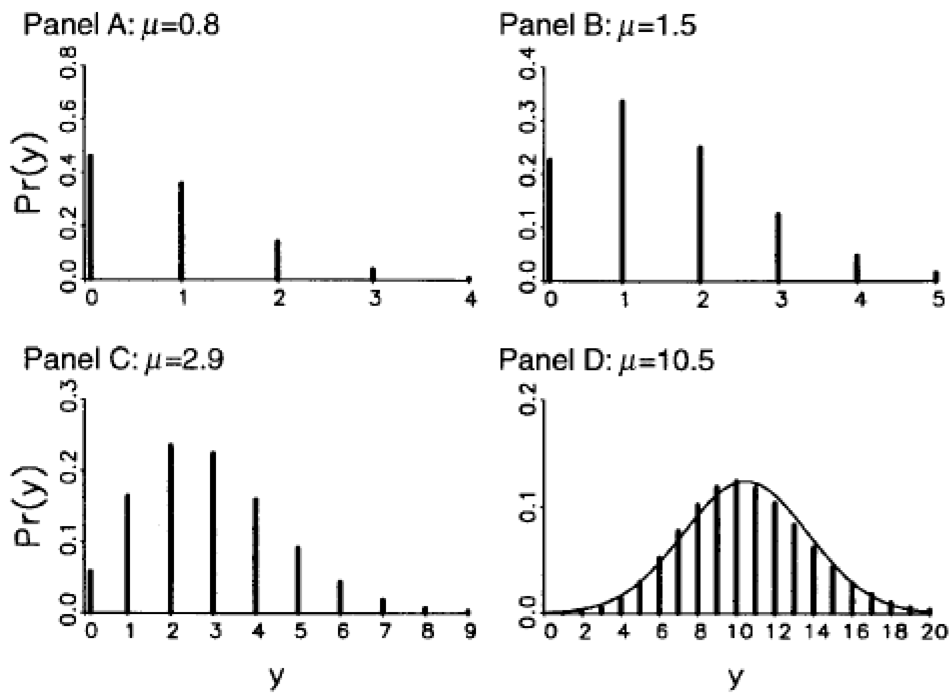
SOCY7704: Regression Models for Categorical Data
Instructor: Natasha Sarkisian

Poisson Regression

Count variables are often treated as though they are continuous, and OLS is used. OLS in this case can result in inefficient, inconsistent, and biased estimates. Need to use models that are developed specifically for count data. Poisson model is the most basic of them.

Poisson distributions:

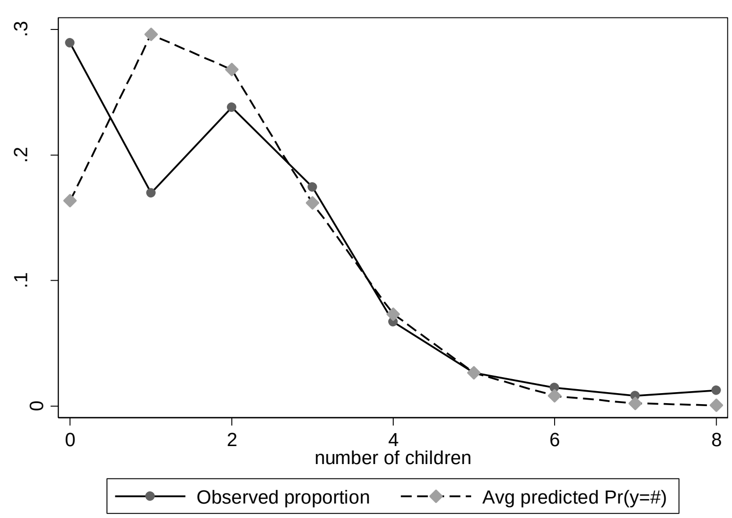


Characteristics of Poisson distribution:

1. E(y) = μ
2. The variance equals the mean: Var(y)=E(y)= μ -- equidispersion. In practice, the variance is often larger than μ: this is called overdispersion. The main reason for overdispersion is heterogeneity – if there are different groups within data that have different means and all of them are actually equal to their variances, when you put all of these groups together, the resulting combination will have variance larger than the mean. Therefore, we need to control for all those sources of heterogeneity. Thus, when using Poisson regression, we need to ensure that the conditional variance equals to the mean – that is Var(y|X)=E(y|X).
3. As μ increases, the probability of zeros decreases. But for many count variables, there are more observed zeros than would be predicted from Poisson distribution
4. As μ increases, the Poisson distribution approximates normal.
5. The assumption of independence of events – past outcomes don't affect future outcomes.

We usually start by examining the raw distribution and comparing it with Poisson:

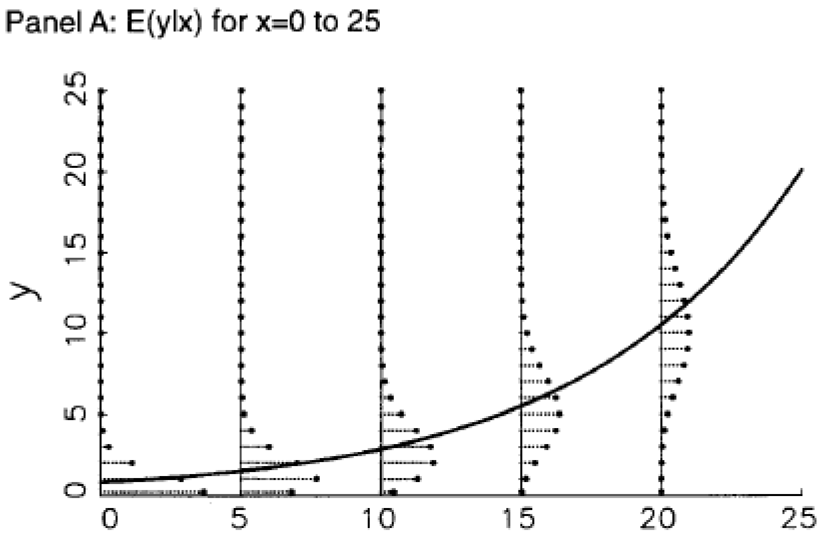
```
. tab childs
    number of |
     children |      Freq.     Percent        Cum.
--------------+-----------------------------------
         none |        799       28.95       28.95
          one |        469       16.99       45.94
          two |        657       23.80       69.75
        three |        481       17.43       87.17
         four |        185        6.70       93.88
         five |         73        2.64       96.52
          six |         40        1.45       97.97
        seven |         22        0.80       98.77
eight or more |         34        1.23      100.00
--------------+-----------------------------------
        Total |      2,760      100.00

. poisson childs
Iteration 0:   log likelihood = -5096.6865
Iteration 1:   log likelihood = -5096.6865
Poisson regression                                Number of obs   =       2760
                                                  LR chi2(0)      =       0.00
                                                  Prob > chi2     =          .
Log likelihood = -5096.6865                       Pseudo R2       =     0.0000
------------------------------------------------------------------------------
      childs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .5936071   .0141464    41.96   0.000     .5658807    .6213334
------------------------------------------------------------------------------

. mgen, pr(0/8) meanpred stub(poi_)
Predictions from:
Variable   Obs Unique      Mean        Min       Max  Label
----------------------------------------------------------------------------------------
poi_val      9      9         4          0         8  number of children
poi_obeq     9      9  .1111111    .007971  .2894928  Observed proportion
poi_oble     9      9  .7988325   .2894928         1  Observed cum. proportion
poi_preq     9      9  .1110984   .0004684  .2961468  Avg predicted Pr(y=#)
poi_prle     9      9  .7988352   .1635711  .9998854  Avg predicted cum. Pr(y=#)
poi_ob_pr    9      9  .0000127  -.1262192  .1259216  Observed - Avg Pr(y=#)
----------------------------------------------------------------------------------------

. graph twoway connected poi_obeq poi_preq poi_val
```



Overdispersion results in Poisson distribution underpredicting the outcomes in the two ends of the distribution – it underpredicts zeros and outcomes of 6 and larger. Fitting this kind of unconditional Poisson distribution does not take heterogeneity into account – the average number of children varies according to some characteristics of respondents. Next, we have to allow for that – need to incorporate the observed heterogeneity. A multivariate Poisson regression model does just that. It models the average count, μ:

$\mu=E(y|x)=\exp(Xb)$

We exponentiate to force the values to be positive–counts cannot be below 0. We get a nonlinear model that looks like this:



Let's run a multivariate Poisson model:

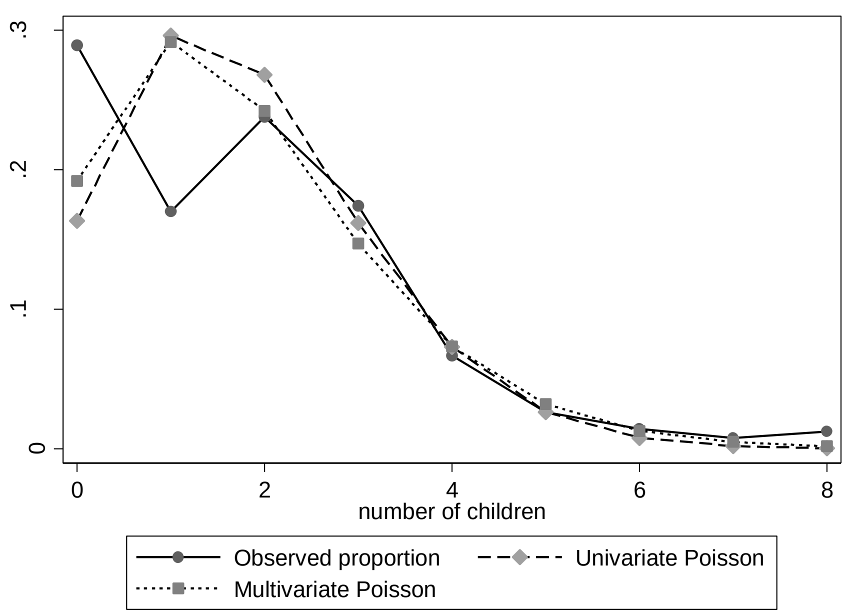
```
. poisson childs sex married sibs  born educ

Iteration 0:   log likelihood = -4784.5123
Iteration 1:   log likelihood = -4784.5079
Iteration 2:   log likelihood = -4784.5079

Poisson regression                                Number of obs   =       2745
                                                  LR chi2(5)      =     572.66
                                                  Prob > chi2     =     0.0000
Log likelihood = -4784.5079                       Pseudo R2       =     0.0565

------------------------------------------------------------------------------
      childs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |    .195229   .0289993     6.73   0.000     .1383915    .2520665
     married |   .4486183   .0288777    15.54   0.000      .392019    .5052176
        sibs |   .0385556    .004219     9.14   0.000     .0302865    .0468246
        born |  -.2209195   .0522438    -4.23   0.000    -.3233154   -.1185235
        educ |   -.061697   .0048163   -12.81   0.000    -.0711369   -.0522572
       _cons |   .9547179   .1010692     9.45   0.000     .7566258     1.15281
------------------------------------------------------------------------------
```

Can interpret sign and significance – to interpret the size, we will exponentiate the coefficients – generating so-called incidence-rate ratios (comparable to odds ratios). But we'll return to that later.

Model fit, hypothesis testing and model comparisons

Once again, to assess how well our model predicts counts, we can graphically examine the predicted probabilities for different counts (these are probabilities for someone average on all characteristics):

```
. mgen, pr(0/8) meanpred stub(mpoi_)
Predictions from:
Variable    Obs Unique      Mean        Min       Max  Label
----------------------------------------------------------------------------------------
mpoi_val      9      9         4          0         8  number of children
mpoi_obeq     9      9  .1111111   .0080146  .2892532  Observed proportion
mpoi_oble     9      9  .7987047   .2892532         1  Observed cum. proportion
mpoi_preq     9      9   .110982   .0018631  .2918259  Avg predicted Pr(y=#)
mpoi_prle     9      9  .7987926    .192048  .9988381  Avg predicted cum. Pr(y=#)
mpoi_ob_pr    9      9  .0001291  -.1216984  .0972052  Observed - Avg Pr(y=#)
----------------------------------------------------------------------------------------

. lab var mpoi_preq "Multivariate Poisson"

. graph twoway connected poi_obeq poi_preq mpoi_preq poi_val, ylabel(0 (.1) .3)
ytitle("Probability of Count")
```



Multivariate Poisson offers a slight improvement over univariate Poisson – it explains some heterogeneity. But it still doesn't fit very well – underpredicts zeros, overpredicts ones, etc.

Just to clarify this, we can also obtain the probabilities presented in this graph using mtable:

```
. mtable, pr(0/8)
Expression: Pr(childs), predict(pr())
     none       one       two     three      four      five       six     seven  eight_or_more
----------------------------------------------------------------------------------------------
    0.192     0.292     0.242     0.147     0.073     0.032     0.013     0.005          0.002
Specified values where .n indicates no values specified with at()
          |  No at()
 ----------+---------
   Current |       .n
```

So we examined model fit graphically. We can also obtain a goodness-of-fit test (there are two versions of it, one based on deviance residuals, one is based on Pearson residuals; they usually produce similar results):

```
. estat gof

         Deviance goodness-of-fit =  4279.437
         Prob > chi2(2739)        =    0.0000

         Pearson goodness-of-fit  =   3943.17
         Prob > chi2(2739)        =    0.0000
```

Since the probability is below .05, this suggests that predicted counts are significantly different from the observed ones, and therefore Poisson model doesn't fit well. We will deal with that later.

In addition to this, we have all the same tools for hypothesis tests and model comparisons – we can use estat ic after poisson to get information criteria and use BIC comparisons to compare models, especially non-nested ones; we can also use lrtest to compare nested models. And we can use test command to get Wald tests for specific hypotheses (e.g., if deciding whether to combine categories of dummies).

Interpretation of Poisson models

A. *Incidence rate ratios:*

First, as mentioned above, we can calculate incidence rate ratios:

```
. poisson childs sex married sibs  born educ, irr
Poisson regression                                Number of obs   =       2745
                                                  LR chi2(5)      =     572.66
                                                  Prob > chi2     =     0.0000
Log likelihood = -4784.5079                       Pseudo R2       =     0.0565
------------------------------------------------------------------------------
      childs |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   1.215589   .0352512     6.73   0.000     1.148425    1.286682
     married |   1.566147   .0452267    15.54   0.000     1.479966    1.657346
        sibs |   1.039308   .0043848     9.14   0.000      1.03075    1.047938
        born |   .8017812   .0418881    -4.23   0.000     .7237455    .8882309
        educ |   .9401677   .0045282   -12.81   0.000     .9313344    .9490847
------------------------------------------------------------------------------
```

So the number of children for women is 1.22 times (or 22%) higher than the number for men, the number of children for married is 1.57 times (or 57%) higher than for those not currently married, each additional sibling increases the number of children by almost 4%, being foreign born decreases the number of children by almost 10%, and each year of education reduces the number of children by 6%.

We can also obtain incidence rate ratios using listcoef – this will also allow us to see standardized ratios describing the change per one standard deviation of each variable.

```
. listcoef
poisson (N=2745): Factor Change in Expected Count
 Observed SD: 1.6887584
----------------------------------------------------------------------
      childs |      b         z     P>|z|    e^b    e^bStdX      SDofX
-------------+--------------------------------------------------------
         sex |   0.19523    6.732   0.000   1.2156   1.1019     0.4970
```

```
     married |   0.44862   15.535   0.000   1.5661   1.2506     0.4985
        sibs |   0.03856    9.139   0.000   1.0393   1.1227     3.0008
        born |  -0.22092   -4.229   0.000   0.8018   0.9381     0.2893
        educ |  -0.06170  -12.810   0.000   0.9402   0.8324     2.9741
----------------------------------------------------------------------
```

And we can get these as percents:

```
. listcoef, percent
poisson (N=2745): Percentage Change in Expected Count
 Observed SD: 1.6887584
----------------------------------------------------------------------
      childs |      b         z     P>|z|      %      %StdX      SDofX
-------------+--------------------------------------------------------
         sex |   0.19523    6.732   0.000     21.6     10.2     0.4970
     married |   0.44862   15.535   0.000     56.6     25.1     0.4985
        sibs |   0.03856    9.139   0.000      3.9     12.3     3.0008
        born |  -0.22092   -4.229   0.000    -19.8     -6.2     0.2893
        educ |  -0.06170  -12.810   0.000     -6.0    -16.8     2.9741
----------------------------------------------------------------------
Marriage and education seem to have the largest effects.
```

Listcoef with reverse option doesn't work after Poisson because we are now dealing with incidence rate ratios rather than odds ratios, so it doesn't make sense to report them. To compare the effect sizes between positive and negative effects, you can still calculate them, e.g., for education:

```
.di exp(.06170*2.9741)
1.2014173
```

So the effect of marriage is still stronger than that of education.

If we have multicategory variables, pwcompare may be useful, e.g.,

```
. poisson childs i.sex i.marital sibs  i.born educ
Iteration 0:   log likelihood = -4395.7525
Iteration 1:   log likelihood = -4394.6057
Iteration 2:   log likelihood = -4394.6042
Iteration 3:   log likelihood = -4394.6042
Poisson regression                                Number of obs   =       2745
                                                  LR chi2(8)      =    1352.47
                                                  Prob > chi2     =     0.0000
Log likelihood = -4394.6042                       Pseudo R2       =     0.1334
--------------------------------------------------------------------------------
        childs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------+----------------------------------------------------------------
           sex |
       female  |   .0959266   .0295251     3.25   0.001     .0380584    .1537948
               |
       marital |
      widowed  |   .1476474   .0437708     3.37   0.001     .0618583    .2334365
     divorced  |  -.1411699   .0391833    -3.60   0.000    -.2179677    -.064372
    separated  |   -.004274    .069557    -0.06   0.951    -.1406031    .1320551
never married  |  -1.393685   .0547016   -25.48   0.000    -1.500898   -1.286472
               |
          sibs |   .0317327   .0042583     7.45   0.000     .0233866    .0400788
               |
          born |
           no  |  -.1795889   .0523534    -3.43   0.001    -.2821996   -.0769782
          educ |  -.0472726   .0048688    -9.71   0.000    -.0568153   -.0377299
         _cons |   1.266891   .0752322    16.84   0.000     1.119439    1.414343
--------------------------------------------------------------------------------

. pwcompare marital, eform
```

```
Pairwise comparisons of marginal linear predictions
Margins      : asbalanced
-----------------------------------------------------------------------------
                            |                                 Unadjusted
                            |     exp(b)   Std. Err.     [95% Conf. Interval]
----------------------------+------------------------------------------------
childs                      |
                    marital |
        widowed vs married  |   1.159104   .0507349      1.063812    1.262933
       divorced vs married  |   .8683418   .0340245      .8041514    .9376561
      separated vs married  |   .9957351   .0692603      .8688341    1.141171
  never married vs married  |   .2481592   .0135747      .2229299    .2762437
       divorced vs widowed  |   .7491491   .0386616      .6770801    .8288892
      separated vs widowed  |   .8590558   .0656894      .7394905    .9979532
  never married vs widowed  |   .2140957   .0138223      .1886485    .2429755
     separated vs divorced  |   1.146709    .086079      .9898211    1.328463
 never married vs divorced  |   .2857852   .0176029      .2532854    .3224551
never married vs separated  |   .2492221   .0210122      .2112617    .2940034
-----------------------------------------------------------------------------
```

B. *Predicted rates and changes in rates*

Next, we can examine predicted rates for various groups. For example, back to simpler model:

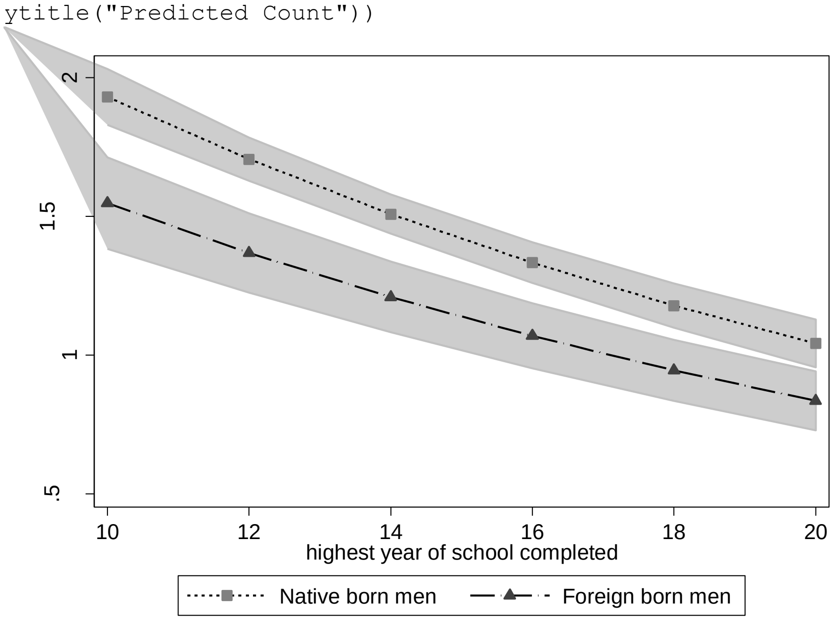
```
. qui poisson childs i.sex i.married sibs  i.born educ
. mtable, at(married=(0 1) sex=(1 2) born=1) atmeans
Expression: Predicted number of childs, predict()
           |      sex   married        mu
 ----------+-----------------------------
         1 |        1         0     1.276
         2 |        1         1     1.998
         3 |        2         0     1.551
         4 |        2         1     2.429

Specified values of covariates
           |     sibs      born      educ
 ----------+-----------------------------
   Current |      3.6         1      13.4
```

We can see that for an average native-born woman, the average number of children she has if she is single is 1.55 and if she is married 2.43. An average native born man has 1.27 children on average if he is single and approximately 2 children if he is married.

We can also use graphs when continuous variables are involved, e.g., to look at effects of education for native born and foreign born men:

```
. mgen, at(sex=1 born=1 educ=(10(2)20)) stub(nbm_)
Predictions from: margins, at(sex=1 born=1 educ=(10(2)20))
Variable   Obs Unique      Mean       Min       Max  Label
---------------------------------------------------------------------------------------
nbm_mu       6      6  1.496709  1.075388  1.993023  mean childs from margins
nbm_ll       6      6  1.413758  .9871829  1.890828  95% lower limit
nbm_ul       6      6   1.57966  1.163594  2.095217  95% upper limit
nbm_educ     6      6        15        10        20  highest year of school completed
---------------------------------------------------------------------------------------
Specified values of covariates
      sex       born
---------------------
        1         1

. mgen, at(sex=1 born=2 educ=(10(2)20)) stub(fbm_) atmeans
Predictions from: margins, at(sex=1 born=2 educ=(10(2)20)) atmeans
Variable   Obs Unique      Mean       Min       Max  Label
```

```
-------------------------------------------------------------------------------------------
fbm_mu       6      6  1.161688  .8346751  1.546907  mean childs from margins
fbm_ll       6      6  1.033278  .7287982  1.381353  95% lower limit
fbm_ul       6      6  1.290098  .9405519  1.712462  95% upper limit
fbm_educ     6      6        15        10        20  highest year of school completed
-------------------------------------------------------------------------------------------
Specified values of covariates
      sex    married       sibs       born
-------------------------------------------
        1    .459745   3.601821          2

. lab var nbm_mu "Native born men"
. lab var fbm_mu "Foreign born men"

. graph twoway (rarea  nbm_ul nbm_ll nbm_educ, color(gs12) ) (rarea  fbm_ul fbm_ll
fbm_educ, color(gs12) ) (connected nbm_mu fbm_mu nbm_educ, legend(order(3 4))
ytitle("Predicted Count"))
```



In addition to rates themselves, we can also examine how such predicted rates change per change of each independent variable – like in logit, we can examine discrete changes or marginal changes.

```
. mchange, amount(all)
poisson: Changes in mu | Number of obs = 2745
Expression: Predicted number of childs, predict()
             |    Change    p-value
-------------+----------------------
sex          |
      0 to 1 |     0.287      0.000
          +1 |     0.391      0.000
         +SD |     0.185      0.000
       Range |     0.349      0.000
    Marginal |     0.354      0.000
married      |
      0 to 1 |     0.819      0.000
          +1 |     1.026      0.000
         +SD |     0.454      0.000
       Range |     0.819      0.000
    Marginal |     0.813      0.000
sibs         |
      0 to 1 |     0.061      0.000
          +1 |     0.071      0.000
```

```
          +SD |     0.222      0.000
        Range |     2.682      0.000
     Marginal |     0.070      0.000
born          |
       0 to 1 |    -0.457      0.000
           +1 |    -0.359      0.000
          +SD |    -0.112      0.000
        Range |    -0.366      0.000
     Marginal |    -0.400      0.000
educ          |
       0 to 1 |    -0.242      0.000
           +1 |    -0.108      0.000
          +SD |    -0.304      0.000
        Range |    -2.871      0.000
     Marginal |    -0.112      0.000

Average prediction
    1.812
```

To make this more interpretable, let’s indicate which variables are dummies:

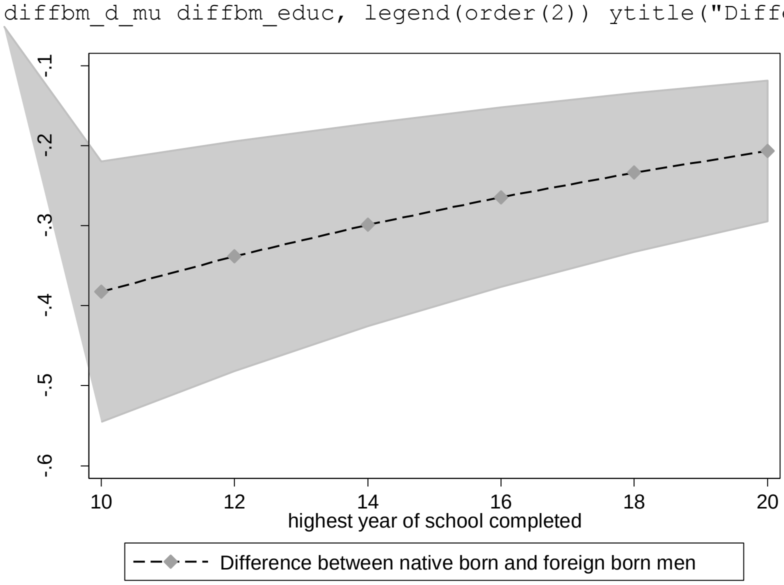
```
. poisson childs i.sex i.married sibs  i.born educ
Iteration 0:   log likelihood = -4784.5123
Iteration 1:   log likelihood = -4784.5079
Iteration 2:   log likelihood = -4784.5079
Poisson regression                                Number of obs   =       2745
                                                  LR chi2(5)      =     572.66
                                                  Prob > chi2     =     0.0000
Log likelihood = -4784.5079                       Pseudo R2       =     0.0565
------------------------------------------------------------------------------
      childs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
     female  |    .195229   .0289993     6.73   0.000     .1383915    .2520665
   1.married |   .4486183   .0288777    15.54   0.000      .392019    .5052176
        sibs |   .0385556    .004219     9.14   0.000     .0302865    .0468246
             |
        born |
         no  |  -.2209195   .0522438    -4.23   0.000    -.3233154   -.1185235
        educ |   -.061697   .0048163   -12.81   0.000    -.0711369   -.0522572
       _cons |   .9290274   .0724785    12.82   0.000     .7869721    1.071083
------------------------------------------------------------------------------

. mchange, amount(all)
poisson: Changes in mu | Number of obs = 2745
Expression: Predicted number of childs, predict()
                |    Change    p-value
----------------+----------------------
sex             |
 female vs male |     0.349      0.000
married         |
         1 vs 0 |     0.819      0.000
sibs            |
         0 to 1 |     0.061      0.000
             +1 |     0.071      0.000
            +SD |     0.222      0.000
          Range |     2.682      0.000
       Marginal |     0.070      0.000
born            |
      no vs yes |    -0.366      0.000
educ            |
         0 to 1 |    -0.242      0.000
             +1 |    -0.108      0.000
            +SD |    -0.304      0.000
```

```
        Range |    -2.871      0.000
     Marginal |    -0.112      0.000

Average prediction
    1.812
```

So for an average person, each additional sibling increases the number of children by .07, and each additional year of education decreases it by .11. Marriage increases the number of kids by .82, etc.

We can also look at changes in predicted rates graphically, e.g., to examine the difference (i.e., change when moving between categories) between native born and foreign born men depending on the value of education variable:

```
. mgen, dydx(born)  at(sex=1 educ=(10(2)20)) stub(diffbm_) atmeans
Predictions from: margins, dydx(born) at(sex=1 educ=(10(2)20)) atmeans
Variable     Obs Unique       Mean        Min        Max  Label
------------------------------------------------------------------------------------------
diffbm_d_mu    6      6  -.2871959  -.3824311  -.2063509  d_mean childs from margins
diffbm_ll      6      6  -.4093363  -.5452841  -.2946197  95% lower limit
diffbm_ul      6      6  -.1650556  -.2195781   -.118082  95% upper limit
diffbm_educ    6      6         15         10         20  highest year of school
completed
------------------------------------------------------------------------------------------

Specified values of covariates
                   1.                     2.
      sex    married       sibs       born
-------------------------------------------
        1    .459745   3.601821   .0921676

. lab var diffbm_d_mu "Difference between native born and foreign born men"

. graph twoway (rarea  diffbm_ul diffbm_ll diffbm_educ, color(gs12) ) (connected
diffbm_d_mu diffbm_educ, legend(order(2)) ytitle("Difference in Predicted Counts"))
```



C. Predicted probabilities of counts and changes in probabilities

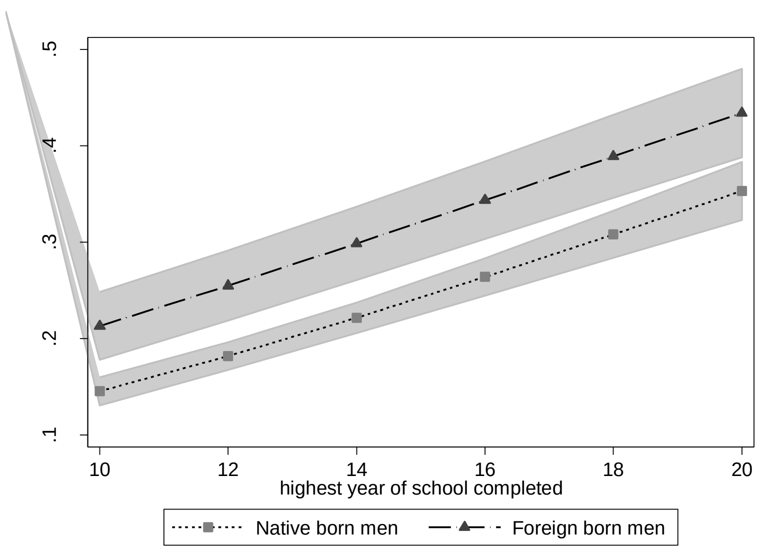
In addition to predicted rates themselves, we can also obtain predicted probabilities for each count for specific combinations of independent variables, as well as changes in such probabilities. This is especially helpful if there are some count values that are of particular interest (e.g., 0, or 1, or 2); we wouldn't usually do this for each count value. Let's look at predicted probabilities by gender and marital status for counts 0-4 kids.

```
. mtable, at(married=(0 1) sex=(1 2)) atmeans pr(0/4)
Expression: Pr(childs), predict(pr())
           |      sex   married      none       one       two     three      four
 ----------+----------------------------------------------------------------------
         1 |        1         0     0.286     0.358     0.224     0.093     0.029
         2 |        1         1     0.141     0.276     0.271     0.177     0.086
         3 |        2         0     0.219     0.332     0.253     0.128     0.049
         4 |        2         1     0.093     0.220     0.262     0.208     0.124

Specified values of covariates
           |     sibs      born      educ
 ----------+------------------------------
   Current |      3.6      1.09      13.4
```

Graphs can once again be helpful for continuous variables (or a combination of continuous and a categorical):

```
. mgen, at(sex=1 born=1 educ=(10(2)20)) stub(nbmp_) atmeans pr(0/4)
Predictions from: margins, at(sex=1 born=1 educ=(10(2)20)) atmeans predict(pr(4))
Variable   Obs Unique      Mean       Min       Max  Label
----------------------------------------------------------------------------------------
nbmp_pr0     6      6  .2455561  .1452442  .3530922  pr(y=none) from margins
nbmp_ll0     6      6  .2258325  .1306376    .32287  95% lower limit
nbmp_ul0     6      6  .2652798  .1598509  .3833145  95% upper limit
nbmp_educ    6      6        15        10        20  highest year of school completed
nbmp_Cpr0    6      6  .2455561  .1452442  .3530922  pr(y<=none)
nbmp_pr1     6      6     .3343  .2802253  .3675782  pr(y=one) from margins
nbmp_ll1     6      6  .3270483  .2666508  .3663383  95% lower limit
nbmp_ul1     6      6  .3415518  .2937998  .3688181  95% upper limit
nbmp_Cpr1    6      6  .5798562  .4254695  .7206704  pr(y<=one)
nbmp_pr2     6      6  .2375513  .1913292  .2703247  pr(y=two) from margins
nbmp_ll2     6      6  .2299111  .1762435  .2693291  95% lower limit
nbmp_ul2     6      6  .2451915  .2064149  .2713204  95% upper limit
nbmp_Cpr2    6      6  .8174075  .6957943  .9119996  pr(y<=two)
nbmp_pr3     6      6  .1174587  .0663929  .1738493  pr(y=three) from margins
nbmp_ll3     6      6  .1078026  .0556992  .1641472  95% lower limit
nbmp_ul3     6      6  .1271147  .0770866  .1835515  95% upper limit
nbmp_Cpr3    6      6  .9348661  .8696436  .9783925  pr(y<=three)
nbmp_pr4     6      6  .0453594  .0172792  .0838535  pr(y=four) from margins
nbmp_ll4     6      6   .039475  .0130754   .074803  95% lower limit
nbmp_ul4     6      6  .0512438   .021483  .0929041  95% upper limit
nbmp_Cpr4    6      6  .9802255  .9534971  .9956717  pr(y<=four)
----------------------------------------------------------------------------------------

Specified values of covariates
      sex    married       sibs       born
-------------------------------------------
        1    .459745   3.601821          1

. mgen, at(sex=1 born=2 educ=(10(2)20)) stub(fbmp_) atmeans pr(0/4)
Predictions from: margins, at(sex=1 born=2 educ=(10(2)20)) atmeans predict(pr(4))
Variable   Obs Unique      Mean       Min       Max  Label
----------------------------------------------------------------------------------------
fbmp_pr0     6      6  .3221438  .2129054  .4340155  pr(y=none) from margins
fbmp_ll0     6      6  .2822726   .177658  .3880633  95% lower limit
fbmp_ul0     6      6  .3620149  .2481528  .4799677  95% upper limit
fbmp_educ    6      6        15        10        20  highest year of school completed
```

```
fbmp_Cpr0    6      6  .3221438  .2129054  .4340155  pr(y<=none)
fbmp_pr1     6      6  .3558725  .3293449   .367287  pr(y=one) from margins
fbmp_ll1     6      6  .3469756  .3100678  .3648889  95% lower limit
fbmp_ul1     6      6  .3647694   .348622   .369859  95% upper limit
fbmp_Cpr1    6      6  .6780163  .5422503  .7962774  pr(y<=one)
fbmp_pr2     6      6  .2052781  .1511855  .2547331  pr(y=two) from margins
fbmp_ll2     6      6  .1869009  .1288374  .2423807  95% lower limit
fbmp_ul2     6      6  .2236552  .1735336  .2670854  95% upper limit
fbmp_Cpr2    6      6  .8832944  .7969834  .9474629  pr(y<=two)
fbmp_pr3     6      6  .0823727  .0420636  .1313495  pr(y=three) from margins
fbmp_ll3     6      6   .066514  .0305101  .1109228  95% lower limit
fbmp_ul3     6      6  .0982313  .0536171  .1517762  95% upper limit
fbmp_Cpr3    6      6   .965667  .9283328  .9895265  pr(y<=three)
fbmp_pr4     6      6  .0257938  .0087774  .0507964  pr(y=four) from margins
fbmp_ll4     6      6  .0182029  .0052531  .0374604  95% lower limit
fbmp_ul4     6      6  .0333847  .0123016  .0641323  95% upper limit
fbmp_Cpr4    6      6  .9914608  .9791292  .9983039  pr(y<=four)
-----------------------------------------------------------------------------------------

Specified values of covariates
      sex    married       sibs       born
-------------------------------------------
        1    .459745   3.601821          2


. lab var nbmp_pr0 "Native born men"

. lab var fbmp_pr0 "Foreign born men"

. graph twoway (rarea  nbmp_ll0 nbmp_ul0 nbmp_educ , color(gs12) ) (rarea  fbmp_ll0
fbmp_ul0 fbmp_educ , color(gs12) )  (connected nbmp_pr0  fbmp_pr0 nbmp_educ,
legend(order(3 4)) ytitle("Probability of 0 kids"))
```



Note that above, we also generated cumulative probabilities of each count or below; we can graph those as well if that is more meaningful for our variable.

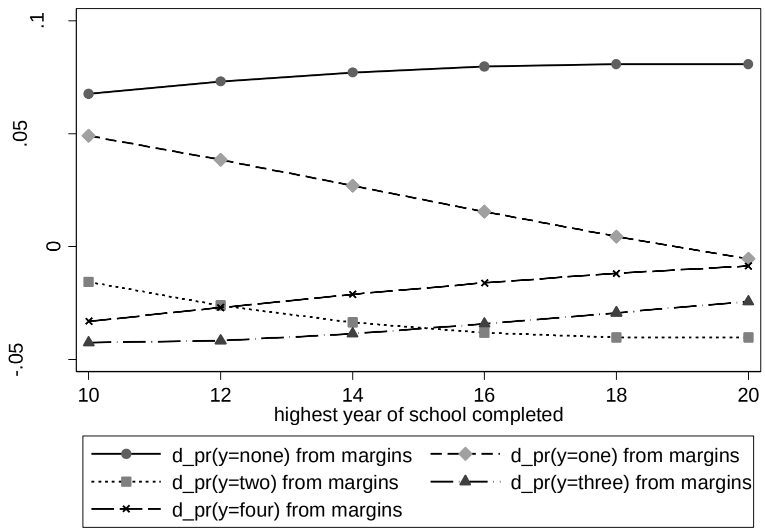
Similarly, we can examine changes in predicted probabilities of 0-4 counts:

```
. mchange married, at(sex=1 born=1) atmeans pr(0/4)
poisson: Changes in Pr(y) | Number of obs = 2745
Expression: Pr(childs), predict(pr())
```

```
             |          0           1           2           3           4
-------------+------------------------------------------------------------
married      |
          +1 |     -0.123      -0.116       0.002       0.078       0.078
     p-value |      0.000       0.000       0.676       0.000       0.000
         +SD |     -0.068      -0.051       0.014       0.043       0.034
     p-value |      0.000       0.000       0.000       0.000       0.000
    Marginal |     -0.147      -0.083       0.050       0.086       0.057
     p-value |      0.000       0.000       0.000       0.000       0.000

Predictions at base value
             |          0           1           2           3           4
-------------+------------------------------------------------------------
  Pr(y|base) |      0.208       0.327       0.256       0.134       0.053

Base values of regressors
             |        sex     married        sibs        born        educ
-------------+------------------------------------------------------------
          at |          1         .46         3.6           1        13.4

1: Estimates with margins option atmeans.
```

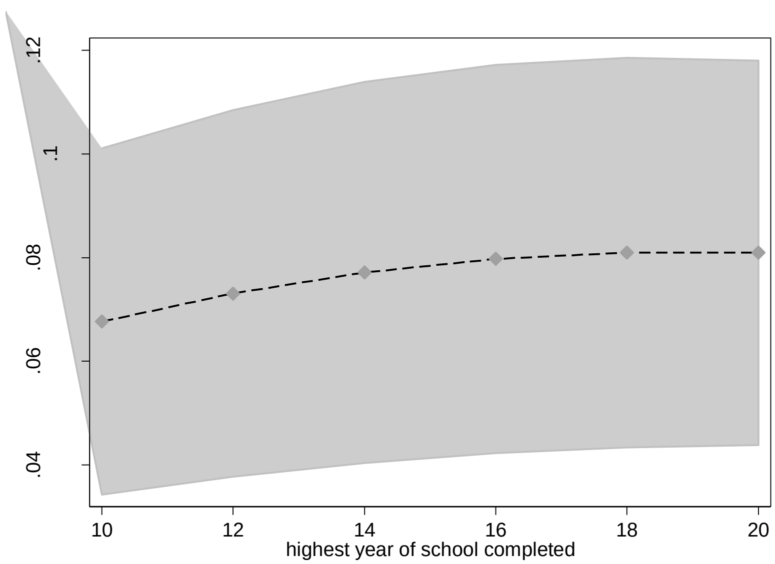
And we can examine changes in predicted probabilities of counts graphically:

```
. mgen, dydx(born) at(sex=1 educ=(10(2)20)) stub(nfbdiffp_) atmeans pr(0/4)
Predictions from: margins, dydx(born) at(sex=1 educ=(10(2)20)) atmeans predict(pr(4))
Variable      Obs Unique       Mean         Min        Max  Label
------------------------------------------------------------------------------------------
nfbdiffp_d~0    6      6   .0765876   .0676611   .0809817  d_pr(y=none) from margins
nfbdiffp_ll0    6      6   .0402904   .0342612   .0437966  95% lower limit
nfbdiffp_ul0    6      6   .1128848   .1010611   .1185632  95% upper limit
nfbdiffp_e~c    6      6         15         10         20  highest year of school
completed
nfbdiffp_C~0    6      6   .0765876   .0676611   .0809817  pr(y<=none)
nfbdiffp_d~1    6      6   .0215725  -.0053163   .0491196  d_pr(y=one) from margins
nfbdiffp_ll1    6      6    .010961   -.013708   .0295499  95% lower limit
nfbdiffp_ul1    6      6    .032184   .0030754   .0686893  95% upper limit
nfbdiffp_C~1    6      6   .0981601    .075607   .1167808  pr(y<=one)
nfbdiffp_d~2    6      6  -.0322732  -.0401783  -.0155917  d_pr(y=two) from margins
nfbdiffp_ll2    6      6   -.048958  -.0588846  -.0276518  95% lower limit
nfbdiffp_ul2    6      6  -.0155884  -.0220525  -.0035316  95% upper limit
nfbdiffp_C~2    6      6   .0658869   .0354632   .1011891  pr(y<=two)
nfbdiffp_d~3    6      6   -.035086  -.0424998  -.0243293  d_pr(y=three) from margins
nfbdiffp_ll3    6      6  -.0501805  -.0620046  -.0343967  95% lower limit
nfbdiffp_ul3    6      6  -.0199915   -.023345  -.0142619  95% upper limit
nfbdiffp_C~3    6      6   .0308009   .0111339   .0586892  pr(y<=three)
nfbdiffp_d~4    6      6  -.0195656  -.0330572  -.0085018  d_pr(y=four) from margins
nfbdiffp_ll4    6      6  -.0273416  -.0464739  -.0120411  95% lower limit
nfbdiffp_ul4    6      6  -.0117895  -.0196404  -.0049625  95% upper limit
nfbdiffp_C~4    6      6   .0112353   .0026321    .025632  pr(y<=four)
------------------------------------------------------------------------------------------

Specified values of covariates
                   1.                   2.
      sex    married       sibs       born
-------------------------------------------
        1    .459745   3.601821   .0921676

. graph twoway  (connected nfbdiffp_d_pr0 nfbdiffp_d_pr1 nfbdiffp_d_pr2 nfbdiffp_d_pr3
nfbdiffp_d_pr4 nfbdiffp_educ), ytitle("Diff. in prob. of 0-4 kids; native vs foreign
born")
```



Or focusing on one count, with confidence intervals:

```
. graph twoway (rarea  nfbdiffp_ll0 nfbdiffp_ul0 nfbdiffp_educ , color(gs12) )
(connected nfbdiffp_d_pr0  nfbdiffp_educ, legend(off) ytitle("Diff. in probability of 0
kids; native vs foreign born men"))
```



Diagnostics:

In terms of diagnostics, we can test for multicollinearity the same way we did with logistic models. To test for linearity and additivity, we can use Box-Tidwell test and mrunning and lowess using a log of the original count variable (add 1 to the count before logging it; otherwise zeros will become missing):

```
. gen countlg=log(childs+1)
```

We can also look at robust standard errors to compare them to the regular ones. We can also get residuals and leverage statistics to assess the outliers; however, to do that, we need to estimate the same model using generalized linear models command – GLM. Unfortunately, predict after Poisson is very limited, but after GLM version of Poisson we can get a range of statistics.

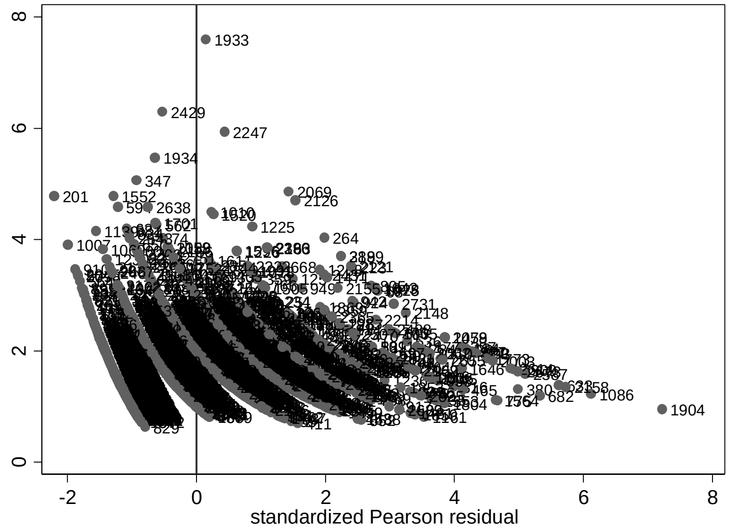
```
. glm childs sex married sibs  born educ, family(poisson)
Generalized linear models                          No. of obs      =       2745
Optimization     : ML                              Residual df     =       2739
                                                   Scale parameter =          1
Deviance         =  4279.437048                    (1/df) Deviance =   1.562409
Pearson          =  3943.169972                    (1/df) Pearson  =   1.439639
Variance function: V(u) = u                        [Poisson]
Link function    : g(u) = ln(u)                    [Log]
                                                   AIC             =   3.490352
Log likelihood   =  -4784.50787                    BIC             =   -17406.7
------------------------------------------------------------------------------
             |                 OIM
      childs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |    .195229   .0289993     6.73   0.000     .1383915    .2520665
     married |   .4486183   .0288777    15.54   0.000      .392019    .5052176
        sibs |   .0385556    .004219     9.14   0.000     .0302865    .0468246
        born |  -.2209195   .0522438    -4.23   0.000    -.3233154   -.1185235
        educ |   -.061697   .0048163   -12.81   0.000    -.0711369   -.0522572
       _cons |   .9547179   .1010692     9.45   0.000     .7566258     1.15281
------------------------------------------------------------------------------
```

Here's what we can obtain by using predict after this (among other statistics):

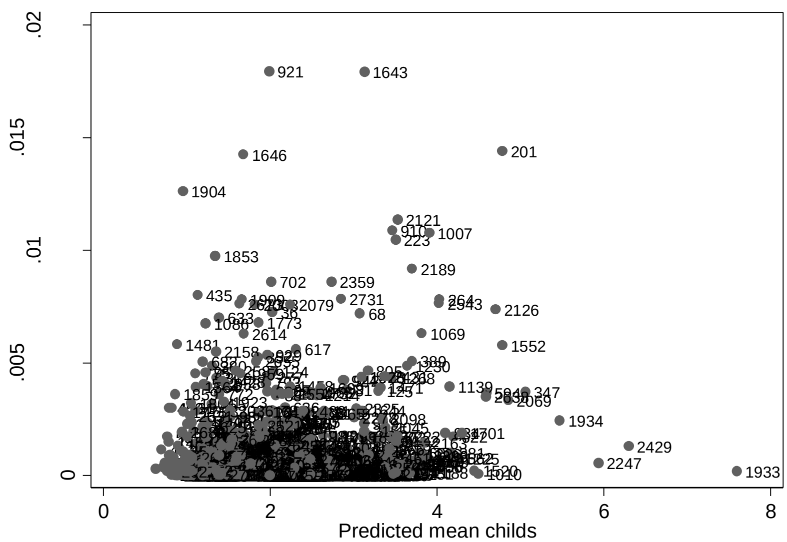
```
    cooksd calculates Cook's distance, which measures the aggregate change in
        the estimated coefficients when each observation is left out of the
        estimation.

    deviance calculates the deviance residuals.  Deviance residuals are
        recommended by McCullagh and Nelder and by others as having the best
        properties for examining the goodness of fit of a GLM.  They are
        approximately normally distributed if the model is correct.  They may
        be plotted against fitted values or against a covariate to inspect the
        model's fit.  Also see the pearson option below.

    hat calculates the diagonals of the "hat" matrix as an analog to simple
        linear regression.

    pearson calculates the Pearson residuals.  Be aware that Pearson residuals
        often have markedly skewed distributions for non-normal family
        distributions.  Also see the deviance option above.

    ----+ Options +-------------------------------------------------------------

  standardized requests that the residual be multiplied by the factor
        (1-h)^[-1/2], where h is the diagonal of the hat matrix.  This is done
        to account for the correlation between depvar and its predicted value.

    studentized requests that the residual be multiplied by one over the
        square root of the estimated scale parameter.
```

We can use these the same way we have used them after logit, e.g.:

```
. predict p
(option mu assumed; predicted mean childs)
(19 missing values generated)
. predict rs, pearson standard
(20 missing values generated)
. predict cooksd, cooksd
(20 missing values generated)
. scatter p rs, xline(0) mlabel(id)
```




```
. scatter cooksd p, mlabel(id)
```



Would have a look at 1904, 921, 1643, 1646, 201.

Models Adjusted for Exposure

Models for count data also allow controlling for so-called exposure – that is usually a variable that indicates how long there has been an opportunity to accumulate counts. E.g. an 20 y.o. and a 40 y.o. had different time available to have kids, and that will likely be reflected in their number of children. So we can control for the duration of reproductive age – that's the amount of exposure one had. Let's assume reproductive age to start at 15 and end at 45 (these numbers of course will vary individually, and it would be best to get a variable with individual data on that, but this is our best approximation):

```
. gen reprage=age-15
(14 missing values generated)
. replace reprage=30 if age>45 & age~=.
```

```
(1312 real changes made)
. poisson childs sex married sibs  born educ, exposure(reprage)
Poisson regression                                Number of obs   =       2734
                                                  LR chi2(5)      =     365.33
                                                  Prob > chi2     =     0.0000
Log likelihood = -4474.7807                       Pseudo R2       =     0.0392
------------------------------------------------------------------------------
      childs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .1829962   .0291302     6.28   0.000     .1259021    .2400902
     married |   .3223659   .0290622    11.09   0.000      .265405    .3793267
        sibs |   .0249154   .0042745     5.83   0.000     .0165375    .0332933
        born |  -.1354091   .0522745    -2.59   0.010    -.2378651    -.032953
        educ |  -.0575382    .004645   -12.39   0.000    -.0666423   -.0484341
       _cons |  -2.218856   .1006406   -22.05   0.000    -2.416108   -2.021604
     reprage | (exposure)
------------------------------------------------------------------------------
```

What this actually does is: ln(reprage) is entered in the model, but its coefficient is constrained to 1. If we don't control for exposure, it's assumed that all cases have had the same exposure.
You can get the same result by using a log of exposure variable and specifying it using offset option: essentially, exposure option enters log of the variable specified into the model, while offset enters the variable as it is (so typically you would use an already logged variable with this option); both constrain the coefficient to 1, however.

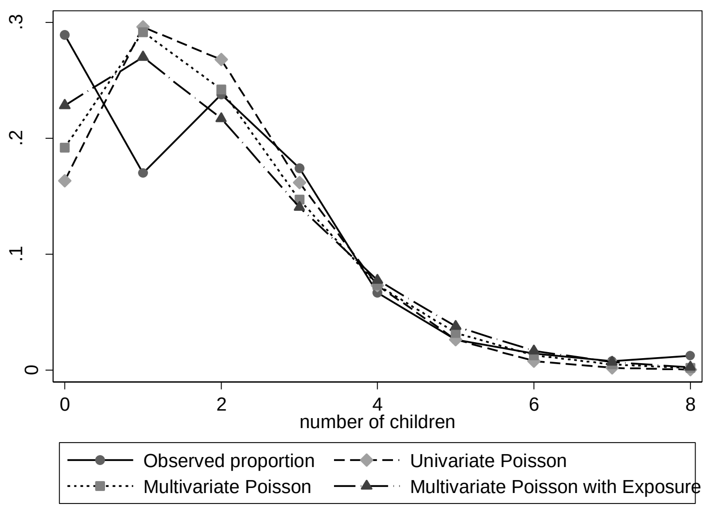
```
. gen repragelog=log(reprage)
. poisson childs sex married sibs  born educ, offset(repragelog)
Poisson regression                                Number of obs   =       2734
                                                  LR chi2(5)      =     365.33
                                                  Prob > chi2     =     0.0000
Log likelihood = -4474.7807                       Pseudo R2       =     0.0392
------------------------------------------------------------------------------
      childs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .1829962   .0291302     6.28   0.000     .1259021    .2400902
     married |   .3223659   .0290622    11.09   0.000      .265405    .3793267
        sibs |   .0249154   .0042745     5.83   0.000     .0165375    .0332933
        born |  -.1354091   .0522745    -2.59   0.010    -.2378651    -.032953
        educ |  -.0575382    .004645   -12.39   0.000    -.0666423   -.0484341
       _cons |  -2.218856   .1006406   -22.05   0.000    -2.416108   -2.021604
  repragelog |   (offset)
------------------------------------------------------------------------------.
```

We can manually replicate what these options are doing by setting a constraint on our model -- first, we specify that constraint #1 will mean repragelog coefficient should be 1, and then estimate the model adding repragelog and using constraint 1:

```
. constraint 1 repragelog=1
. poisson childs sex married sibs  born educ repragelog, constraints(1)
Poisson regression                                Number of obs   =       2734
                                                  Wald chi2(5)    =     371.72
Log likelihood = -4474.7807                       Prob > chi2     =     0.0000
 ( 1)  [childs]repragelog = 1
------------------------------------------------------------------------------
      childs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .1829962   .0291302     6.28   0.000     .1259021    .2400902
     married |   .3223659   .0290622    11.09   0.000      .265405    .3793267
        sibs |   .0249154   .0042745     5.83   0.000     .0165375    .0332933
        born |  -.1354091   .0522745    -2.59   0.010    -.2378651    -.032953
        educ |  -.0575382    .004645   -12.39   0.000    -.0666423   -.0484341
  repragelog |          1          .        .       .            .           .
       _cons |  -2.218856   .1006406   -22.05   0.000    -2.416108   -2.021604
```

After any of these models (regardless of the option), we can graphically examine model fit:

```
. mgen, pr(0/8) meanpred stub(expmpoi_)
Predictions from:
Variable      Obs Unique      Mean       Min       Max  Label
------------------------------------------------------------------------------------------
expmpoi_val     9      9         4         0         8  number of children
expmpoi_obeq    9      9  .1111111  .0080468  .2882224  Observed proportion
expmpoi_oble    9      9  .7985451  .2882224         1  Observed cum. proportion
expmpoi_preq    9      9  .1109258  .0027285   .270141  Avg predicted Pr(y=#)
expmpoi_prle    9      9  .7986562  .2284905   .998332  Avg predicted cum. Pr(y=#)
expmpoi_ob~r    9      9  .0001853  -.099329  .0597319  Observed - Avg Pr(y=#)
------------------------------------------------------------------------------------------

. lab var expmpoi_preq "Multivariate Poisson with Exposure"
. graph twoway connected poi_obeq poi_preq mpoi_preq expmpoi_preq poi_val, ylabel(0 (.1)
.3) ytitle("Probability of Count")
```



This model fits somewhat better but still has the same problems. Further, when we think that our measure of exposure is not a perfect measure of how much time one had to accumulate counts, we may just enter log of exposure variable it into the model without constraining the coefficient to 1:

```
. poisson childs sex married sibs  born educ repragelog
Poisson regression                                Number of obs   =       2734
                                                  LR chi2(6)      =    1151.72
                                                  Prob > chi2     =     0.0000
Log likelihood = -4473.9245                       Pseudo R2       =     0.1140
------------------------------------------------------------------------------
      childs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .1835258   .0291282     6.30   0.000     .1264356     .240616
     married |   .3266819   .0292507    11.17   0.000     .2693516    .3840121
        sibs |   .0254587   .0042934     5.93   0.000     .0170438    .0338737
        born |  -.1396764   .0523749    -2.67   0.008    -.2423293   -.0370235
        educ |  -.0577855   .0046561   -12.41   0.000    -.0669113   -.0486597
  repragelog |   .9417878   .0441539    21.33   0.000     .8552478    1.028328
       _cons |  -2.028168   .1760539   -11.52   0.000    -2.373228   -1.683109
------------------------------------------------------------------------------
```

Here it has a coefficient not significantly different from 1 (the confidence interval includes 1), so reprage seems to be a good estimate of exposure time. If it would be significantly different from 1, and we would have substantive reasons to believe that our measure of exposure is imperfect, we might use this model instead of the one with exposure option or offset option.

In terms of diagnostics and model fit for models with exposure, everything works the same except Box-Tidwell test which does not work with exposure or offset option, but does work with constraints – but now we need two of them:

```
. constraint 1 repragelog=1
. constraint 2 Irepr__1 =1
. boxtid poisson childs educ sex married sibs born repragelog, constraints(1 2)
Poisson regression                                Number of obs   =       2734
                                                  Wald chi2(8)    =     852.30
Log likelihood = -4472.2691                       Prob > chi2     =     0.0000
 ( 1)  [childs]Irepr__1 = 1
------------------------------------------------------------------------------
      childs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    Ieduc__1 |  -.5378193   .1694153    -3.17   0.002    -.8698671   -.2057714
    Ieduc_p1 |  -.0004982   .1461062    -0.00   0.997    -.2868611    .2858646
    Isibs__1 |   .3208799   .1250397     2.57   0.010     .0758066    .5659532
    Isibs_p1 |   .0009448     .12313     0.01   0.994    -.2403854    .2422751
    Irepr__1 |          1          .        .       .            .           .
    Irepr_p1 |   1.577093   .0834845    18.89   0.000     1.413467     1.74072
     Isex__1 |    .182057   .0292248     6.23   0.000     .1247774    .2393366
     married |   .3238347   .0292863    11.06   0.000     .2664346    .3812348
    Iborn__1 |  -.1413858   .0530498    -2.67   0.008    -.2453614   -.0374101
       _cons |   .2499269   .0315315     7.93   0.000     .1881263    .3117275
------------------------------------------------------------------------------
educ     |  -.0582693   .0046599    -12.504   Nonlin. dev. 0.069   (P = 0.793)
      p1 |   1.067851   .2711467      3.938
------------------------------------------------------------------------------
sibs     |   .0255532   .0042942      5.951   Nonlin. dev. 0.742   (P = 0.389)
      p1 |   .7165476   .3622967      1.978
------------------------------------------------------------------------------
repragelog|         1          0          .   Nonlin. dev. 4.167   (P = 0.041)
      p1 |   .2074807   .4305246      0.482
------------------------------------------------------------------------------
Deviance: 8944.406.
```

For those statistics that are obtained using predict after GLM, we need to use offset option with GLM (exposure option doesn't work for that):

```
. glm childs sex married sibs  born educ, family(poisson) offset(repragelog)
Generalized linear models                          No. of obs      =      2734
Optimization     : ML                              Residual df     =      2728
                                                   Scale parameter =         1
Deviance         =  3675.111598                    (1/df) Deviance =  1.347182
Pearson          =  3353.513369                    (1/df) Pearson  =  1.229294
Variance function: V(u) = u                        [Poisson]
Link function    : g(u) = ln(u)                    [Log]
                                                   AIC             =  3.277821
Log likelihood   = -4474.780694                    BIC             = -17912.97
------------------------------------------------------------------------------
             |                 OIM
      childs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .1829962   .0291302     6.28   0.000     .1259021    .2400902
     married |   .3223659   .0290622    11.09   0.000      .265405    .3793267
        sibs |   .0249154   .0042745     5.83   0.000     .0165375    .0332933
        born |  -.1354091   .0522745    -2.59   0.010    -.2378651    -.032953
        educ |  -.0575382    .004645   -12.39   0.000    -.0666423   -.0484341
       _cons |  -2.218856   .1006406   -22.05   0.000    -2.416108   -2.021604
  repragelog |   (offset)
------------------------------------------------------------------------------
```

After that, we can obtain residuals etc.