727956    .6406966    0.11  0.910   -1.182947   1.328538
SWEDEN  | -.1271326   .3248615   -0.39  0.696   -.7638494   .5095842
_cons   |  3.646691    .2339019   15.59  0.000    3.188252    4.10513

-----+-----
      /ln_p | .1703841   .0500171    3.41  0.001    .0723523    .2684158

-----+-----
      p |  1.18576   .0593083    1.075034   1.307891
      1/p | .8433409   .0421815    .7645898   .9302031
-----+-----
```

We get three different estimates of the same quantity. The model estimates $\ln(p)$, but we're usually interested either in p or $1/p$ and Stata provides us with all of these quantities. To see whether the Weibull model is significantly better than the exponential model, we need

to test $H_0: \ln(p) = 0$. Since the p -value is quite small (<0.05), we can reject the null in favor of the alternative that $p \neq 0$, so the Weibull model is significantly better. We can get Stata to tell us the predicted mean all of the observations with the following command:

```
. predict mtw, mean time
```

Since we know we're only getting one prediction per country (since no other cabinet-specific effects are in the model), we can look at the prediction for all of the countries as follows:

```
. table country, contents(mean mtw )
```

```
-----
      country | mean(mtw)
-----+-----
      belgium | 18.31613
        canada | 39.76334
      denmark | 21.59574
      finland | 14.37232
        france | 4.761634
      iceland | 42.36686
      ireland | 31.44225
        israel | 16.01296
         italy | 10.04136
netherlands | 34.93079
        norway | 28.39825
      portugal | 10.66089
         spain | 38.91733
        sweden | 31.8651
          uk   | 36.18497
-----
```

We could also calculate this “by hand” in Stata. Below is the calculation for Belgium.

```
. dis exp(3.646691-.6808617)*exp(lngamma(1+1/1.18576))
18.316136
```

EXERCISE

Add `FORMAT` to the Weibull model estimated above. Test to see whether the Weibull model is statistically better than the exponential model. What is the null hypothesis we're testing?

What is the expected survival time for a cabinet in the Netherlands with the median number of formation attempts in the Netherlands? What is the effect of formation attempts in the Netherlands?

Have Stata represent the model in accelerated failure time format. Interpret the coefficient for formation in terms of the percentage change in survival time for a one-unit increase.

8. CONCLUDING REMARKS

The models we've talked about today are of the "proportional hazards" type – where the effects of covariates are the same over time. There are a set of non-proportional models that we didn't talk about. We also didn't spend any time on time-varying covariates. If you're interested in these topics or learning more about survival models, Indridi Indridason is teaching an option in Trinity Term dealing exclusively with these types of models.