

## Models for Count Outcomes, Part I

These notes borrow heavily (sometimes verbatim) from Long 1997, Regression Models for Categorical and Limited Dependent Variables, and Long & Freese, 2003 Regression Models for Categorical Dependent Variables Using Stata, Revised Edition, and also the 2006 2<sup>nd</sup> edition of Long & Freese. Materials prepared by my former Soc 592/593 teaching assistant Jamie Przybysz are also incorporated in these notes.

Variables that count the # of times something happens are common in the Social Sciences.

- Hausman looked at effect of R & D expenditures on # of patents received by US companies
- Grogger examined deterrent effects of capital punishment on daily homicides
- King examined effect of # of alliances on the # of nations at war
- Long looked at # of publications of scientists

Count variables are often treated as though they are continuous and the linear regression model is applied; but this can result in inefficient, inconsistent and biased estimates. Fortunately, there are many models that deal explicitly with count outcomes.

- The most basic is the *Poisson Regression Model* (PRM). In the PRM the probability of a count is determined by a Poisson distribution, where the mean of the distribution is a function of the IVs. The conditional mean of the outcome is equal to the conditional variance.
- In practice, however, the conditional variance often exceeds the conditional mean. The *Negative Binomial Regression Model* (NBRM) deals with this problem by allowing the variance to exceed the mean.
- A second problem with the PRM is that the # of 0's in a sample often exceeds the # predicted by either the PRM or the NBRM. *Zero Modified Count Models* explicitly model the # of predicted 0s, and also allow the variance to differ from the mean.
- A third problem is that many count variables are only observed after the first count occurs. This requires a *Truncated Count Model*.

The Poisson Distribution.

Let  $y$  be a random variable indicating the # of times an event has occurred during an interval of time.  $y$  has a Poisson distribution with parameter  $\mu > 0$  if

$$\Pr(y | \mu) = \frac{\exp(-\mu)\mu^y}{y!} \quad \text{for } y = 0, 1, 2, \dots$$

# of occurrences	Pr(y=# of occurrences  $\mu$ )
0	$\text{Exp}(-\mu)$
1	$\text{Exp}(-\mu) \mu$
2	$\text{Exp}(-\mu) \mu^2/2$
3	$\text{Exp}(-\mu) \mu^3/6$
4	$\text{Exp}(-\mu) \mu^4/24$

So, for example, with 50 events occurring to 100 units, we find the following:

Prop(0) =  $[(.5^0)*(e^{-.5})/1] = .61$  (61 of the 100 units will experience no events)

Prop(1) =  $[(.5^1)*(e^{-.5})/1] = .30$  (30 of the 100 units will experience 1 event)

Prop(2) =  $[(.5^2)*(e^{-.5})/(2*1)] = .08$  (8 of the 100 units will experience 2 events)

Prop(3) =  $[(.5^3)*(e^{-.5})/(3*2*1)] = .01$  (1 of the 100 units will experience 3 events)

Prop(4) =  $[(.5^4)*(e^{-.5})/(4*3*2*1)] = .002$  (not substantively meaningful here, as it is too small,)

Prop(5) =  $[(.5^5)*(e^{-.5})/(5*4*3*2*1)] = .0002$  (but presented to show the example calculations )

This figure shows what the Poisson distribution looks like for different values of  $\mu$

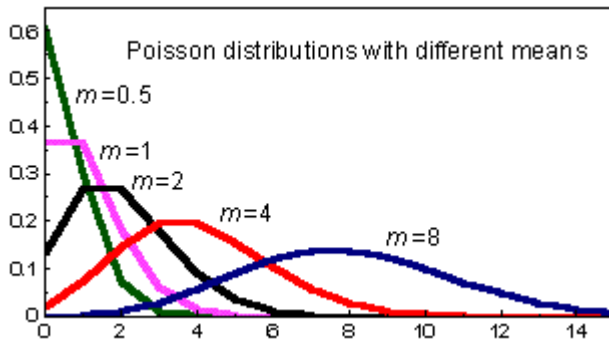


Image copied from <http://www.cmh.edu/stats/model/poiss10.htm>

Key properties of the Poisson distribution:

- As  $\mu$  increases, the mass of the distribution shifts to the right. Specifically,  $E(y) = \mu$ . The parameter  $\mu$  is known as the rate since it is the expected # of times that an event has occurred per unit of time.  $\mu$  can also be thought of as the mean or expected count.
- The variance equals the mean. The equality of the mean and the variance is known as *equidispersion*. In practice, count variables often have a variance that is greater than the mean, which is called *overdispersion*. The development of many models for count data is an attempt to account for overdispersion.
- As  $\mu$  increases, the probability of 0s decreases. For  $\mu = .8$ , the probability of a 0 is .45. For  $\mu = 1.5$ , it is .22, for  $\mu = 2.9$ , it is .05; and for  $\mu = 10.5$ , the probability is .00002. For many count variables, there are more observed 0s than predicted by the Poisson distribution.
- As  $\mu$  increases, the Poisson distribution approximates a normal distribution.

A critical assumption of a Poisson process is that events are independent; this means that when an event occurs it does not affect the probability of an event occurring in the future. For example,

this implies that when a scientist publishes a paper, her rate of publication does not change. Past success in publishing does not affect future success.

As noted, the actual variance is often larger than a Poisson process would suggest. One likely explanation is that  $\mu$  differs across individuals, e.g. not all scientists are equally productive. This is known as heterogeneity. For example, suppose that for men, mean productivity =  $\mu + \delta$ , and for women it is  $\mu - \delta$ . If the number of men and women is equal, the mean productivity will be  $\mu$ , but the variance will exceed  $\mu$ . In general, failure to account for heterogeneity among individuals in the rate of a count variable leads to overdispersion. This leads to the Poisson Regression Model which introduces heterogeneity based on *observed* characteristics.

## Poisson Regression Model

In the PRM, the # of events  $y$  has a Poisson distribution with a conditional mean that depends on an individual's characteristics:

$$\mu_i = E(y_i | x_i) = \exp(x_i\beta)$$

Note the exponentiation forces the expected count to be positive. It can also be written as (and this is more consistent with the way we have written all our other models)

$$\ln(\mu_i) = x_i\beta$$

Under this model, as  $\mu$  increases, the conditional variance of  $y$  increases, the proportion of predicted 0s decreases and the distribution around the expected value becomes approximately normal.

The PRM can be thought of as a non-linear regression model with errors equal to  $\varepsilon = y - E(y|x)$ . The errors have a Poisson distribution. But, we cannot use OLS as the regression technique for data that resemble a Poisson distribution because in the Poisson, the mean ( $\mu$ ) = Variance of  $x$ . As  $\mu$  increases, so does the variance around it. (You'll recall that OLS assumes a constant variance.) The dispersion of data increases as  $\mu$  increases. Since the level of the DV affects dispersion, the errors in a Poisson regression are inherently heteroskedastic. The PRM is, in fact, another case of the Generalized Linear Model that we have been talking about and is estimated via maximum likelihood. The family is Poisson (errors have a Poisson distribution) and the link is log (the log of  $E(Y)$  is the dependent variable).

You can use the parameters to compute the probability distribution for a given level of the IVs. For a given  $x$ , the probability that  $y = m$  is

$$\hat{\Pr}(y = m | x) = \frac{\exp(-\hat{\mu})\hat{\mu}^m m!}{m!} \quad \text{where } \hat{\mu} = \exp(x\hat{\beta})$$

The PRM model should do better than a univariate Poisson distribution. Still, it can under predict 0s and have a variance that is greater than the conditional mean. Hence, other models have been developed which we will discuss shortly.

*Estimating the PRM in Stata.* The `poisson` command is used to estimate Poisson Regression Models. Long and Freese present an analysis of the number of publications produced by Ph.D. biochemists:

```
. use http://www.nd.edu/~rwilliam/xsoc73994/long2006/couart2.dta
(Academic Biochemists / S Long)
. des
```

```
Contains data from http://www.nd.edu/~rwilliam/xsoc694/long2003/couart2.dta
  obs:          915          Academic Biochemists / S Long
  vars:           6          30 Jan 2001 10:49
  size:        11,895 (99.9% of memory free)  (_dta has notes)
```

variable name	storage type	display format	value label	variable label
art	byte	%9.0g		Articles in last 3 yrs of PhD
fem	byte	%9.0g	sexlbl	Gender: 1=female 0=male
mar	byte	%9.0g	marlbl	Married: 1=yes 0=no
kid5	byte	%9.0g		Number of children < 6
phd	float	%9.0g		PhD prestige
ment	byte	%9.0g		Article by mentor in last 3 yrs

```
Sorted by:  art
```

```
. sum, sep(6)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
art	915	1.692896	1.926069	0	19
fem	915	.4601093	.4986788	0	1
mar	915	.6622951	.473186	0	1
kid5	915	.495082	.76488	0	3
phd	915	3.103109	.9842491	.755	4.62
ment	915	8.767213	9.483916	0	77

Note that the mean # of articles published is 1.69. Note too that the variance is  $1.926^2$ , which is substantially more than the mean.

We now estimate a simple model with constant-only. If this model is valid, then every academic biochemist has the same rate of productivity.

```
. poisson art, nolog
```

```
Poisson regression          Number of obs =          915
                           LR chi2(0)      =           0.00
                           Prob > chi2     =            .
Log likelihood = -1742.5735   Pseudo R2      =          0.0000
```

art	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
_cons	.5264408	.0254082	20.72	0.000	.4766416 .57624

Note that the coefficient for the constant is .52664408. Further, note that  $\exp(.52664408) = 1.693$ , the same as the mean given in the earlier descriptive statistics.

Your intuition probably tells you that this model does not make much sense – but how do you test it? You can do so with the `estat gof` post-estimation command (the older `poisgof` command also works)

```
. estat gof
```

```
Goodness-of-fit chi2 = 1817.405
Prob > chi2(914)    = 0.0000
```

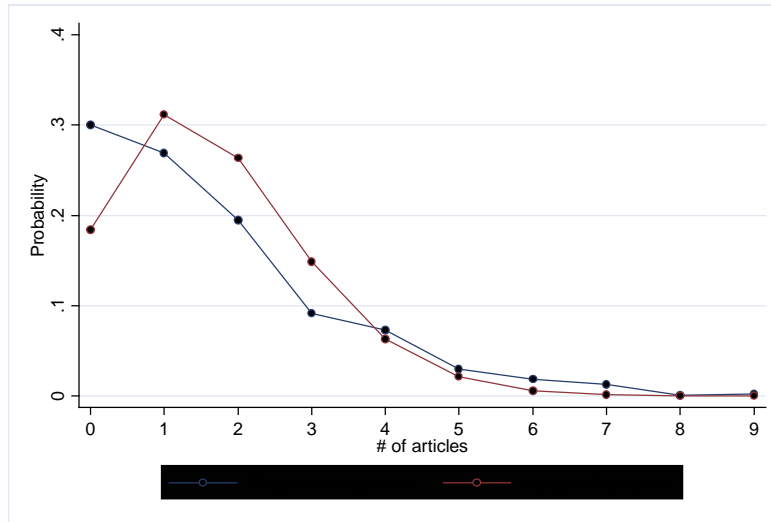
This command compares the observed distribution with the distribution predicted by a Poisson distribution. The highly significant test statistic indicates that this is not a very good model. Long and Freese describe a procedure for comparing the predicted with the observed distribution. Their post-estimation command `prcounts` computes the predicted rate and predicted probabilities of each count from 0 to the specified maximum for every observation:

```
. prcounts psn, plot max(9)
. label var psnobeq "Observed Proportion"
. label var psnpreq "Poisson Prediction"
. label var psnval "# of articles"
. list psnval psnobeq psnpreq in 1/10
```

	psnval	psnobeq	psnpreq
1.	0	.3005464	.1839859
2.	1	.2688525	.311469
3.	2	.1945355	.2636423
4.	3	.0918033	.148773
5.	4	.073224	.0629643
6.	5	.0295082	.0213184
7.	6	.0185792	.006015
8.	7	.0131148	.0014547
9.	8	.0010929	.0003078
10.	9	.0021858	.0000579

As you can see, when the mean is 1.69, a Poisson distribution predicts that 18.39% of the cases will be zeros; but in reality more than 30% are. You also see more people than predicted in the 3+ range. If you want to graph this (and can remember the command!):

```
. graph twoway connected psnobeq psnpreq psnval, ytitle("Probability") ylabel(0(.1).4)
xlabel(0(1)9) ysize(2.7051) xsize(4.0421)
```



Of course, we never believed in that model anyway. Productivity may differ by gender, marital status, number of young children, prestige of the graduate program, and the number of articles written by a scientist's mentor. If so, mixing together scientists who differ in their rate of productivity can cause the univariate distribution of the articles to be overdispersed, i.e. have a variance greater than its mean. To account for these differences we add IVs to our model:

```
. poisson art fem mar kid5 phd ment
```

```
Iteration 0: log likelihood = -1651.4574
Iteration 1: log likelihood = -1651.0567
Iteration 2: log likelihood = -1651.0563
Iteration 3: log likelihood = -1651.0563
```

```
Poisson regression                               Number of obs   =          915
                                                  LR chi2(5)      =          183.03
                                                  Prob > chi2     =           0.0000
Log likelihood = -1651.0563                    Pseudo R2       =           0.0525
```

art	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
fem	-.2245942	.0546138	-4.11	0.000	-.3316352 -.1175532
mar	.1552434	.0613747	2.53	0.011	.0349512 .2755356
kid5	-.1848827	.0401272	-4.61	0.000	-.2635305 -.1062349
phd	.0128226	.0263972	0.49	0.627	-.038915 .0645601
ment	.0255427	.0020061	12.73	0.000	.0216109 .0294746
_cons	.3046168	.1029822	2.96	0.003	.1027755 .5064581

```
. estat gof
```

```
Goodness-of-fit chi2 = 1634.371
Prob > chi2(909)    = 0.0000
```

Alas, the fit still isn't very good. Repeating our earlier procedure (I dropped all the variables created last time so I could give the same commands again):

```

. prcounts psn, plot max(9)

. label var psnobeq "Observed Proportion"
. label var psnpreq "Poisson Prediction"
. label var psnval "# of articles"
. list psnval psnobeq psnpreq in 1/10

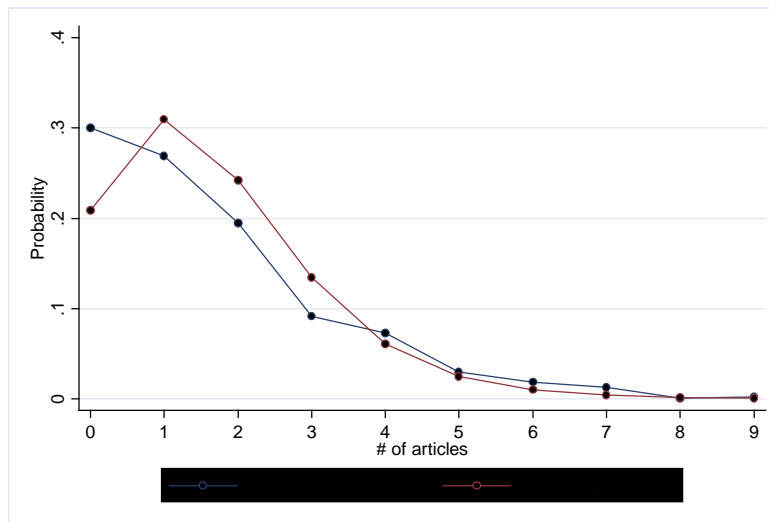
```

	psnval	psnobeq	psnpreq
1.	0	.3005464	.2092071
2.	1	.2688525	.3098447
3.	2	.1945355	.242096
4.	3	.0918033	.1346656
5.	4	.073224	.0611696
6.	5	.0295082	.0249554
7.	6	.0185792	.0099346
8.	7	.0131148	.0041384
9.	8	.0010929	.001877
10.	9	.0021858	.0009304

```

. graph twoway connected psnobeq psnpreq psnval, ytitle("Probability") ylabel(0(.1).4)
xlabel(0(1)9) ysize(2.7051) xsize(4.0421)

```



Again, we see more observed zeroes than predicted zeros. We'll talk about some alternatives to this model, but first we'll talk about how to interpret the parameters we have got.

**Relationship to the Generalized Linear Model.** As noted before, Poisson Regression models are a special case of the Generalized Linear Model. Therefore they can also be estimated with the `glm` command:

```
. glm art fem mar kid5 phd ment, family(poisson) link(log)
```

```
Iteration 0: log likelihood = -1670.3221
Iteration 1: log likelihood = -1651.1048
Iteration 2: log likelihood = -1651.0563
Iteration 3: log likelihood = -1651.0563
```

```
Generalized linear models          No. of obs      =          915
Optimization      : ML: Newton-Raphson  Residual df    =          909
                                                Scale parameter =           1
Deviance          = 1634.370984         (1/df) Deviance = 1.797988
Pearson          = 1662.54655          (1/df) Pearson = 1.828984
```

```
Variance function: V(u) = u          [Poisson]
Link function      : g(u) = ln(u)     [Log]
Standard errors   : OIM
```

```
Log likelihood = -1651.056316        AIC              = 3.621981
BIC            = -4564.030991
```

art	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
fem	-.2245942	.0546138	-4.11	0.000	-.3316352 -.1175532
mar	.1552434	.0613747	2.53	0.011	.0349512 .2755356
kid5	-.1848827	.0401272	-4.61	0.000	-.2635305 -.1062349
phd	.0128226	.0263972	0.49	0.627	-.038915 .0645601
ment	.0255427	.0020061	12.73	0.000	.0216109 .0294746
_cons	.3046168	.1029822	2.96	0.003	.1027755 .5064581

Interpreting the Results of the PRM. In their current form, the beta coefficients tell us how much a 1 unit increase in each X causes the log of  $\mu$  to increase. Since that isn't the most intuitive idea in the world, it will be useful to exponentiate the coefficients. We can do this by adding the `irr` parameter (which, mathematically, does the exact same thing as the odds ratio parameter we have used in the past; but `irr` stands for *incident rate ratio*, with the idea being that the coefficient tells you how changes in X affect the rate at which Y occurs (keeping in mind that the terms rate and mean stand for the same thing here.)

```
. quietly poisson art fem mar kid5 phd ment
. poisson, irr
```

```
Poisson regression          Number of obs   =          915
LR chi2(5)                 =          183.03
Prob > chi2                =          0.0000
Pseudo R2                  =          0.0525
Log likelihood = -1651.0563
```

art	IRR	Std. Err.	z	P> z	[95% Conf. Interval]
fem	.7988403	.0436277	-4.11	0.000	.7177491 .8890932
mar	1.167942	.0716821	2.53	0.011	1.035569 1.317236
kid5	.8312018	.0333538	-4.61	0.000	.7683342 .8992134
phd	1.012905	.0267379	0.49	0.627	.9618325 1.06669
ment	1.025872	.002058	12.73	0.000	1.021846 1.029913

These coefficients tell us that, on an all other things equal basis,

- Females publish 80% as many articles as males, i.e. are 20% less productive
- Married people are about 17% more productive than unmarried people
- Each additional child multiplies the rate of productivity by .83, e.g. somebody with one child will only produce 83% as many articles as somebody with no children.
- The prestige of the PHD institution doesn't have much effect
- For each additional article a mentor publishes, productivity gets multiplied by 1.025872, i.e. there is about a 2.6% increase per article. (But remember, you do compounding, not addition, as you figure the effect of increases in X that are greater than one.

The margins command is also helpful. Note that the default asobserved is being used instead of atmeans.

```
. quietly poisson art i.fem i.mar kid5 phd ment
. margins fem mar
```

```
Predictive margins                                Number of obs =          915
Model VCE      : OIM
```

```
Expression   : Predicted number of events, predict()
```

	Margin	Delta-method Std. Err.	z	P> z	[95% Conf. Interval]	
-----						
fem						
0	1.863249	.062788	29.68	0.000	1.740187	1.986312
1	1.488439	.0614126	24.24	0.000	1.368072	1.608805
mar						
0	1.526787	.0742234	20.57	0.000	1.381312	1.672263
1	1.7832	.0576126	30.95	0.000	1.670281	1.896118

```
. margins, dydx(*)
```

```
Average marginal effects                          Number of obs =          915
Model VCE      : OIM
```

```
Expression   : Predicted number of events, predict()
dy/dx w.r.t. : 1.fem 1.mar kid5 phd ment
```

	dy/dx	Delta-method Std. Err.	z	P> z	[95% Conf. Interval]	
-----						
1.fem	-.3748107	.0900846	-4.16	0.000	-.5513733	-.1982481
1.mar	.256412	.0990332	2.59	0.010	.0623105	.4505135
kid5	-.3129872	.068395	-4.58	0.000	-.447039	-.1789354
phd	.0217073	.0446911	0.49	0.627	-.0658857	.1093003
ment	.0432412	.0035694	12.11	0.000	.0362454	.0502371

Note: dy/dx for factor levels is the discrete change from the base level.

The results tell us that, after controlling for other variables, on average woman publish .375 fewer articles than men; and on average, married people publish .256 more articles.

The `listcoef` command can also be used here:

```
. listcoef, help
```

```
poisson (N=915): Factor Change in Expected Count
```

```
Observed SD: 1.926069
```

art	b	z	P> z	e^b	e^bStdX	SDofX
fem	-0.22459	-4.112	0.000	0.7988	0.8940	0.4987
mar	0.15524	2.529	0.011	1.1679	1.0762	0.4732
kid5	-0.18488	-4.607	0.000	0.8312	0.8681	0.7649
phd	0.01282	0.486	0.627	1.0129	1.0127	0.9842
ment	0.02554	12.733	0.000	1.0259	1.2741	9.4839

```
b = raw coefficient
z = z-score for test of b=0
P>|z| = p-value for z-test
e^b = exp(b) = factor change in expected count for unit increase in X
e^bStdX = exp(b*SD of X) = change in expected count for SD increase in X
SDofX = standard deviation of X
```

The main additional piece of information you are gaining here is the effect on productivity of a 1 standard deviation increase in X. Alternatively, we can get the percent change produced by changes in X with the following:

```
. listcoef, help percent
```

```
poisson (N=915): Percentage Change in Expected Count
```

```
Observed SD: 1.926069
```

art	b	z	P> z	%	%StdX	SDofX
fem	-0.22459	-4.112	0.000	-20.1	-10.6	0.4987
mar	0.15524	2.529	0.011	16.8	7.6	0.4732
kid5	-0.18488	-4.607	0.000	-16.9	-13.2	0.7649
phd	0.01282	0.486	0.627	1.3	1.3	0.9842
ment	0.02554	12.733	0.000	2.6	27.4	9.4839

```
b = raw coefficient
z = z-score for test of b=0
P>|z| = p-value for z-test
% = percent change in expected count for unit increase in X
%StdX = percent change in expected count for SD increase in X
SDofX = standard deviation of X
```

prchange continues to be useful:

```
. prchange
```

```
poisson: Changes in Predicted Rate for art
```

	min->max	0->1	++1/2	++sd/2	MargEfct
fem	-0.3591	-0.3591	-0.3624	-0.1804	-0.3616
mar	0.2440	0.2440	0.2502	0.1183	0.2500
kid5	-0.7512	-0.2978	-0.2981	-0.2279	-0.2977
phd	0.0794	0.0200	0.0206	0.0203	0.0206
ment	7.9124	0.0333	0.0411	0.3910	0.0411

```
exp(xb): 1.6101
```

	fem	mar	kid5	phd	ment
x=	.460109	.662295	.495082	3.10311	8.76721
sd(x)=	.498679	.473186	.76488	.984249	9.48392

Exposure time. So far we have implicitly assumed that each observation was “at risk” of an event occurring for the same amount of time. This need not be true; for example, scientists may have received their Ph.D.s in different years. Amount of time in career will certainly affect the number of publications. Further, if exposure time is correlated with our variables, e.g. men have had the Ph.D.s longer than women have, we may get very misleading results.

Since the data from our example do not include exposure data, we will make some up. The variable profage corresponds to the scientists professional age which corresponds to the amount of time a scientist has been exposed to the risk of publishing. In the following, men have an average professional age of 30, while women have an average professional age of 15:

```
. set seed 123456
. gen profage = (10 + invnorm(uniform())) * 3 if fem == 0
(421 missing values generated)

. set seed 1234567

. replace profage = (5 + invnorm(uniform())) * 3 if fem == 1
(421 real changes made)

. bysort fem: sum profage
```

```
-----
-> fem = Men

  Variable |      Obs      Mean   Std. Dev.   Min      Max
-----+-----
  profage |      494   29.98072   2.888995   21.81031   39.3049
-----+-----

-> fem = Women

  Variable |      Obs      Mean   Std. Dev.   Min      Max
-----+-----
  profage |      421   14.73158    3.16198    6.919592   24.08156
-----
```

As Long and Freese note, there are at least three ways to incorporate exposure time into Poisson models. The simplest may be to use the `exposure` option. (See Long and Freese for the other alternatives if for some reason you prefer them.)

```
. poisson art fem mar kid5 phd ment, nolog exposure(profage) irr
```

Poisson regression

Number of obs	=	915
LR chi2(5)	=	239.61
Prob > chi2	=	0.0000
Pseudo R2	=	0.0670

Log likelihood = -1667.0565

art	IRR	Std. Err.	z	P> z	[95% Conf. Interval]
fem	1.629648	.0888685	8.96	0.000	1.464454 1.813476
mar	1.160387	.071134	2.43	0.015	1.029016 1.308528
kid5	.8360909	.0334668	-4.47	0.000	.7730042 .9043262
phd	1.005985	.026486	0.23	0.821	.9553903 1.05926
ment	1.026454	.002059	13.02	0.000	1.022427 1.030498
profage	(exposure)				

Notice how this dramatically changes our estimate of the effect of gender; once we control for exposure time, women are much more productive than men. In other words, their lower productivity is due to the fact that they haven't had their Ph.Ds as long. Hence, failing to control for exposure time could create a very misleading impression.

### Negative Binomial Regression Model

The PRM accounts for observed heterogeneity (i.e. observed differences among sample members) by specifying the rate  $\mu$  as a function of the observed Xs. In practice, the PRM rarely fits, because of overdispersion. That is, the model underestimates the amount of dispersion in the outcome. If the mean structure from the PRM is correct, but there is overdispersion in the estimates,

- PRM estimates are consistent, but inefficient
- Standard errors will be biased downward resulting in spuriously large z-values

The NBRM adds a parameter that allows the conditional variance of  $y$  to exceed the conditional mean. In the NBRM, the mean  $\mu$  is replaced with the random variable  $\tilde{\mu}$ :

$$\tilde{\mu}_i = \exp(x_i\beta + \varepsilon_i)$$

where  $\varepsilon$  is a random error that is assumed to be uncorrelated with  $x$ . You can think of  $\varepsilon$  as either the combined effects of unobserved variables that have been omitted from the model or as another source of pure randomness.

Put another way, in the PRM, variation in  $\mu$  is introduced through *observed heterogeneity*. In the NBRM, you also have variation due to *unobserved heterogeneity*. For a given combination of xs there is a distribution of  $\mu$ s rather than a single  $\mu$ . The conditional mean is still  $\mu$ , but the variance will be greater because of the error term.

The relationship between mu-squiggle and mu is

$$\tilde{\mu}_i = \exp(x_i\beta) \exp(\varepsilon_i) = \mu_i \exp(\varepsilon_i) = \mu\delta_i$$

The NBRM is not identified without an assumption about the mean of the error term, and the most convenient assumption is that the mean is 1. (This is analogous to assuming in OLS regression that the mean of the residuals is 0). Hence,

$$\tilde{\mu}_i = \exp(x_i\beta) \exp(\varepsilon_i) = \mu_i \exp(\varepsilon_i) = \mu\delta_i = \mu_i$$

What is the distribution of delta? The most common assumption is that delta has a gamma distribution with parameter v. If delta has a gamma distribution, then E(delta) = 1 and Var(delta) = 1/v.

The expected value of y for the Negative Binomial distribution is the same as for the Poisson distribution, but the conditional variance differs:

$$\text{Var}(y_i | x) = \mu_i \left( 1 + \frac{\mu_i}{v_i} \right) = \exp(x_i\beta) \left( 1 + \frac{\exp(x_i\beta)}{v_i} \right)$$

Since mu and v are positive, the conditional variance of y in the NBRM must exceed the conditional mean exp(xB).

The larger conditional variance in y increases the relative frequency of low and high counts. The NB distribution corrects a number of sources of poor fit that are often found when the Poisson distribution is used:

- The variance of the NB distribution exceeds the variance of the Poisson distribution for a given mean
- The increased variance in the NBRM results in substantially larger probabilities for small counts.
- There are slightly larger probabilities for larger counts in the NB distribution.

If v varies by individuals, then there are more parameters than there are observations. The most common identifying assumption is that v is the same for all individuals (again note the similarities with OLS):

$$v_i = \alpha^{-1} \text{ for } \alpha > 0$$

$\alpha$  is known as the *dispersion parameter* since increasing  $\alpha$  increases the conditional variance of y. Substituting back into our formula for the conditional variance of y,

$$\text{Var}(y_i | x) = \mu_i \left( 1 + \frac{\mu_i}{\alpha^{-1}} \right) = \exp(x_i\beta) \left( 1 + \frac{\exp(x_i\beta)}{v_i} \right) = \mu_i (1 + \alpha\mu_i) = \mu_i + \alpha\mu_i^2$$

Note that, if  $\alpha = 0$ , the mean and variance become one and the same, and you have a Poisson model.

*Heterogeneity and Contagion.* Our discussion so far has motivated the NB distribution by talking about unobserved heterogeneity. An alternative derivation is based on the idea of *contagion*. Contagion occurs when individuals with a given set of Xs have the same probability of an event occurring, but this probability changes as events occur. For example, suppose a scientist publishes a paper. Her rate of productivity may go up as a result of contagion from the initial publication. She might receive additional resources as a result of her success which will lead to further increases in productivity. A second scientist, who had the same initial rate of productivity, would have his rate stay the same so long as he did not publish. The process is contagious in the sense that success in publishing increases the rate of future publishing. Contagion violates the independence assumption of the Poisson distribution.

Unobserved heterogeneity and contagion can both generate the same NB distribution of observed counts. Consequently, heterogeneity is sometimes referred to as “spurious” or “apparent” contagion, as opposed to “true” contagion. *With cross-sectional data, it is impossible to determine whether the observed distribution of counts arose from true or spurious contagion.*

Testing for overdispersion. Remember that, with the PRM, if overdispersion is present then estimates are inefficient and standard errors are biased downward. It is therefore important to test for overdispersion. There are various ways to do this. The approaches described below take advantage of the fact that the PRM is a special case of the NBRM, when  $\alpha = 0$ .

1. You can do a 1-tailed test of  $H_0: \alpha = 0$ . (The test is one-tailed, because  $\alpha$  cannot be less than zero.) Stata’s nbreg routine reports this for you automatically:

```
. use http://www.nd.edu/~rwilliam/xsoc73994/long2006/couart2.dta, clear
(Academic Biochemists / S Long)

. nbreg art fem mar kid5 phd ment, nolog
```

```
Negative binomial regression          Number of obs   =          915
LR chi2(5)                            =          97.96
Prob > chi2                            =          0.0000
Log likelihood = -1560.9583            Pseudo R2       =          0.0304
```

art	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
fem	-.2164184	.0726724	-2.98	0.003	-.3588537	-.0739832
mar	.1504895	.0821063	1.83	0.067	-.0104359	.3114148
kid5	-.1764152	.0530598	-3.32	0.001	-.2804105	-.07242
phd	.0152712	.0360396	0.42	0.672	-.0553652	.0859075
ment	.0290823	.0034701	8.38	0.000	.0222811	.0358836
_cons	.256144	.1385604	1.85	0.065	-.0154294	.5277174
/lnalpha	-.8173044	.1199372			-1.052377	-.5822318
alpha	.4416205	.0529667			.3491069	.5586502

```
Likelihood-ratio test of alpha=0:  chibar2(01) = 180.20 Prob>=chibar2 = 0.000
```

As we see from the last line of the printout, alpha significantly differs from 0. Incidentally, what the program actually estimates is ln(alpha). This forces the estimated alpha to be positive.

2. You can do a Wald test of ln(alpha) = 1 (which corresponds to a test of alpha = 0):

```
. test [lnalpha]_cons = 1
( 1) [lnalpha]_cons = 1
      chi2( 1) = 229.59
      Prob > chi2 = 0.0000
```

To confirm this:

$$\frac{-0.8173044 - 1}{0.1199372} = \frac{-1.8173044}{0.1199372} = 15.15213295$$

Square the above and you get 229.59

3. You can do the LR chi-square test yourself by estimating both the Poisson and NBRM:

```
. quietly poisson art fem mar kid5 phd ment, nolog
. est store poisson
. quietly nbreg art fem mar kid5 phd ment, nolog
. est store nbreg
. lrtest poisson nbreg, stats force

likelihood-ratio test                                LR chi2(1) = 180.20
(Assumption: poisson nested in nbreg)              Prob > chi2 = 0.0000

-----+-----
Model      |   nobs   ll(null)   ll(model)   df       AIC       BIC
-----+-----
      poisson |   915   -1742.573   -1651.056    6       3314.113   3343.026
      nbreg   |   915   -1609.937   -1560.958    7       3135.917   3169.649
-----+-----
```

Clearly, overdispersion is a problem with the PRM in this case, and the NBRM should be preferred. This side by side comparison of the PRM and NBRM further illustrates the point:

```
. est table poisson nbreg, t label varwidth(32) stats(alpha N) b(%9.3f)
```

Variable		poisson	nbreg
art			
Gender: 1=female 0=male		-0.225	-0.216
		-4.11	-2.98
Married: 1=yes 0=no		0.155	0.150
		2.53	1.83
Number of children < 6		-0.185	-0.176
		-4.61	-3.32
PhD prestige		0.013	0.015
		0.49	0.42
Article by mentor in last 3 yrs		0.026	0.029
		12.73	8.38
Constant		0.305	0.256
		2.96	1.85
lnalpha			
Constant			-0.817
			-6.81
Statistics			
alpha			0.442
N		915.000	915.000

legend: b/t

As we see, the Poisson distribution consistently has higher t values than the NBREG distribution. The Poisson estimates are less precise and you are more likely to conclude that an effect differs from zero when in reality it does not.

Interpretation. Interpretation of the NBRM is pretty much the same as the PRM. Using the margins command,

```
. quietly nbreg art i.fem i.mar kid5 phd ment
. margins fem mar
```

```
Predictive margins                                Number of obs =          915
Model VCE      : OIM
```

```
Expression   : Predicted number of events, predict()
```

	Delta-method		z	P> z	[95% Conf. Interval]	
	Margin	Std. Err.				
fem						
0	1.868735	.0869613	21.49	0.000	1.698294	2.039176
1	1.505076	.0823171	18.28	0.000	1.343737	1.666414
mar						
0	1.542236	.1002205	15.39	0.000	1.345808	1.738665
1	1.7927	.079988	22.41	0.000	1.635926	1.949474

```
. margins, dydx(*)
```

```
Average marginal effects          Number of obs   =          915  
Model VCE      : OIM
```

```
Expression      : Predicted number of events, predict()  
dy/dx w.r.t.    : 1.fem 1.mar kid5 phd ment
```

```
-----  
          |                Delta-method  
          |          dy/dx   Std. Err.      z    P>|z|    [95% Conf. Interval]  
-----+-----  
 1.fem | - .3636591   .1211958   -3.00   0.003   - .6011984   - .1261197  
 1.mar |  .2504638   .1337954    1.87   0.061   - .0117703    .512698  
 kid5  | - .3007755   .0914704   -3.29   0.001   - .4800543   - .1214967  
  phd  |  .0260362   .0614472    0.42   0.672   - .0943981    .1464706  
  ment |  .0495833   .0065477    7.57   0.000    .0367501    .0624166  
-----
```

Note: dy/dx for factor levels is the discrete change from the base level.

Using the `listcoef` command,

```
. listcoef
```

```
nbreg (N=915): Factor Change in Expected Count
```

```
Observed SD: 1.926069
```

```
-----  
      art |          b          z    P>|z|    e^b    e^bStdX    SDofX  
-----+-----  
      fem | -0.21642   -2.978    0.003    0.8054    0.8977    0.4987  
      mar |  0.15049    1.833    0.067    1.1624    1.0738    0.4732  
 kid5   | -0.17642   -3.325    0.001    0.8383    0.8738    0.7649  
      phd |  0.01527    0.424    0.672    1.0154    1.0151    0.9842  
      ment |  0.02908    8.381    0.000    1.0295    1.3176    9.4839  
-----  
 ln alpha | -0.81730  
 alpha   |  0.44162    SE(alpha) = 0.05297  
-----  
 LR test of alpha=0: 180.20    Prob>=LRX2 = 0.000  
-----
```

Perhaps the most helpful column is  $e^b$  (which you can also get by specifying the `irr` option on `nbreg`). If you prefer, you can get equivalent results with the `percent` option:

```
. listcoef, percent
```

```
nbreg (N=915): Percentage Change in Expected Count
```

```
Observed SD: 1.926069
```

art	b	z	P> z	%	%StdX	SDofX
fem	-0.21642	-2.978	0.003	-19.5	-10.2	0.4987
mar	0.15049	1.833	0.067	16.2	7.4	0.4732
kid5	-0.17642	-3.325	0.001	-16.2	-12.6	0.7649
phd	0.01527	0.424	0.672	1.5	1.5	0.9842
ment	0.02908	8.381	0.000	3.0	31.8	9.4839
ln alpha	-0.81730					
alpha	0.44162	SE(alpha) = 0.05297				

LR test of alpha=0: 180.20 Prob>=LRX2 = 0.000

Looking at the % column, we see that, on an all other things equal basis, women are 19.5% less productive than men; married people are 16.2% more productive; each additional child lowers productivity by 16.2% (again, remember to compound, not add, for units greater than 1, e.g. somebody with 3 kids would have a rate  $.8383^2 = 58.9\%$  as great as somebody with no children); each additional article by a mentor adds 3% productivity.

See Long and Freese (2003, 2006) for additional examples, or else try the commands yourself. There aren't many new surprises here given what we have gone over before.