Models In Epidemiology And Biostatistics
Gordon Hilton Fick

Session 12 :
Count Data, Incidence, Person-Time & Models for Counts

A famous cohort study

In a famous age-specific study of coronary disease deaths among male British doctors, Doll and Hill (1966) reported the following data

	Smokers		Nonsmokers	
Age	Deaths	Person-years	Deaths	Person-years
35–44	32	52,407	2	18,790
45–54	104	43,248	12	10,673
55–64	206	28,612	28	5,710
65–74	186	12,663	28	2,585
75–84	102	5,317	31	1,462

Stratifying on age, the estimates of the incidence rate ratios (IRR) would be:

```
. list

     +-----------------------------------+
     | agecat   smokes   deaths   pyears |
     |-----------------------------------|
  1. |      1        1       32   52,407 |
  2. |      2        1      104   43,248 |
  3. |      3        1      206   28,612 |
  4. |      4        1      186   12,663 |
  5. |      5        1      102    5,317 |
     |-----------------------------------|
  6. |      1        0        2   18,790 |
  7. |      2        0       12   10,673 |
  8. |      3        0       28    5,710 |
  9. |      4        0       28    2,585 |
 10. |      5        0       31    1,462 |
     +-----------------------------------+

. ir deaths smokes pyears, by(agecat)

          agecat |      IRR       [95% Conf. Interval]   M-H Weight
-----------------+------------------------------------------------
               1 |   5.736638      1.463557   49.40468     1.472169 (exact)
               2 |   2.138812      1.173714   4.272545     9.624747 (exact)
               3 |    1.46824      .9863624   2.264107     23.34176 (exact)
               4 |    1.35606      .9081925   2.096412     23.25315 (exact)
               5 |   .9047304      .6000757   1.399687     24.31435 (exact)
-----------------+------------------------------------------------
           Crude |   1.719823      1.391992    2.14353                (exact)
    M-H combined |   1.424682      1.154703   1.757784
  Test of homogeneity (M-H)    chi2(4) =     10.41  Pr>chi2 = 0.0340
```

The 'classical' analysis...

.... suggests that age [category] modifies the incidence rate - smoking relationship. Indeed, the highest estimated incidence rate ratio estimate is with the youngest age category. The incidence rate ratio estimates decline with age category.

The [omnibus] test for homogeneity of incidence rate ratios has p-value of 0.0340 also indicating evidence of modification

Model based method

If y is the number of deaths and PY is the corresponding person-years, then the incidence rate is: $\frac{E(y)}{PY}$

One can model $\log \frac{E(y)}{PY} = \log E(y) - \log(PY)$

with:

$$\log E(y) - \log(PY) = \sum_{i=0}^{k} \beta_i x_i$$

$$\log E(y) = \log(PY) + \sum_{i=0}^{k} \beta_i x_i$$

Offset

Notice the addition of the term log(PY) on the right hand side of the equation.

This term does not have a regression coefficient in front of it. Such terms appear in many other situations and are typically called 'offsets'.

Stata adds the offset using the “exposure' option. Using this option, one does not take the logarithm of PY (Stata does this for you)

Log of Expected Number of Deaths

The regression coefficients, here, now involve the log of the incidence rate and accordingly a difference between 2 log incidence rates is the log of the incidence rate ratio.

The distribution for the counts (the deaths, here) is 'typically' taken as the Poisson distribution:

$$f(y) = e^{-\lambda} \frac{\lambda^{y}}{y!} \quad \text{where} \quad E(y) = \lambda$$

The Poisson assumption

Consider the Binomial distribution with a large number trials (n) and small probability (p) with success on any trial. Let the expected value be:

$$\lambda = np \quad \text{so that} \quad p = \lambda / n$$

so that

$$\binom{n}{y} p^y (1-p)^{n-y} = \frac{n^{(y)}}{y!} \left(\frac{\lambda}{n}\right)^y \left(1-\frac{\lambda}{n}\right)^{n-y} = \frac{n^{(y)}}{n^y} \left(1-\frac{\lambda}{n}\right)^{-y} \left(1-\frac{\lambda}{n}\right)^n \frac{\lambda^y}{y!}$$

$$\text{now} \quad \left(1-\frac{\lambda}{n}\right)^n \text{ gets close to } e^{-\lambda} \text{ when n is large ¿}$$

Mean linked to Variance

From the Binomial distribution, we know that

$$Var(y) = np(1-p) = n\frac{\lambda}{n}(1-\frac{\lambda}{n}) = \lambda(1-\frac{\lambda}{n})$$

which will be λ for large n.

So that, for the Poisson distribution, E(y)=Var(y)

Using this distribution to fit a model in E(y) completely determines the distribution.

```
. gen ageg=6-agecat
. poisson deaths smokes##i.ageg, exposure(pyears) irr

Poisson regression                                      Number of obs   =         10
                                                        LR chi2(9)      =     935.07
                                                        Prob > chi2     =     0.0000
Log likelihood =  -27.53397                             Pseudo R2       =     0.9444

------------------------------------------------------------------------------
      deaths |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.smokes |   .9047304   .1855513    -0.49   0.625     .6052658     1.35236
             |
        ageg |
          2  |    .510838   .1331832    -2.58   0.010     .3064517    .8515384
          3  |   .2312638    .060294    -5.62   0.000     .1387352    .3855038
          4  |    .053025   .0180278    -8.64   0.000     .0272322    .1032472
          5  |   .0050198   .0036623    -7.26   0.000     .0012014    .0209745
             |
 smokes#ageg |
       1 2   |   1.498855   .4322128     1.40   0.160      .851737     2.63763
       1 3   |   1.622848   .4664966     1.68   0.092      .923837    2.850758
       1 4   |   2.364032   .8686353     2.34   0.019     1.150508    4.857549
       1 5   |   6.340714   4.801021     2.44   0.015     1.437576    27.96698
      pyears |  (exposure)
------------------------------------------------------------------------------

. disp 6.340714*0.9047304
5.7366367
```

The Modeling Process

All of the techniques developed for modeling can be used here.

Adding the option 'irr' takes the exponent of the coefficients enabling the direct viewing of estimates of incidence rate ratios (and as always) ratios of incidence rate ratios.

A likelihood ratio test is the model based equivalent for the test for homogeneity. The procedure is the same as always.

More distributions for counts

Fitting with the poisson distribution, forces the Variance to be the same as the Mean.

In health research contexts, this often turns out to be a poor distribution choice.

A much more flexible distribution is now seeing considerable use in health research.

The Negative Binomial (or sometimes called the Polya) distribution is now quite easily used.

Negative Binomial

The Negative Binomial distribution comes from the focus of attention on the distribution of the number of failures needed to achieve r successes.

Allowing r to take on any positive number, one will see this distribution called the Polya distribution.

$$E(y) = r\frac{q}{p} \quad \text{and} \quad Var(y) = \frac{rq}{p^2}$$

Dispersion

Dispersion is the ratio of the variance to the mean: $\frac{Var(y)}{E(y)}$

For the Poisson distributions, the dispersion is always one.

For the Negative Binomial distributions, the dispersion is always greater than one.

Constant Dispersion: These models allow for varying (conditional) means but fixed estimable dispersion

With this approach, using nbreg, one models

$$\log(\mu)=\sum_{i=0}^{k}\beta_i x_i \quad \text{and estimates} \quad \delta \quad \text{where} \quad 1+\delta \quad \text{is the dispersion}$$

$$E(y)=\mu \quad \text{and} \quad Var(y)=\mu(1+\delta)$$

$$\text{then} \quad r=\frac{\mu}{\delta} \quad \text{and} \quad p=\frac{1}{1+\delta}$$

$$\text{or} \quad \mu=r\left(\frac{1}{p}-1\right) \quad \text{and} \quad \delta=\frac{1}{p}-1$$

A test for 'overdispersion' is enabled with testing $\delta>0$

Mean Dispersion: These models allow for varying means and varying dispersion dependent on the mean

With this approach, using nbreg, we model:

$$\log(\mu)=\sum_{i=0}^{k}\beta_i x_i \quad \text{and estimate } \alpha \quad \text{where } \quad 1+\alpha\mu \quad \text{is the dispersion}$$

$$E(y)=\mu \quad \text{and} \quad Var(y)=\mu(1+\alpha\mu)$$

$$\text{then} \quad r=\frac{1}{\alpha} \quad \text{and} \quad p=\frac{1}{1+\alpha\mu}$$

$$\text{or} \quad \mu=r\left(\frac{1}{p}-1\right) \quad \text{and} \quad \alpha=\frac{1}{r}$$

Notice here that: $\log(\alpha)=-\log(r)$

Now, testing $\alpha>0$ is testing for 'overdispersion'

Generalized Mean Dispersion

For this approach, using gnbreg, one models both:

$$\log(\mu)=\sum_{i=0}^{k}\beta_i x_i \quad \text{and} \quad \log(\alpha)=\sum_{j=0}^{l}\gamma_j z_j$$

$$\text{where} \quad 1+\alpha\mu \quad \text{is the dispersion}$$

Here, the varying dispersion may be modelled by a (possibly different) set of 'covariates'.

At this time, there is no test for overdispersion available with gnbreg

It is not uncommon to posit a Poisson regression model and observe a lack of model fit. The following data appeared in Rodrıguez (1993):

```
. use http://www.statapress.com/data/r11/rod93
```

```
. list

     +---------------------------------------+
     | cohort   age_mos   deaths   exposure |
     |---------------------------------------|
  1. |      1       0.5      168      278.4 |
  2. |      1       2.0       48      538.8 |
  3. |      1       4.5       63      794.4 |
  4. |      1       9.0       89    1,550.8 |
  5. |      1      18.0      102    3,006.0 |
     |---------------------------------------|
  6. |      1      42.0       81    8,743.5 |
  7. |      1      90.0       40   14,270.0 |
  8. |      2       0.5      197      403.2 |
  9. |      2       2.0       48      786.0 |
 10. |      2       4.5       62    1,165.3 |
     |---------------------------------------|
 11. |      2       9.0       81    2,294.8 |
 12. |      2      18.0       97    4,500.5 |
 13. |      2      42.0      103   13,201.5 |
 14. |      2      90.0       39   19,525.0 |
 15. |      3       0.5      195      495.3 |
     |---------------------------------------|
 16. |      3       2.0       55      956.7 |
 17. |      3       4.5       58    1,381.4 |
 18. |      3       9.0       85    2,604.5 |
 19. |      3      18.0       87    4,618.5 |
 20. |      3      42.0       70    9,814.5 |
     |---------------------------------------|
 21. |      3      90.0       10    5,802.5 |
     +---------------------------------------+
```

Very briefly, to start out: Stata examples :

```
. poisson deaths i.cohort, exposure(exposure)
. estat gof
. nbreg deaths i.cohort,exposure(exposure)
. nbreg deaths i.cohort,exposure(exposure) dispersion(constant)
. gnbreg deaths age_mos, lnalpha(i.cohort) exposure(exposure)
. test 2.cohort 3.cohort
```

UCLA Website

Another example with annotation can be found at:

http://www.ats.ucla.edu/stat/stata/output/stata_nbreg_output.htm

Caveat: The terminology used here is based on the conventions adopted by Stata. Other software systems, notably R and SAS, use different conventions and definitions for negative binomial modelling

Zero Counts : Inflation or Truncation

Sometimes the zero counts need special consideration

Inflation: There can be two regimes : one regime that contributes only zeroes and another regime that may be handled by the Poisson or the Negative Binomial [number of surfaces with tooth decay]

Truncation: A zero count may be excluded [days in the ICU]

Zero Inflation

We can model each regime with explanatory variables [with commands : zip or zinb]

We can assess whether there is zero inflation [with the Vuong test]

Indications of zero inflation can be seen with:

plausible regimes from the context

very high zero counts and sometimes overdispersion

Prenatal visit data

the data is in bw5k.dta

the rate of visits per week is to be modelled with:

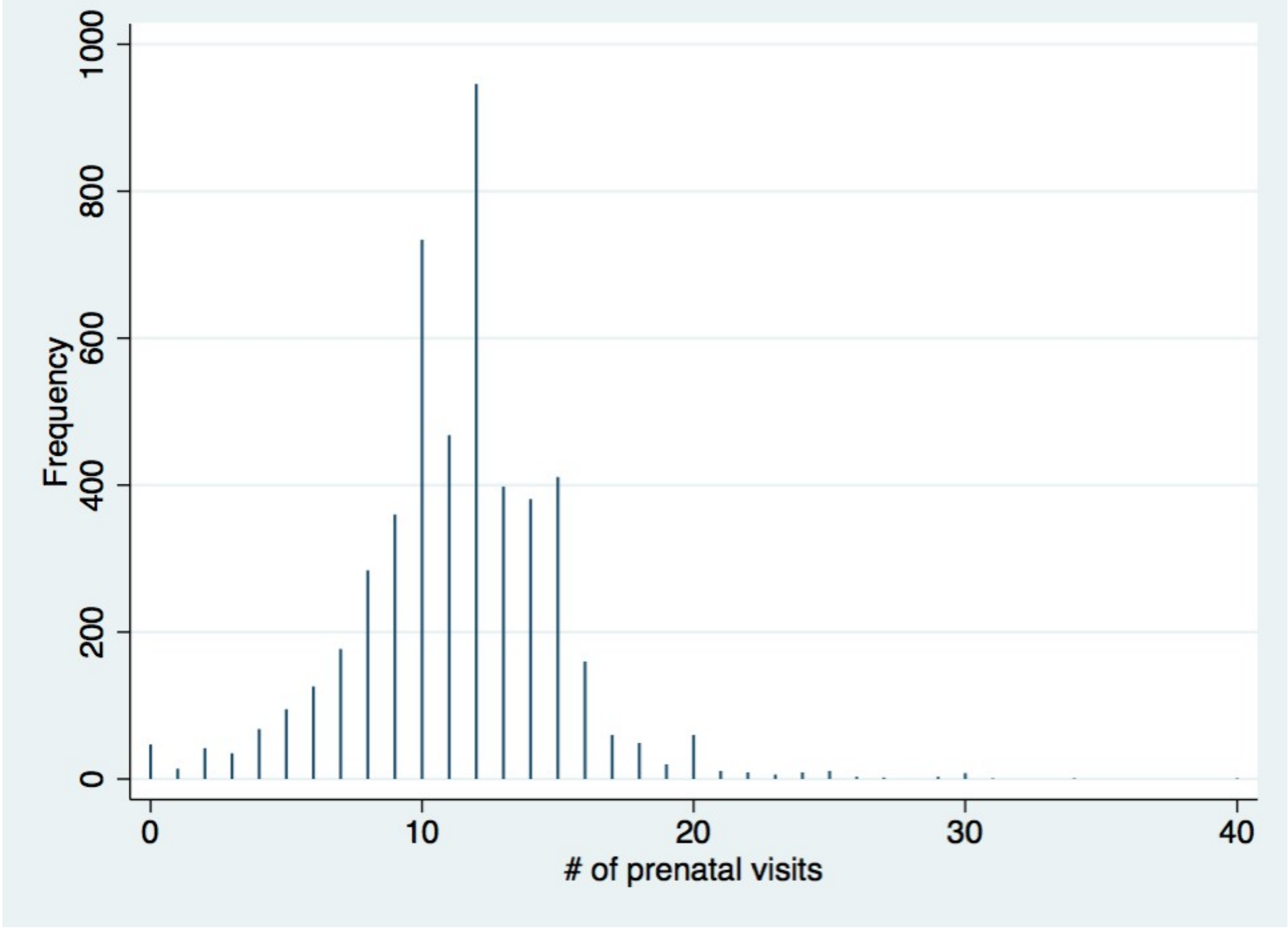
white : mother's race (0=no , 1=yes)

mage_28 : mother's age [centred at 28 yrs)

tbo_1 : parity (0=1st , 1=2nd ...)

mother's education

```
use bw5k.dta
// from www.cdc.gov/nchs
gen mage_28=mage-28
gen tbo_1=tbo-1
gen white=(mrace_c3==2)
spikeplot previs
// notice the 'bump' at zero visits
zinb  previs white mage_28 tbo_1 i.meduc_c4, inflate(i.meduc_c4) exposure(gest) vuong
// following Dohoo, Martin & Stryhn
```


Frequency
0
200
400
600
800
1000
0
10
20
30
40
# of prenatal visits

```
Zero-inflated negative binomial regression        Number of obs   =       5000
                                                  Nonzero obs     =       4953
                                                  Zero obs        =         47

Inflation model = logit                           LR chi2(6)      =     103.49
Log likelihood  = -13560.19                       Prob > chi2     =     0.0000

------------------------------------------------------------------------------
      previs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
previs       |
       white |    .031533   .0095333     3.31   0.001      .012848     .050218
     mage_28 |   .0037196   .0009207     4.04   0.000     .0019151     .005524
       tbo_1 |  -.0120256   .0034051    -3.53   0.000    -.0186995   -.0053517
             |
    meduc_c4 |
     hs dip  |   .0440167   .0141647     3.11   0.002     .0162544     .071779
   some col  |   .0632298   .0150188     4.21   0.000     .0337936    .0926661
   univ deg  |   .0618504   .0150638     4.11   0.000      .032326    .0913748
             |
       _cons |  -1.265611   .0130318   -97.12   0.000    -1.291153   -1.240069
    ln(gest) |          1  (exposure)
-------------+----------------------------------------------------------------
inflate      |
    meduc_c4 |
     hs dip  |  -1.097442   .4259315    -2.58   0.010    -1.932252   -.2626317
   some col  |  -1.048664   .4458699    -2.35   0.019    -1.922553   -.1747748
   univ deg  |   -1.04811   .3650916    -2.87   0.004    -1.763676   -.3325433
             |
       _cons |  -3.909495   .2323357   -16.83   0.000    -4.364865   -3.454125
-------------+----------------------------------------------------------------
    /lnalpha |  -4.700879   .2119302   -22.18   0.000    -5.116254   -4.285503
-------------+----------------------------------------------------------------
       alpha |   .0090873   .0019259                     .0059984    .0137667
------------------------------------------------------------------------------
Vuong test of zinb vs. standard negative binomial: z =     5.77  Pr>z = 0.0000
```

A recent use of Zero Inflated methods

Measuring the short-term impact of fluoridation cessation on dental caries in Grade 2 children using tooth surface indices [2016]

Lindsay McLaren et al

Community Dentistry and Oral Epidemiology

Truncation

If a zero count for the outcome [y] is not possible, then one analysis option is to subtract one from the outcome

Then, y-1 might be assumed Poisson or Negative Binomial [poisson or nbreg]

You could then add one back to the confidence intervals for E(y) = E(y-1)+1

You can use tpoisson or ztnb to analyze y directly [the distributional assumption is not the same]

A study of length of hospital stay

```
use hospital_stay.dta
tab stay
gen staym1=stay-1
nbreg staym1 age hmo died
ztnb stay age hmo died
```

The analyses are quite similar here. The choice might be dictated by your literature review.