

Intermediate Social Statistics Hilary 2009 Lecture 7: Event Count Models

Raymond Duch

Nuffield College

March 2, 2010

Definition of Count Data

- Count data is data where the dependent variable assumes nonnegative integer values $(0, 1, 2, \dots)$ for each of n observations.
- These values represent the number of times an event occurs within a fixed observation period.
- Examples of count data would include
 - ▶ number of presidential vetos per congressional session
 - ▶ annual number of presidential nominations for the Supreme Court
 - ▶ and the number of military conflicts between countries.

Definition of Count Data

- Events occur at an unobserved expected rate of event occurrence during the observation period.
- We get to see the number of events that occurred during the period only at the end of the period.

Definition of Count Data

Least squares regression does not handle these kinds of data very well:

- The linearity assumption is inappropriate for count data.
- OLS does not constrain the expected number of events to be positive.
- Count data typically come from distributions that are heteroskedastic.
- As a result least squares tends to be inefficient and it gives inconsistent standard errors.

Definition of Count Data

All of our models will feel like probability models, of the sort:

$$\text{Prob}(\text{event } j \text{ occurs}) = \text{Prob}(Y=j) = F[\text{stochastic component, systematic component}]$$

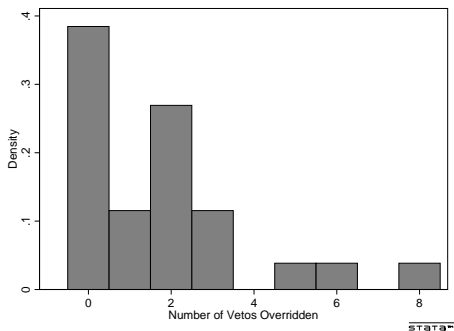
Example of Count Data: Annual Presidential Veto Overrides

Table: Presidential Veto Overrides

	Frequency	Percentage
0	10	39
1	3	12
2	7	27
3	3	12
5	1	4
6	1	4
8	1	4

Example of Count Data: Annual Presidential Veto Overrides

Figure: Number of Annual Presidential Veto Overrides



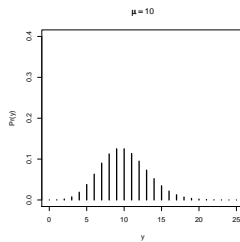
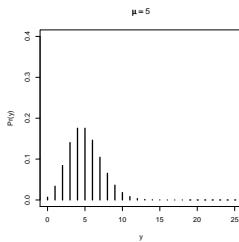
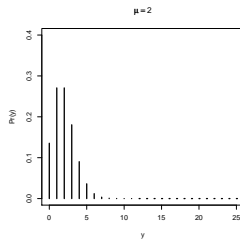
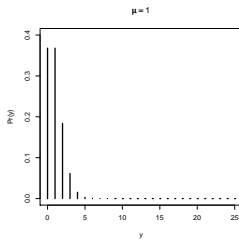
Poisson Distribution

The most basic data generating process we can use to model count data – and the veto override example above – is the Poisson distribution.

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

- as μ increases the mass of the distribution shifts to the right
- note that the $\text{Var}(y) = \text{E}(y) = \mu$

Poisson Distribution: Example

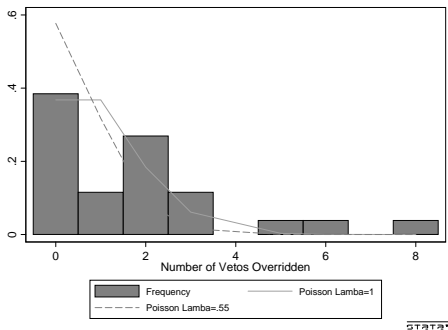


Poisson Distribution: Presidential Veto

- Modeling veto overrides as a simple poisson distribution
- with two different values: $\mu = 1$ and $\mu = .55$
- does a reasonably good job of fitting the data.

Poisson Distribution: Presidential Veto

Figure: Number of Annual Presidential Veto Overrides



The Data: Poisson Distribution

```
stset nover
```

```
      failure event: (assumed to fail at time=nover)
obs. time interval: (0, nover]
      exit on or before: failure
```

```
-----
      26 total obs.
      10 obs. end on or before enter()
```

```
-----
      16 obs. remaining, representing
      16 failures in single record/single failure data
      45 total analysis time at risk, at risk from t =          0
              earliest observed entry t =          0
              last observed exit t =          8
```

```
. poisson nover
```

```
Iteration 0:  log likelihood = -52.513187
```

```
Iteration 1:  log likelihood = -52.513187
```

```
Poisson regression
```

```
Number of obs   =          26
LR chi2(0)      =          -0.00
Prob > chi2     =          .
Pseudo R2      =          -0.0000
```

```
Log likelihood = -52.513187
```

```
-----
      nover |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      _cons |    .548566   .1490712    3.68   0.000    .2563918    .8407401
-----
```

Poisson Regression: Systematic Component

- In the Poisson regression model the number of events, y_i , has a Poisson distribution, as above
- except now it has a conditional mean that depends on an individual's characteristics
- hence we have added a systematic component to the data generating process, or to the distribution function
- The systematic component of the model can be specified as:

$$\mu_i = E(y_i|x_i) = \exp(\mathbf{x}_i\beta)$$

Poisson Regression Model

- these μ are by construction positive and then drive the probability calculations

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

Likelihood for the Poisson Regression Model

$$L(\beta|\mathbf{y}, \mathbf{X}) = \prod_{i=1}^N P(y_i|\mu_i) = \prod_{i=0}^N \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}$$

where $\mu_i = \exp(\mathbf{X}\beta)$

and assuming there is independence across observations of different counts

Poisson Regression of Veto Overrides

```
. poisson nover nveto janpop preshmaj pressmaj
```

```
Iteration 0: log likelihood = -37.96409
Iteration 1: log likelihood = -37.908086
Iteration 2: log likelihood = -37.907938
Iteration 3: log likelihood = -37.907938
```

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

Log likelihood = -37.907938
```

```
-----+-----
      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
      preshmaj | -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
-----+-----
```

Poisson Regression Model Fit

- What do we gain by incorporating this systematic component to the poisson distribution function?
- By adding this systematic component we are taking into account the possibility that μ_i varies across years (or more accurately political contexts).
- In other words the rate of presidential overrides varies by political contexts.
- This is referred to as capturing heterogeneity in the sample

Poisson Regression Model Fit: Log Likelihood Ratio Test

The log likelihood ratio test is simply:

$$LRI = 2 * (\ln L - \ln L_0) \quad (1)$$

- where L_0 is the log-likelihood computed only with a constant, and $\ln L$ is log-likelihood with the systematic component included.
- This has a chi square distribution with degrees of freedom equal to the number of restrictions imposed.

Poisson Regression Model Fit: Log Likelihood Ratio Test

- So in the Poisson example above the first Poisson estimate without any systematic component provides $\ln L_0$ which is -52.51 and the estimate with the systematic component provides the unconstrained $\ln L$ which is -37.90 .
- Hence the log likelihood ratio equals $2 * (-37.9 - (-52.5)) = 29.2$
- The Chi-square p value associated with 4 degrees and a value of 29.2 is 0.000.
- Hence adding this systematic component to the model significantly improves on the fit between the predictions from the Poisson distribution and the actual data.

Variables in the PRM for Presidential Vetos

First of all what are the variables in the systematic component of the model.

```
. summarize nover nveto janpop preshmaj 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
preshmaj	26	.3846154	.4961389	0	1
pressmaj	26	.5	.509902	0	1

Predictions: Setting Independent Variables to Interesting Values

- In order to generate a prediction we need values of independent variable associated with the prediction we want to make
- For example, what is the predicted rate of veto overrides if:
 - ▶ the number of vetos equal the mean: 16
 - ▶ the president's popularity level is at its mean: 58
 - ▶ the president has a majority in the house: 1
 - ▶ the president has a majority in the senate: 1

Predictions: Calculating Systematic Component

- First calculate the value of the systematic term in the Poisson Regression Model which is
$$\mu = \exp[1.72 + .04(16) - .029(58) - 1.17(1) + 0.008(1)] = .616$$
- Then calculate the probability using the Poisson distribution including the the μ term we generated.

$$Prob(y_i = 2 | \mu_i = .616) = \frac{\exp(-.616) * (-.616)^2}{2!} = .10$$

Generating Predictions and Standard Errors with Clarify

```
. simqi, prval(0 1 2 3 4 5 6 7 8)
```

Quantity of Interest	Mean	Std. Err.	[95% Conf. Interval]	
Pr(nover=0)	.5503067	.1307383	.2807208	.7684717
Pr(nover=1)	.3123263	.0492409	.2023782	.3677523
Pr(nover=2)	.1036074	.0541755	.0266483	.226528
Pr(nover=3)	.0265253	.0252612	.0023393	.0959267
Pr(nover=4)	.0058239	.0087028	.000154	.0304662
Pr(nover=5)	.0011542	.0025519	8.11e-06	.0077408
Pr(nover=6)	.0002122	.0006716	3.56e-07	.001639
Pr(nover=7)	.0000368	.0001617	1.34e-08	.0002975
Pr(nover=8)	6.06e-06	.0000358	4.41e-10	.0000472

Predictions: Generating Expected Outcomes

- We can also generate the expected value of y , or the incidence rate, for a given value of \mathbf{x}_i .
- where $\mu = E(y|\mathbf{X}) = \exp(\mathbf{X}\beta)$
- Rather than calculating the probability that the number of veto overrides equal 2 for a given set of values on the independent variables \mathbf{X}
- we can calculate the expected number of veto overrides associated with a political contexts corresponding to values on the independent variables that we used earlier, i.e.,

Predictions: Generating Expected Outcomes for Values of Independent Variable

- the number of vetos equal the mean: 16
- the president's popularity level is at its mean: 58
- the president has a majority in the house: 1
- the president has a majority in the senate: 1

The expected number of veto overrides is calculated as:

$$E(y|\mathbf{X}) = \exp(\mathbf{X}\beta) = .616$$

which is about a half a veto.

Interpreting the effect of changes in the independent variable: factor change

The impact of a δ factor change in one of the independent variables (\mathbf{x}_k) on the expected incidence rate is calculated as follows

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

Interpreting the effect of changes in the independent variable: factor change

- Take, for example, a $\delta = 1$ (or a unit change in) number of vetos (nveto)
- This results in an increase in the expected veto override count by a factor of $e^{(.04X1)}$ which is 1.04, holding all other variables constant
- You can request that Stata generate these factor changes associated with a 1 unit change in each of the independent variables by specifying the irr option.

Factor Change

```
. poisson nover nveto janpop preshmaj pressmaj, irr
```

```
Iteration 0:  log likelihood = -37.96409
Iteration 1:  log likelihood = -37.908086
Iteration 2:  log likelihood = -37.907938
Iteration 3:  log likelihood = -37.907938
```

Poisson regression

```
Number of obs   =      26
LR chi2(4)      =     29.21
Prob > chi2     =     0.0000
Pseudo R2      =     0.2781
```

Log likelihood = -37.907938

nover	IRR	Std. Err.	z	P> z	[95% Conf. Interval]	
nveto	1.041535	.0094753	4.47	0.000	1.023128	1.060273
janpop	.9707421	.0153473	-1.88	0.060	.9411233	1.001293
preshmaj	.3109959	.1963091	-1.85	0.064	.0902512	1.071658
pressmaj	1.008526	.477919	0.02	0.986	.3984012	2.553015

Recall the assumption about data generation that is made for Poisson

- The rate of event occurrence within the "bin" – during the year for example – is constant
- The probability that an event occurs inside a bin is independent of whether any other events have already occurred
- This is the basis for assuming that the conditional variance equals the conditional mean.
- If this is violated we move to other count models such as the negative binomial.

Negative Binomial Regression Model

$$Pr(y_i|x_i) = \frac{\Gamma(y_i + \nu_i)}{y_i! \Gamma(\nu_i)} \left(\frac{\nu_i}{\nu_i + \mu_i}\right)^{\nu_i} \left(\frac{\mu_i}{\nu_i + \mu_i}\right)^{y_i} \quad (2)$$

where $E(y_i|\mathbf{x}_i) = \exp(\mathbf{x}_i\beta) = \mu_i$

$$Var(y_i|\mathbf{x}_i) = \mu(1 + \frac{\mu_i}{\nu_i}) = \exp(\mathbf{x}_i\beta)(1 + \frac{\exp(\mathbf{x}_i\beta)}{\nu_i}) \quad (3)$$

where we assume that $\nu_i = \alpha^{-1}$

$$Var(y_i|\mathbf{x}_i) = \mu_i(1 + \frac{\mu_i}{\alpha^{-1}}) = \mu_i(1 + \alpha\mu_i) = \mu_i + \alpha\mu_i^2 \quad (4)$$

Negative Binomial Regression of Veto Overrides

```
. nbreg nover nveto janpop preshmaj pressmaj
```

Negative binomial regression

Number of obs = 26

LR chi2(4) = 18.03

Dispersion = mean

Prob > chi2 = 0.0012

Log likelihood = -37.609976

Pseudo R2 = 0.1933

```
-----+-----
```

nover	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
nveto	.0425453	.01128	3.77	0.000	.0204368	.0646538
janpop	-.035034	.0188896	-1.85	0.064	-.072057	.0019889
preshmaj	-1.149386	.6660737	-1.73	0.084	-2.454866	.1560944
pressmaj	-.0048036	.5338629	-0.01	0.993	-1.051156	1.041549
_cons	1.983813	1.044684	1.90	0.058	-.0637302	4.031356
-----+-----						
/lnalpha	-2.083274	1.598841			-5.216946	1.050398
-----+-----						
alpha	.1245219	.1990907			.0054239	2.858787

Likelihood-ratio test of alpha=0: chibar2(01) = 0.60 Prob>=chibar2 = 0.220