Notes for Session 11

The Poisson distribution

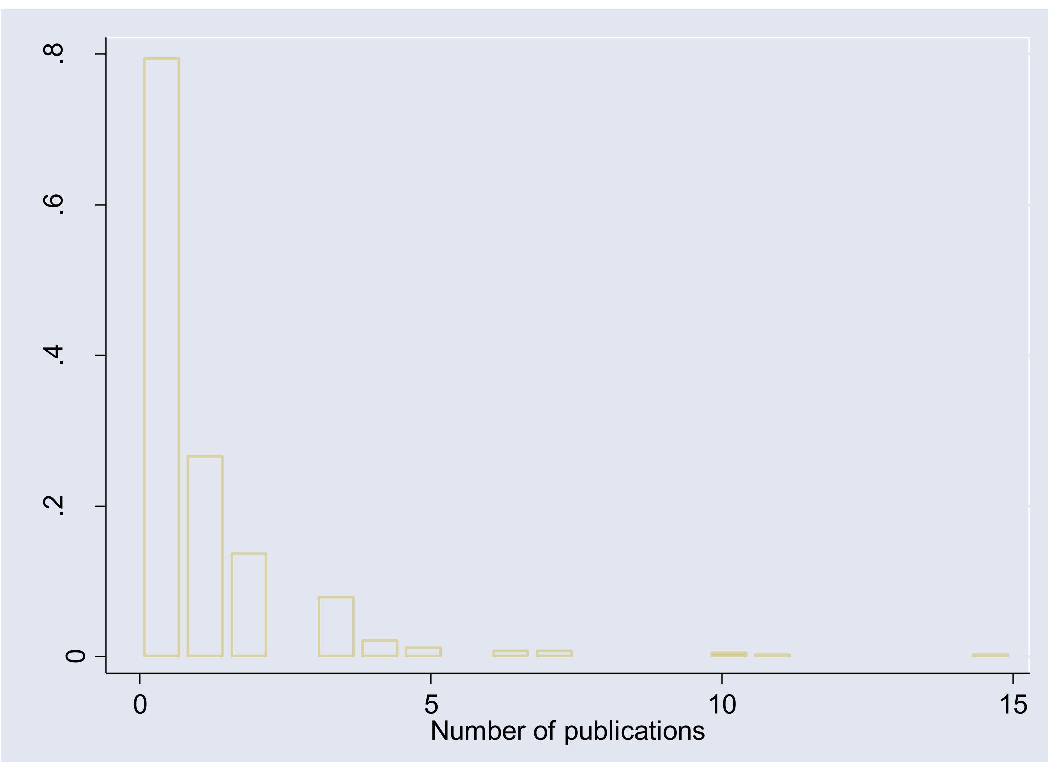
Scientific productivity example (McGinnis, Allison and Long, 1982, Allison, 1999)

An example of a data set that can be analyzed by Poisson methods is as follows: 557 male biochemists received their doctoral degree from 106 American universities in the late 1950s and 1960s.

PDOC	1 if received postdoctoral training, 0 otherwise
AGE	Age in years at completion of Ph.D.
MAR	1 if married, 0 otherwise
DOC	Measure of the prestige of the doctoral institution
UND	Measure of the selectivity of the undergraduate institution
AG	1 if degree is from an agricultural department, 0 otherwise
ARTS	Number of articles published while a graduate student
CITS	Number of citations to published articles
DOCID	ID number of the doctoral institution

The frequency distribution of the number of publications is given as follows:

```
. hist arts, xlab() ylab()  bin(20) gap(20)
```



The goodness of fit test for a Poisson distribution however, is highly significant (i.e., does not support a Poisson-distributed variable). Notice that you must run a Poisson model before `poisgof`.

```
. quietly poisson arts
. poisgof
        Goodness of fit chi-2 =  1087.821
        Prob > chi2(556)      =    0.0000
```

Analysis with a Poisson GLM

In the case of the Poisson mean, because λ is always positive, the function $g(.)$ is chosen so that the linear predictor $\eta = \beta_o + \beta_1 X_1 + \cdots + \beta_1 X$, that can take any real-number value, gets mapped into the positive real numbers. A good candidate function (link) for the Poisson GLM is the logarithm as follows:

$$\log(\lambda) = \beta_o + \beta_1 X_1 + \cdots + \beta_p X_p = \eta$$

We carry out the Poisson regression using either the `poisson` or `glm` command in STATA. Here we prefer the `glm` command, because it produces the deviance that will be useful in the following.

```
. xi: glm   arts age i.mar doc  und i.ag , nolog fam(poisson)
i.mar             _Imar_0-1           (naturally coded; _Imar_0 omitted)
i.ag              _Iag_0-1            (naturally coded; _Iag_0 omitted)

Generalized linear models                          No. of obs      =       557
Optimization     : ML: Newton-Raphson              Residual df     =       551
                                                   Scale parameter =         1
Deviance         =  1078.905935                    (1/df) Deviance =  1.958087
Pearson          =   1497.36098                    (1/df) Pearson  =  2.717534

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

Log likelihood   =  -817.464978                    AIC             =  2.956786
BIC              = -2404.827512

------------------------------------------------------------------------------
        arts |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0165613   .0101663    -1.63   0.103    -.0364868    .0033642
     _Imar_1 |  -.0153611   .1300267    -0.12   0.906    -.2702088    .2394865
         doc |  -.0000399   .0004551    -0.09   0.930    -.0009319    .0008521
         und |   .0723311   .0303235     2.39   0.017      .012898    .1317641
      _Iag_1 |   .0421593    .099889     0.42   0.673    -.1536195    .2379381
       _cons |  -.0401208   .3897092    -0.10   0.918    -.8039369    .7236953
------------------------------------------------------------------------------
```

Interpretation of the coefficients

The coefficients $\beta_1,\cdots,\beta_p$ denote the change in log(λ) for each one-unit change in the corresponding explanatory variable. In our example, the only significant variable is UND, the selectivity index of the under graduate institution. So, if two observations *i* and *j* have a difference of one unit in explanatory variable X_4 (UND), that is $X_{4i} - X_{4j} = 1$, while all the other explanatory variables are the same, then the difference in log(λ) will be β_4.

a. Given this information please calculate the impact of (thus interpret) the coefficients that were produced by the model.

Overdispersion

By the assumptions of the Poisson model, the expected value (mean) of the Poisson distribution is theoretically equal to its variance. Frequently this is not the case and the variance is much higher than the mean. In that situation, we have what is called *overdispersion*. In this case, the scaled deviance value of 1.96 and scaled Pearson chi-square of 2.72 point to a potential problem with the model.

One way to deal with overdispersion is to divide the chi-square statistic that tests the significance of each variable by the scaled deviance or scaled Pearson chi-square (or equivalently multiply each standard error by the square root of the scaled deviance or scaled Pearson chi-square; Agresti, 1996).

To carry out the method suggested in Agresti (1996) by STATA we proceed as follows:

We divide each *z* statistic in the output above by the square root of the scaled Pearson chi-square statistic and re-calculate its significance. Here we go:

- `age`

```
. di 2*norm(-1.629/sqrt(2.717532))
.32306707
```

b. Perform the calculations for the remaining coefficients.

We can do this a lot more simply by the following command:

```
. xi: glm   arts age i.mar doc  und i.ag , nolog fam(poisson) scale(x2)
i.mar             _Imar_0-1           (naturally coded; _Imar_0 omitted)
i.ag              _Iag_0-1            (naturally coded; _Iag_0 omitted)

Generalized linear models                          No. of obs      =       557
Optimization     : ML: Newton-Raphson              Residual df     =       551
                                                   Scale parameter =         1
Deviance         =  1078.905935                    (1/df) Deviance =  1.958087
Pearson          =   1497.36098                    (1/df) Pearson  =  2.717534

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

Log likelihood   =  -817.464978                    AIC             =  2.956786
BIC              = -2404.827512

------------------------------------------------------------------------------
        arts |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0165613    .016759    -0.99   0.323    -.0494084    .0162858
     _Imar_1 |  -.0153611   .2143483    -0.07   0.943    -.4354761    .4047538
         doc |  -.0000399   .0007502    -0.05   0.958    -.0015103    .0014305
         und |   .0723311   .0499882     1.45   0.148    -.0256439    .1703061
      _Iag_1 |   .0421593   .1646664     0.26   0.798    -.2805809    .3648995
       _cons |  -.0401208   .6424335    -0.06   0.950    -1.299267    1.219026
------------------------------------------------------------------------------
(Standard errors scaled using square root of Pearson X2-based dispersion)

.
    arts |      Coef.   Std. Err.       z      P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |  -.0165613    .016759     -0.988   0.323      -.0494084    .0162858
  Imar_1 |  -.0153611   .2143482     -0.072   0.943      -.4354759    .4047536
     doc |  -.0000399   .0007502     -0.053   0.958      -.0015103    .0014305
     und |   .0723311   .0499881      1.447   0.148      -.0256439     .170306
   Iag_1 |   .0421593   .1646663      0.256   0.798      -.2805808    .3648994
   _cons |  -.0401209    .642433     -0.062   0.950      -1.299266    1.219025
------------------------------------------------------------------------------
(Standard errors scaled using square root of Pearson X2-based dispersion)
```

The results are identical to the calculations above.

Accounting for overdispersion: The Negative Binomial distribution

The negative binomial model is fit in STATA either by the `nbreg` command, or the `glm` command by specifying `family(nbinom)` as the family of distributions. The default is a log link.

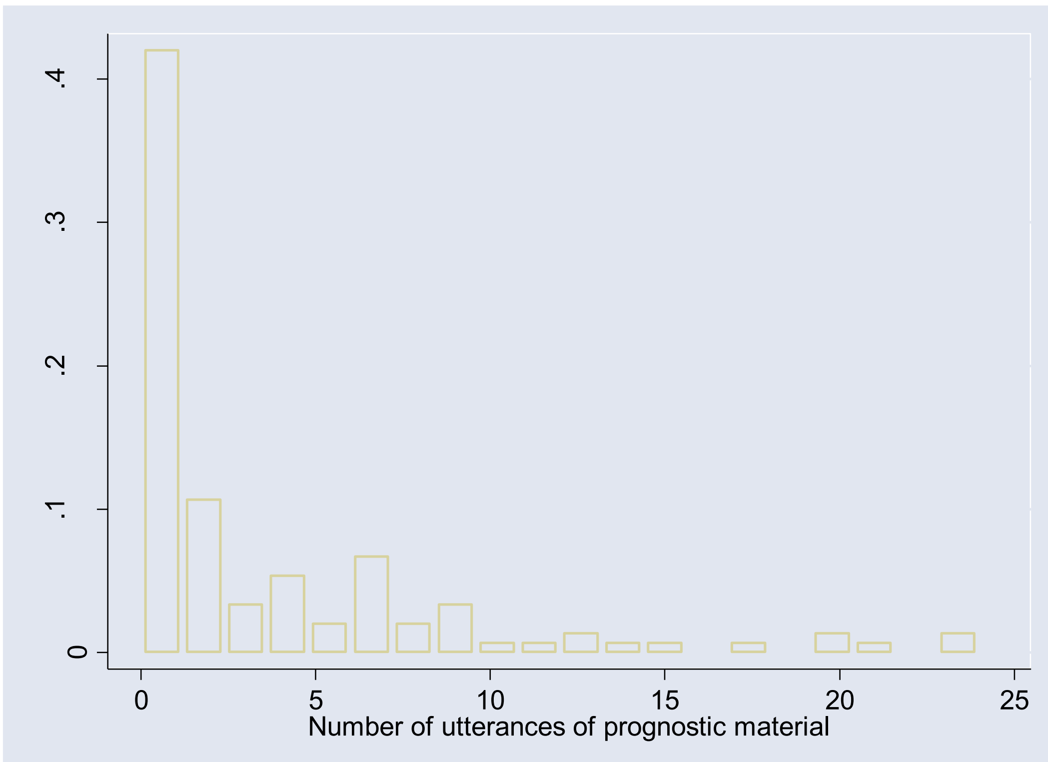
```
. xi: glm  arts  age i.mar doc und i.ag , family(nbinom) nolog
i.mar             _Imar_0-1           (naturally coded; _Imar_0 omitted)
i.ag              _Iag_0-1            (naturally coded; _Iag_0 omitted)

Generalized linear models                          No. of obs      =       557
Optimization     : ML: Newton-Raphson              Residual df     =       551
                                                   Scale parameter =         1
Deviance         =  602.3390862                    (1/df) Deviance =  1.093174
Pearson          =  805.4735374                    (1/df) Pearson  =  1.461839

Variance function: V(u) = u+(1)u^2                 [Neg. Binomial]
Link function    : g(u) = ln(u)                    [Log]
Standard errors  : OIM

Log likelihood   = -705.6949434                    AIC             =  2.555458
BIC              = -2881.394361

------------------------------------------------------------------------------
        arts |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0179105    .014638    -1.22   0.221    -.0466004    .0107794
     _Imar_1 |  -.0082171    .179777    -0.05   0.964    -.3605735    .3441394
         doc |   .0000457   .0006079     0.08   0.940    -.0011458    .0012373
         und |   .0709433   .0404426     1.75   0.079    -.0083228    .1502094
      _Iag_1 |   .0463118   .1362171     0.34   0.734    -.2206688    .3132925
       _cons |  -.0272584     .54196    -0.05   0.960     -1.08948    1.034964
------------------------------------------------------------------------------
```

The coefficients are similar to those generated by the poisson regression model, and the dispersion value is a great deal closer to 1.0. The undergraduate selectivity index is significant at the 10% level but not the 5% level in this analysis. No other factors are significant.

Number of utterances about prognosis

The data set of Christakis and Levinson, 1998) describes the analysis of the number of utterances concerning prognosis by a doctor during a patient visit. The relevant variables and information were given in the lecture. The dataset is `prognosis.dta.`
The frequency distribution of the LENGTHPX variable is given below:

```
. hist  lengthpx, bin(20) xlab() ylab()  gap(20)
```




```
. quietly poisson  lengthpx

. poisgof

         Goodness of fit chi-2 =  793.6129
         Prob > chi2(124)      =    0.0000
```

We see that the data are highly skewed with a substantial proportion of observations at zero. The goodness of fit test however, is significant, implying that the marginal (i.e., without considering the explanatory variables) distrbution of `lengthpx` is not Poisson.

Offset

The way we incorporate the length of observation (duration of visit) is by adding what is called an "offset" variable to the model, that is, $\log E(Y_i) = \log(t_i) + \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$. This is done by adding the option `offset(varname)` or `lnoffset(varname)` in the `glm` command. The latter is what we need if the variable has not been transformed to the logarithmic scale already. The results of the analysis for these data are as follows:

```
. xi: glm  lengthpx ptage i.ptsex  ezcompt mdlikept i.surgeon claims, family(po
> isson) lnoffset( minutes) nolog
i.ptsex           _Iptsex_0-1         (naturally coded; _Iptsex_0 omitted)
i.surgeon         _Isurgeon_0-1       (naturally coded; _Isurgeon_0 omitted)

Generalized linear models                          No. of obs      =       121
Optimization     : ML: Newton-Raphson              Residual df     =       114
                                                   Scale parameter =         1
Deviance         =  682.0299077                    (1/df) Deviance =  5.982718
Pearson          =  899.6256873                    (1/df) Pearson  =  7.891453

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

Log likelihood   = -455.9056163                    AIC             =  7.651333
BIC              =  135.3097855

------------------------------------------------------------------------------
    lengthpx |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ptage |  -.0014421   .0030592    -0.47   0.637    -.0074381    .0045538
   _Iptsex_1 |   .5482447   .1048295     5.23   0.000     .3427827    .7537067
     ezcompt |     .19809   .0760462     2.60   0.009     .0490422    .3471378
    mdlikept |  -.0864474    .074387    -1.16   0.245    -.2322432    .0593484
 _Isurgeon_1 |   1.343119   .1303695    10.30   0.000     1.087599    1.598638
      claims |   .0519112   .0231909     2.24   0.025     .0064579    .0973645
       _cons |  -3.175498   .3188584    -9.96   0.000    -3.800449   -2.550547
     minutes | (exposure)
------------------------------------------------------------------------------
```

c. What kind of offset did we add to this model? Why isn't there any coefficient associated with variable `minutes`?

Interpretation of the analysis results

Almost all variables are significant. It seems that there are 73% more utterances about prognosis when the subject is male ($e^{0.548} - 1 = 0.73$).

d. In a similar fashion with the operations above, interpret the meaning of the rest of the coefficients.

However, the results of this analysis are questionable, as the scaled Pearson chi-square and scaled deviance statistics are much larger than 1.0.

Thus, significant overdispersion is likely present in these data.

Correcting for overdispersion

To correct for overdispersion, we scale the test statistics corresponding to the coefficients by the scaled Pearson chi-square statistic. Only `surgeon` is significant in predicting prognosis utterances.

```
. xi: glm  lengthpx ptage i.ptsex  ezcompt mdlikept i.surgeon claims, family(po
> isson) lnoffset( minutes) nolog scale(x2)
i.ptsex           _Iptsex_0-1         (naturally coded; _Iptsex_0 omitted)
i.surgeon         _Isurgeon_0-1       (naturally coded; _Isurgeon_0 omitted)

Generalized linear models                          No. of obs      =       121
Optimization     : ML: Newton-Raphson              Residual df     =       114
                                                   Scale parameter =         1
Deviance         =  682.0299077                    (1/df) Deviance =  5.982718
Pearson          =  899.6256873                    (1/df) Pearson  =  7.891453

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

Log likelihood   = -455.9056163                    AIC             =  7.651333
BIC              =  135.3097855

------------------------------------------------------------------------------
    lengthpx |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ptage |  -.0014421   .0085938    -0.17   0.867    -.0182857    .0154014
   _Iptsex_1 |   .5482447   .2944842     1.86   0.063    -.0289337    1.125423
     ezcompt |     .19809   .2136269     0.93   0.354    -.2206111    .6167911
    mdlikept |  -.0864474   .2089659    -0.41   0.679     -.496013    .3231182
 _Isurgeon_1 |   1.343119   .3662305     3.67   0.000     .6253199    2.060917
      claims |   .0519112   .0651472     0.80   0.426     -.075775    .1795973
       _cons |  -3.175498   .8957284    -3.55   0.000    -4.931093   -1.419902
     minutes | (exposure)
------------------------------------------------------------------------------
(Standard errors scaled using square root of Pearson X2-based dispersion)
```

Correcting for overdispersion by negative-binomial regression

The previous analysis may be inefficient, so we also undertake a negative binomial regression analysis.

```
. xi: glm  lengthpx ptage i.ptsex  ezcompt mdlikept i.surgeon claims, family(nb
> inom) lnoffset( minutes) nolog
i.ptsex           _Iptsex_0-1         (naturally coded; _Iptsex_0 omitted)
i.surgeon         _Isurgeon_0-1       (naturally coded; _Isurgeon_0 omitted)

Generalized linear models                          No. of obs      =       121
Optimization     : ML: Newton-Raphson              Residual df     =       114
                                                   Scale parameter =         1
Deviance         =  197.2955808                    (1/df) Deviance =  1.730663
Pearson          =  203.1605871                    (1/df) Pearson  =   1.78211

Variance function: V(u) = u+(1)u^2                 [Neg. Binomial]
Link function    : g(u) = ln(u)                    [Log]
Standard errors  : OIM

Log likelihood   =  -276.385523                    AIC             =  4.684058
BIC              = -349.4245414

------------------------------------------------------------------------------
    lengthpx |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ptage |   .0002238   .0078626     0.03   0.977    -.0151867    .0156342
   _Iptsex_1 |   .5829174    .224159     2.60   0.009     .1435738    1.022261
     ezcompt |    .129117   .1464301     0.88   0.378    -.1578809    .4161148
    mdlikept |     -.1062    .153993    -0.69   0.490    -.4080207    .1956208
 _Isurgeon_1 |    1.40706   .2687397     5.24   0.000     .8803401    1.933781
      claims |   .0514741    .055639     0.93   0.355    -.0575763    .1605245
       _cons |  -2.758854   .7874139    -3.50   0.000    -4.302157   -1.215551
     minutes | (exposure)
------------------------------------------------------------------------------
```

e. Interpret the results from this output and compare with the results of the unadjusted and adjusted poisson regressions that were undertaken originally?