14 Poisson Regression

In class we will cover Chapter 12 (Analysis of Rates with Poisson Regression) from Steve Selvin's text *Practical Biostatistical Methods* (1995, Wadsworth). There are many more applications of Poisson regression than covered there, but this chapter has a treatment quite relevant to you.

The appendix in Selvin discusses the Poisson distribution mostly as an approximation to the binomial distribution when n is large and p is small. A broader more detailed perspective can be found at Wikipedia (`http://en.wikipedia.org/wiki/Poisson_distribution`). One entry there states that

> The word **law** is sometimes used as a synonym of probability distribution, and *convergence in law* means convergence in distribution. Accordingly, the Poisson distribution is sometimes called the **law of small numbers** because it is the probability distribution of the number of occurrences of an event that happens rarely but has very many opportunities to happen. *The Law of Small Numbers* is a book by Ladislaus Bortkiewicz about the Poisson distribution, published in 1898. Some historians of mathematics have argued that the Poisson distribution should have been called the Bortkiewicz distribution.

Wikipedia provides several examples of the Poisson as well:

> The Poisson distribution arises in connection with Poisson processes. It applies to various phenomena of discrete nature (that is, those that may happen 0, 1, 2, 3, ... times during a given period of time or in a given area) whenever the probability of the phenomenon happening is constant in time or space. Examples of events that can be modelled as Poisson distributions include:
>
> - The number of cars that pass through a certain point on a road during a given period of time.
> - The number of spelling mistakes a secretary makes while typing a single page.
> - The number of phone calls at a call center per minute.
> - The number of times a web server is accessed per minute. For instance, the number of edits per hour recorded on Wikipedia's Recent Changes page follows an approximately Poisson distribution.
> - The number of roadkill found per unit length of road.
> - The number of mutations in a given stretch of DNA after a certain amount of radiation.
> - The number of unstable nuclei that decayed within a given period of time in a piece of radioactive substance. The radioactivity of the substance will weaken with time, so the total time interval used in the model should be significantly less than the mean lifetime of the substance.
> - The number of pine trees per unit area of mixed forest.
> - The number of stars in a given volume of space.
> - The number of soldiers killed by horse-kicks each year in each corps in the Prussian cavalry. This example was made famous by a book of Ladislaus Josephovich Bortkiewicz (1868-1931).
> - The distribution of visual receptor cells in the retina of the human eye.

- The number of V2 rocket attacks per area in England, according to the fictionalized account in Thomas Pynchon's Gravity's Rainbow.
- The number of light bulbs that burn out in a certain amount of time.

For our purposes two more interesting examples probably are number of deaths in a subpopulation in a certain amount of time and number of cases of a disease in a subpopulation in a fixed period of time. What is key in these and the other examples is that the Poisson distribution describes random counts. Selvin deals with rates (proportions) which probably are more common but require some special consideration.

Mathematical Background

Let the random variable Y have a Poisson distribution with parameter λ. This means that

$$P(Y=k)=\frac{e^{-\lambda}\lambda^k}{k!};\ k=0,1,2,3,\dots$$

and both the mean $E(Y)$ and variance $Var(Y)$ of Y are λ. As with the binomial distribution where we found the logit function and logistic regression to be more useful than considering the binomial distribution directly, we fit regression models to $\log Y$ (natural log), so we get models of the form

$$\log E(Y)=\beta_0+\beta_1x_1+\ldots+\beta_px_p$$

where there are good theoretical reasons for logit with binomial and log for the Poisson. These are special cases of a large class of such models called *generalized linear models* (handled with the `glm` command in Stata in general, more easily handled with the `poisson` command for this case).

Poisson regression fits linear models to log(counts of number of events) – remember, fitting linear models means looking for group differences, interactions, adjusting for covariates and confounders,etc.. The events we are looking at probably will be deaths or diagnosis of disease. In the applications you are likely to encounter most often, we actually want to fit linear models to rates, where rates are usually of the form

$$r=\frac{\text{count of events}}{\text{population size}},\qquad \text{or}\qquad r=\frac{\text{count of events}}{\text{total exposure}}.$$

Either way, we probably want to model r more than we want to model counts directly. The way we do this is to fit a linear model to log of r

$$\begin{aligned}
\log r &= \hat\beta_0+\hat\beta_1x_1+\ldots+\hat\beta_px_p && (1)\\
\log\left(\frac{\text{count of events}}{\text{population size}}\right) &= \hat\beta_0+\hat\beta_1x_1+\ldots+\hat\beta_px_p && (2)\\
\log(\text{count of events})-\log(\text{population size}) &= \hat\beta_0+\hat\beta_1x_1+\ldots+\hat\beta_px_p && (3)\\
\log(\text{count of events}) &= \hat\beta_0+\hat\beta_1x_1+\ldots+\hat\beta_px_p+\log(\text{population size}) && (4)
\end{aligned}$$

Equation (4) shows that we model $\log r$ by writing a linear model for log (count of events), which is regular Poisson regression, except that we have a special new variable with a known coefficient of 1, i.e. log (population size). Such a variable is called an *offset*. The general form would be the same if we had a rate using exposure instead of population size.

In order to interpret coefficients, consider the simple case of two groups (say F and M) where $x_1=\begin{cases}0 & \text{Group} = \text{F}\\ 1 & \text{Group} = \text{M}\end{cases}$, the usual group indicator variable, and we fit a simple model to compare groups, $\log r = \hat{\beta}_0 + \hat{\beta}_1 x_1$. Then F is the reference group and $\hat{\beta}_1$ is the difference between groups M and F in the log scale, just as we usually have in linear models, i.e. $\log r_F = \hat{\beta}_0$ and $\log r_M = \hat{\beta}_0 + \hat{\beta}_1$, so $\log r_M - \log r_F = \hat{\beta}_1$ and $\frac{r_M}{r_F} = e^{\log r_M - \log r_F} = e^{\hat{\beta}_1}$. Similar to the way we obtained estimated Odds Ratios in logistic regression, we obtain estimated *incidence-rate ratios* by exponentiating estimated regression coefficients. Factors with multiple levels and continuous predictor variables are handled similarly to the way we have handled then in least squares regression and in logistic regression.

Stata Implementation

A portion of the help viewer in Stata for the Poisson command shows

```
Syntax
        poisson depvar [indepvars] [if] [in] [weight] [, options]

    options                        description
    -------------------------------------------------------------------------
    Model
      noconstant                   suppress constant term
      exposure(varname_e)          include ln(varname_e) in model with
                                     coefficient constrained to 1
      offset(varname_o)            include varname_o in model with coefficient
    Reporting
      level(#)                     set confidence level; default is level(95)
      irr                          report incidence-rate ratios

    irr reports estimated coefficients transformed to incidence-rate ratios,
        that is, exp(b) rather than b.  Standard errors and confidence
        intervals are similarly transformed.  This option affects how results
        are displayed, not how they are estimated or stored.  irr may be
        specified at estimation or when replaying previously estimated
        results.

    offset(varname) specifies that varname be included in the model with the
        coefficient constrained to be 1.

    exposure(varname) specifies a variable that reflects the amount of
        exposure over which the depvar events were observed for each
        observation; ln(varname) with coefficient constrained to be 1 is
        entered into the log-link function.
```

From a practical perspective, what difference does it make if you declare a variable an offset or an exposure? Read carefully – you need to already have taken the log for an offset, but Stata will go ahead and take the log of an exposure variable. You can use either form, just be careful you have the variable in the correct form.

Example

Consider the very simple data set

```
       group   deaths   popsize
  1.       F       10     10000
  2.       M       15      8000
```

and compare mortality rates for the two groups. Define `lpopsize` from the command
`gene lpopsize = log(popsize)`. We want to fit using both the `offset` and `exposure` forms of the command to see they agree.

```
. xi:poisson deaths i.group, exposure(popsize)
i.group           _Igroup_1-2         (_Igroup_1 for group==F omitted)
Iteration 0:   log likelihood =   -4.35708
Iteration 1:   log likelihood =   -4.35708
Poisson regression                                Number of obs   =          2
                                                  LR chi2(1)      =       2.43
                                                  Prob > chi2     =     0.1188
Log likelihood =   -4.35708                       Pseudo R2       =     0.2183
------------------------------------------------------------------------------
      deaths |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Igroup_2 |   .6286087   .4082483     1.54   0.124    -.1715433    1.428761
       _cons |  -6.907755   .3162278   -21.84   0.000     -7.52755    -6.28796
     popsize | (exposure)
------------------------------------------------------------------------------
. xi:poisson deaths i.group, offset(lpopsize)
i.group           _Igroup_1-2         (_Igroup_1 for group==F omitted)
Iteration 0:   log likelihood =   -4.35708
Iteration 1:   log likelihood =   -4.35708
Poisson regression                                Number of obs   =          2
                                                  LR chi2(1)      =       2.43
                                                  Prob > chi2     =     0.1188
Log likelihood =   -4.35708                       Pseudo R2       =     0.2183
------------------------------------------------------------------------------
      deaths |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Igroup_2 |   .6286087   .4082483     1.54   0.124    -.1715433    1.428761
       _cons |  -6.907755   .3162278   -21.84   0.000     -7.52755    -6.28796
    lpopsize |   (offset)
------------------------------------------------------------------------------
. poisson,irr
Poisson regression                                Number of obs   =          2
                                                  LR chi2(1)      =       2.43
                                                  Prob > chi2     =     0.1188
Log likelihood =   -4.35708                       Pseudo R2       =     0.2183
------------------------------------------------------------------------------
      deaths |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Igroup_2 |      1.875   .7654656     1.54   0.124     .8423638    4.173523
    lpopsize |   (offset)
------------------------------------------------------------------------------
```

But look what happens if you mix offset and exposure:

```
. xi:poisson deaths i.group, offset(popsize)
Poisson regression                                Number of obs   =          2
                                                  Wald chi2(1)    =  2.401e+07
Log likelihood =   -4.35708                       Prob > chi2     =     0.0000
------------------------------------------------------------------------------
      deaths |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Igroup_2 |   2000.405   .4082483  4899.97   0.000     1999.605    2001.206
       _cons |  -9997.697   .3162278 -3.2e+04   0.000    -9998.317   -9997.078
     popsize |   (offset)
------------------------------------------------------------------------------
```

We will do a number of the examples in Selvin's chapter during class.