POLS585 - Advanced Regression
October 29, 2002
Event Count Models – Poisson Regression

1 Event Count Models

Event count models are models where the dependent variable is a count of events. So, we're considering a variable Y_i where $Y \in \{0, 1, 2, ...\}$. Event count variables are thus *nonnegative integers* – bounded at zero below, unbounded above.

Note a couple things that *aren't* event count data:

- **Ordinal Data**
  - Items such as Likert scales may look like event counts, but they aren't...
  - Use ordered logit/probit instead.
- **Grouped binary data**
  - Data which are the number of "successes" (or "failures") out of some known number of binary trials.
  - E.g.: the number of successful veto overrides in each Congress, or the number of failed coup attempts in a given nation.
  - Grouped binary data can be expressed as event counts, but are not event counts.
    * One should (generally) not use event count models for grouped binary data.
    * The exception is when there are relatively few "successes", relative to the possible number of trials (see below).

Event counts have a few interesting properties; they are discrete, and (as noted above) nonnegative; both of these things can make using OLS on event counts a bad idea...

- King (1988) tells how and why, specifically.
- In general: OLS on event counts yields estimates which are:
  - *Inaccurate* (e.g., can yield negative counts).
  - *Inefficient* (e.g., because they fail to account for the heteroscedastic nature of event counts-more on this in a bit...).

2 Event Count Data

2.1 The Poisson Distribution

A good place to start is with an abstract model of event counts. Suppose we are interested in studying events, and that those events occur over time. We might consider the *constant rate* at which events occur; call this rate λ.

- Think of this as the expected number of events in any particular time "period" of length h.
- Further imagine a process in which events are *independent*; that is, the occurrence of one event has no bearing on the probability that another will occur...
- As the length of the interval $h \to 0$,
  - The probability of an event occurring in the interval $(t, t+h] = \lambda h$
  - The probability of no event occurring in the interval $(t, t+h] = 1 - \lambda h$

Next, consider our outcome variable Y_t as the number of events that have occurred in the interval t of length h. For such a process, the probability that the number of events occurring in $(t, t+h]$ is equal to some value $y \in \{0, 1, 2, 3, ...\}$ is:

$$Pr(Y_t = y) = \frac{exp(-\lambda h)\lambda h^y}{y!} \quad (1)$$

This variable is what is known as a *Poisson process*: events occur independently with a constant probability equal to λ times the length of the interval (that is, λh). If all the intervals are of the same length equal to 1, this reduces to:

$$Pr(Y_t = y) = \frac{exp(-\lambda)\lambda^y}{y!} \tag{2}$$

This is the way we typically see the *Poisson distribution* written.

2.2 Other Motivations

There are other ways of motivating the Poisson distribution; e.g., as a count of "rare events"...

- For a large number of Bernoulli trials, where the probability of an event in any one trial is small, the total number of events observed will follow a Poisson distribution.
- Formally, for M independent Bernoulli trials with (small) probability of success π and where $M\pi \equiv \lambda < 0$,[1] the probability of observing exactly y total "successes" is:

$$\begin{aligned} Pr(Y_i = y) &= \lim_{M \to \infty} \left[\binom{M}{y} \left(\frac{\lambda}{M}\right)^y \left(1 - \frac{\lambda}{M}\right)^{M-y} \right] \\ &= \frac{\lambda^y e^{-\lambda}}{y!} \end{aligned}$$

Cameron and Trivedi (1998, Ch. 1) call this the "Law of Rare Events" motivation; see their book for other ways to motivate the Poisson. (We'll talk about the notion of "counting processes" later in the course, when we do survival analysis...).

[1]Formally, holding λ constant as $M \to \infty$ requires that $\pi \to 0$.

2.3 Characteristics

The Poisson distribution has several important traits:

- It is a discrete probability distribution, with support on the non-negative integers.
- The "rate" can also be interpreted as the expected number of events during an observation period t. In fact, for a Poisson variate Y, $E(Y) = \lambda$.
- As λ increases, several interesting things happen... (see Figure 1):
  1. The *mean/mode* of the distribution gets bigger (no shock there).
  2. The *variance* of the distribution gets larger as well. This also makes sense: since the variable is bounded from below, its variability will necessarily get larger with its mean. In fact, in the Poisson, the mean equals the variance (that is, $E(Y) = Var(Y) = \lambda$).
  3. The distribution becomes more Normal-looking (and, in fact, becomes more Normal, period).

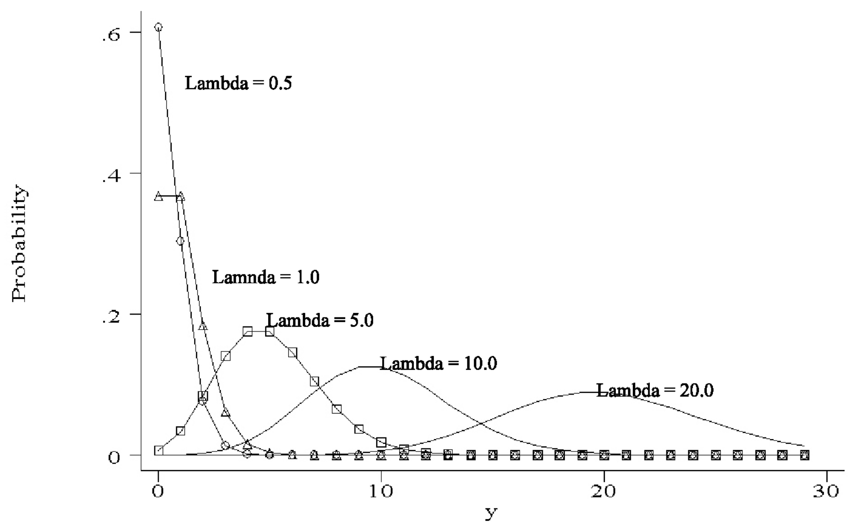
3 A Poisson Model with Covariates

If we assume our event count is distributed according to a Poisson distribution, then the next thing we probably want to know is the effect of covariates $\mathbf{X}_i$ on (the expected value of) Y_i.

- Since we know that $E(Y) > 0$, we need to restrict the "link" to be positive.
- The exponential is the standard one for this...
- So, we have

$$E(Y_i) \equiv \lambda_i = exp(\mathbf{X}_i\beta) \tag{3}$$

Figure 1: Poisson PDFs, with varying λs



- This yields a probability model that looks like this:

$$Pr(Y_i = y|\mathbf{X}_i) = \frac{exp[-exp(\mathbf{X}_i\beta)][exp(\mathbf{X}_i\beta)]^y}{y!} \tag{4}$$

- This Poisson distribution then yields a pretty simple likelihood:

$$L = \prod_{i=1}^{N} \frac{exp[-exp(\mathbf{X}_i\beta)][exp(\mathbf{X}_i\beta)]^{Y_i}}{Y_i!} \tag{5}$$

- and an equally simple log-likelihood:

$$lnL = \sum_{i=1}^{N} \{-exp(\mathbf{X}_i\beta) + Y_i\mathbf{X}_i\beta - ln(Y_i!)\} \tag{6}$$

where the last term $-ln(Y_i!)$ can be omitted because it doesn't vary with β. This log-likelihood is globally concave, and so estimation is really easy and reliable.

3.1 Exposure and Offsets

As noted in (1) above, the general format for the Poisson distribution takes into account the extent of "exposure" of each subject. In the Bernoulli/"rare events" formulation of the Poisson model, above, Y_i is the number of events and M_i is the number of "trials" (that is, the number of possible events). There, we assumed $M \to \infty$, but that isn't always the case.

Consider, for example, a model of the number of Supreme Court decisions in each term in which there is at least one dissenting vote/opinion (*a la* Caldeira and Zorn 1998). There's an upper limit on this number: the number of total decisions by the Court – obviously, we can't have $Y_i = 10$ if the Court only decided 9 cases in term i. This factor is often referred to as an observation's *exposure* (after a number of applications in biometrics).

There are a few things to remember about the notion of exposure:

1. Formally, if each of the observations doesn't have the same exposure (that is, if $M_i \neq M_j \forall i \neq j$), then the expected count is proportional to the exposure (so that $E(Y_i|\mathbf{X}_i, M_i) = \lambda_i M_i$), and so exposure needs to be accounted for in some way.

2. The easiest way to do this is to include M_i as an *offset*:

$$\lambda_i = exp[\mathbf{X}_i\beta + ln(M_i)]$$

   This amounts to including $ln(M_i)$ among the right–hand–side variables, and constraining the coefficient to 1.0. Most software programs have straightforward ways to do this (e.g., in `Stata`, one uses the `-offset-` option to `-poisson-`).

3. In fact, one can even just include $ln(M_i)$ among the covariates, and then test whether or not $\hat{\beta}_M = 1.0$.

4. This might even be of some substantive interest – for example, if we were modeling the aforementioned number of "dissent cases", it might be interesting to know whether, as the Court's workload increased, the ability/propensity for justices to cast dissenting votes or author dissenting opinions (say) decreased.

5. As a practical matter, if (a) all of the observations have the same (or very similar) exposures, or (b) for the most part, Y_i is significantly less than M_i (that is, no observation "comes close" to experiencing its maximum possible number of events) then the issue of exposure is not a big deal, and you can probably safely ignore it.

4 An Example: Judicial review of Congressional acts

We'll consider an example, where Y_i is the number of Acts of Congress overturned by the Supreme Court in each Congress, 1789–1996 ($N = 104$). We'll consider two covariates:

- The *mean tenure* (`tenure`) of the Supreme Court's justices ($\bar{X} = 10.4, \sigma = 3.4, E(\hat{\beta}) > 0$).

- Whether (1) or not (0) there was *unified government* (`unified`) ($\bar{X} = 0.83, E(\hat{\beta}) < 0$).

Here are the results, from Stata's `-poisson-` command:

```
. poisson nulls tenure unified

Iteration 0:   log likelihood = -189.53751
Iteration 1:   log likelihood = -189.53751

Poisson regression                                Number of obs   =        104
                                                  LR chi2(2)      =      14.27
                                                  Prob > chi2     =     0.0008
Log likelihood = -189.53751                       Pseudo R2       =     0.0363

------------------------------------------------------------------------------
  nulls |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
--------+---------------------------------------------------------------------
 tenure |   .0958868   .0256271     3.742   0.000      .0456585     .146115
unified |   .1434999   .2327122     0.617   0.537     -.3126077    .5996074
  _cons |  -.8776214   .3712678    -2.364   0.018     -1.605293   -.1499499
------------------------------------------------------------------------------
```

Note how quickly the results converged, thanks in large part to the aforementioned fact that the likelihood is globally concave...

4.1 Interpretation...

4.1.1 Signs–n–Significance

Signs indicate the effect on the expected number of counts...

- So, positive signs indicate positive effects – higher values of X correspond to higher counts.
- Here, tenure seems to matter, but unified/divided government doesn't.

4.1.2 Incident Rate Ratios

As described here, the Poisson model is a log-linear model not unlike the various flavors of logit we talked about before. This means that (surprise, surprise!) there is an odds–ratio interpretation of Poisson regression coefficients, just as there is for those models.

In the Poisson case, the quantity of interest is known as the *incidence rate* – that is, $\hat{\lambda}$. The natural way to compare two observations, then, is the *incidence rate ratio* (or *IRR*). For e.g. a binary covariate X_D, we can think of the IRR as the ratio

$$\frac{\hat{\lambda}|X_D = 1}{\hat{\lambda}|X_D = 0} = \frac{exp(\hat{\beta}_0 + \bar{X}\hat{\beta} + (X_D = 1)\hat{\beta}_{X_D})}{exp(\hat{\beta}_0 + \bar{X}\hat{\beta} + (X_D = 0)\hat{\beta}_{X_D})} = exp(\hat{\beta}_{X_D})$$

That is, we can tell the relative change in the incidence rate for a one–unit change in any given variable X_k by simply exponentiating its coefficient estimate $\hat{\beta}_k$.

- In the example, this means that the estimated IRR for the `unified` variable is equal to $exp(0.143) = 1.15$.
- This means that the incidence rate under unified government is about 1.15 times that under divided government (that is $\lambda_{unified} = 1/15 \times \lambda_{divided}$), which ain't a large change.
- Similarly, a one–year change in the `tenure` variable corresponds to an estimated IRR of $exp(0.096) = 1.10$ – increasing the average tenure of the Court increases the estimated incidence rate by a factor of 1.1.
- More generally, for a δ–unit change in X_k is $exp(\delta\hat{\beta}_k)$.
  - Thus, a 10–year change in `tenure` corresponds to an estimated IRR of $exp(10 \times 0.096) = exp(0.96) = 2.61$.
  - That is, a Court with an average `tenure` of t years will be expected to have an incidence of judicial review roughly 2.6 times that of a Court with a mean `tenure` equal to $t - 10$.
- Note as well that `Stata` will report IRRs for you automatically; simply specify the `-irr-` option to the `-poisson-` command:

```
. poisson, irr

Poisson regression                                Number of obs   =        104
                                                  LR chi2(2)      =      14.27
                                                  Prob > chi2     =     0.0008
Log likelihood = -189.53751                       Pseudo R2       =     0.0363
------------------------------------------------------------------------------
   nulls |       IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  tenure | 1.100634   .0282061     3.74   0.000     1.046717    1.157329
 unified | 1.154307   .2686212     0.62   0.537     .7315369    1.821404
------------------------------------------------------------------------------
```

4.1.3 Predicted Counts

- The predicted count is just $exp(\bar{\mathbf{X}}\hat{\beta})$.
- This is pretty easy to calculate.
- E.g., for a typical case (that is, a unified government in which the average Court tenure is ten years), we get a predicted count of:

$$
\begin{aligned}
E(Y|\bar{\mathbf{X}}_i) &= exp[-0.878 + (0.096 \times 10) + (0.143 \times 1)] \\
&= exp(0.225) \\
&= 1.25
\end{aligned}
$$

- You can accordingly calculate the change in expected counts by calculating the predicted count for different values of $\mathbf{X}_i$, and taking the difference.
- E.g., the expected count for the same Congress with an average Court `tenure` of 20 years is $exp(1.185) = 3.27$.
- So, a ten-year increase in tenure corresponds to a $(3.27 - 1.25) \approx 2$–act increase in judicial review.
- Note that $\frac{3.27}{1.25} = 2.61$, *which is the same as the IRR for a ten–year change reported above...*

- Predicted counts can be interesting either for *within–sample* or *out–of–sample* predictions.
- Be sure to include measures of uncertainty here as well (*Clarify* will help here...).

To graph out–of–sample predictions as a function of continuous covariates, we adopt the same "dummy dataset" strategy we've been using for logit, etc.:

```
. clear

. set obs 21

. gen unified = 1

. gen tenure=_n-1

. save tensim.dta

. use notes9-1.dta, clear

. poisson nulls unified tenure
```

(output omitted)

```
. use tensim.dta, clear

. predict xb, xb

. predict stdp, stdp

. gen counthat=exp(xb)

. gen ub=exp(xb+(1.96*stdp))

. gen lb=exp(xb-(1.96*stdp))

. gra counthat ub lb tenure, c(lll) s(o..) xlab ylab t1(" ")
```

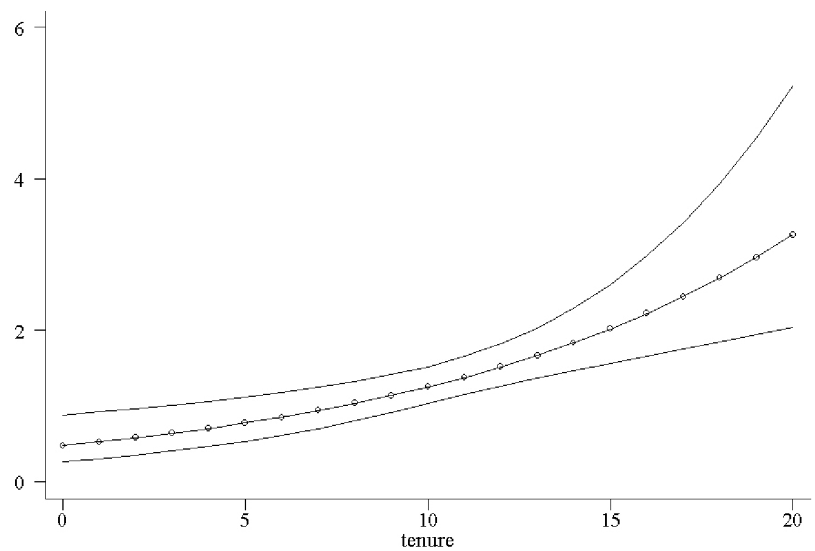


Figure 2: $\widehat{E(Y)}$ as a function of `tenure`, with 95% c.i.s

which yields Figure 2.

4.1.4 Predicted Probabilities

We might be interested in the probability that a particular observation Y_i takes on a particular count value y. We can get this predicted probability by plugging the $\mathbf{X}$ values for that observation, and the estimates of $\hat{\beta}$, into the basic Poisson probability statement:

$$Pr(\widehat{Y_i = y}|\mathbf{X}_i, \hat{\beta}) = \frac{exp[-exp(\mathbf{X}_i\hat{\beta})][exp(\mathbf{X}_i\hat{\beta})]^y}{y!}$$

Example: for the above case, what are the probabilities of counts equalling 0,1,2, or 3?

$$
\begin{aligned}
\widehat{Pr(Y_i = 0|}\mathbf{X}_i, \hat{\beta}) &= \frac{[exp(-1.25)](1.25)^0}{0!} \\
&= \frac{(0.287)(1)}{1} \\
&= 0.287
\end{aligned}
$$

$$
\begin{aligned}
\widehat{Pr(Y_i = 1|}\mathbf{X}_i, \hat{\beta}) &= \frac{[exp(-1.25)](1.25)^1}{1!} \\
&= \frac{(0.287)(1.25)}{1} \\
&= 0.359
\end{aligned}
$$

$$
\begin{aligned}
\widehat{Pr(Y_i = 2|}\mathbf{X}_i, \hat{\beta}) &= \frac{[exp(-1.25)](1.25)^2}{2!} \\
&= \frac{(0.287)(1.563)}{2} \\
&= 0.224
\end{aligned}
$$

$$
\begin{aligned}
\widehat{Pr(Y_i = 3|}\mathbf{X}_i, \hat{\beta}) &= \frac{[exp(-1.25)](1.25)^3}{3!} \\
&= \frac{(0.287)(1.953)}{6} \\
&= 0.093
\end{aligned}
$$

- If you add these you get 0.963, which tells you that these outcomes account for most of the potential outcomes at this level of the covariates.
- Obviously, it would be possible to calculate this for a larger range, and to automate this using Stata or whatever, and to graph them...
- Long also suggests calculating the *mean predicted probability* of each possible count, across all observations, as a measure that can be used to assess model fit...

Next time: Over– and underdispersion and event counts...