SC704: Topics in Multivariate Analysis
Instructor: Natasha Sarkisian
<u>Count Data Models</u>

Negative Binomial Model

Using Poisson, we attempted to account for some sources of heterogeneity – but the model doesn't fit very well. Maybe we didn't take into account all sources of heterogeneity – could try additional variables. That's important to explore, but rarely helps. In practice, Poisson regression models rarely fits due to overdispersion.

There is another process that often creates overdispersion – it is known as contagion – violation of the assumption of the independence of events. This assumption is often unrealistic; e.g. if you have your first child, that increases your chances of having your second.

To better model overdispersion from this and other sources, we can use negative binomial model. It allows taking into account unobserved heterogeneity. To do so, it introduces an additional parameter – alpha, known as the dispersion parameter. Increasing alpha increases conditional variance of X. If alpha is zero, the model becomes regular Poisson model. Here's a comparison of Poisson and negative binomial distributions with different variances for mean count=1 and mean count=10:

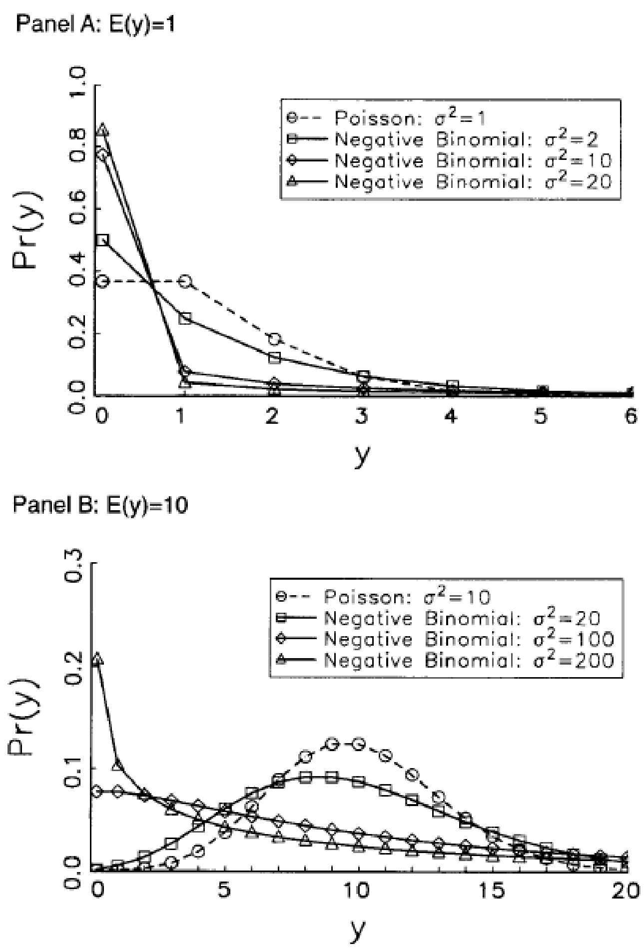


Figure 8.6. Comparisons of the Negative Binomial and Poisson Distributions

And here's an example of regression curves for negative binomial models:

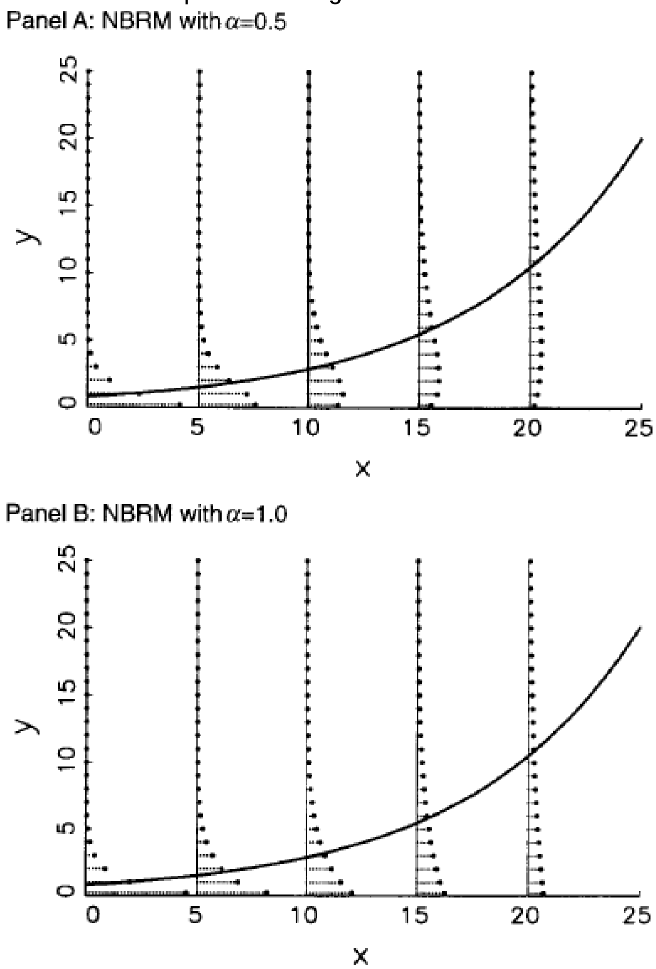


Figure 8.7. Distribution of Counts for the Negative Binomial Regression Model

Now let's run NB model for our data:

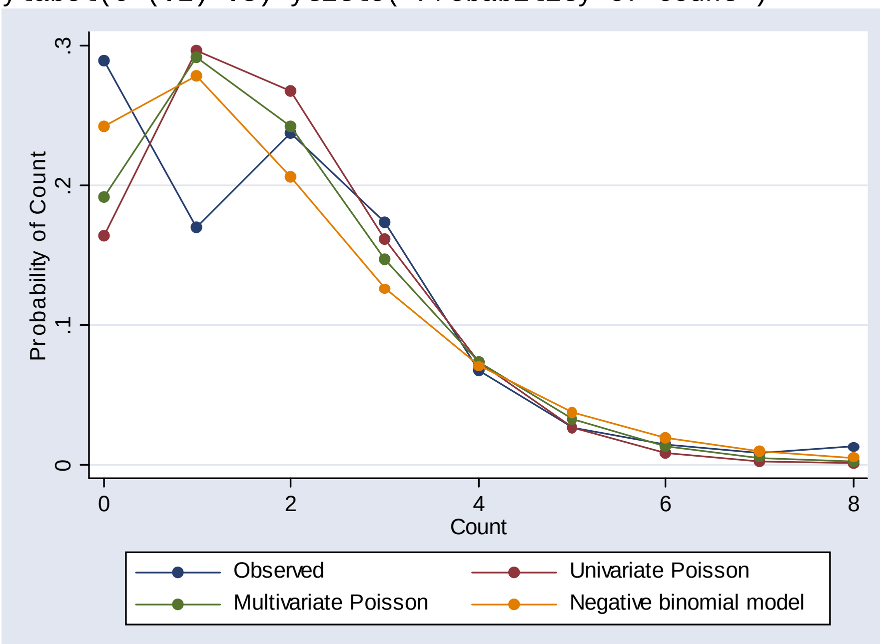
```
. nbreg childs sex married sibs  born educ
Negative binomial regression                      Number of obs   =       2745
                                                  LR chi2(5)      =     380.47
                                                  Prob > chi2     =     0.0000
Log likelihood = -4711.6789                       Pseudo R2       =     0.0388
------------------------------------------------------------------------------
      childs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .2086278   .0346569     6.02   0.000     .1407014    .2765542
     married |    .471206    .034682    13.59   0.000     .4032305    .5391816
        sibs |   .0397041   .0054244     7.32   0.000     .0290725    .0503358
        born |  -.2231164   .0616061    -3.62   0.000    -.3438622   -.1023706
        educ |  -.0616831   .0058316   -10.58   0.000    -.0731129   -.0502534
       _cons |   .9198597   .1211683     7.59   0.000     .6823743    1.157345
-------------+----------------------------------------------------------------
    /lnalpha |  -1.523939   .1086487                     -1.736886   -1.310991
-------------+----------------------------------------------------------------
       alpha |   .2178522   .0236694                      .1760678    .2695528
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) =  145.66 Prob>=chibar2 = 0.000
```

Interpretation of the results for negative binomial model is exactly the same as for Poisson model. But we have an extra line of output to interpret – the likelihood-ratio test. This allows us to see whether NB model should be used in place of regular Poisson. If probability is below the cutoff, it means that there is overdispersion (Alpha is not zero) and we should be using NB model rather than Poisson.

Now let's compare their performance graphically:

```
. prcounts nb, plot max(8)
(19 missing values generated)
. lab var nbpreq "Negative binomial model"

. gr twoway connected poisobeq poispreq prmpreq expopreq nbpreq poisval,
ylabel(0 (.1) .3) ytitle("Probability of Count")
```



The graph confirms the results of the test: NB model does better than regular multivariate Poisson. But it still underpredicts zeros and overpredicts ones. Unfortunately, the goodness of fit tests that are available after Poisson are not available after negative binomial. But the significance test for alpha tells us if Poisson performs better than negative binomial.

The interpretation tools for nbreg are the same as for poisson; we can get IRR and use prtab, prgen, prchange, and prvalue commands, as well as mfx command. We could also estimate this model with exposure.

As for diagnostics, everything is similar to Poisson, except for boxtid which doesn't work with nbreg. To obtain a GLM negative binomial model that's identical to the one estimated to nbreg, you need to specify the exact alpha to use – otherwise it uses the default value of 1 and the results differ. So here we use:

```
. glm childs sex married sibs  born educ, family(nb .2178552)

Generalized linear models                          No. of obs      =       2745
Optimization     : ML                              Residual df     =       2739
                                                   Scale parameter =          1
Deviance         =  3284.463783                    (1/df) Deviance =   1.199147
```

```
Pearson          =  2908.984543                   (1/df) Pearson  =  1.062061

Variance function: V(u) = u+(.2178552)u^2         [Neg. Binomial]
Link function    : g(u) = ln(u)                   [Log]
                                                  AIC             =  3.437289
Log likelihood   = -4711.678905                   BIC             = -18401.67
------------------------------------------------------------------------------
             |                 OIM
      childs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .2086279   .0346384     6.02   0.000     .1407379    .2765179
     married |   .4712062   .0346364    13.60   0.000     .4033201    .5390924
        sibs |   .0397041   .0054238     7.32   0.000     .0290737    .0503346
        born |  -.2231165   .0616059    -3.62   0.000    -.3438618   -.1023712
        educ |  -.0616831   .0058316   -10.58   0.000    -.0731129   -.0502533
       _cons |   .9198593   .1211388     7.59   0.000     .6824317    1.157287
------------------------------------------------------------------------------
```

We can obtain residuals etc. after this.

In addition to regular nbreg where overdispersion is assumed to be constant, we can also use generalized negative binomial regression to model overdispersion:

```
. gnbreg childs sex married sibs  born educ, lnalpha(sex married sibs  born
educ)
Generalized negative binomial regression         Number of obs   =       2745
                                                 LR chi2(5)      =     222.46
                                                 Prob > chi2     =     0.0000
Log likelihood = -4587.1261                      Pseudo R2       =     0.0237

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
childs       |
         sex |    .079685   .0354711     2.25   0.025     .0101628    .1492071
     married |   .3413691   .0387924     8.80   0.000     .2653374    .4174008
        sibs |   .0369471   .0047258     7.82   0.000     .0276847    .0462095
        born |  -.1967968   .0582151    -3.38   0.001    -.3108963   -.0826973
        educ |  -.0514978   .0056236    -9.16   0.000    -.0625199   -.0404758
       _cons |   1.085011   .1189463     9.12   0.000     .8518807    1.318142
-------------+----------------------------------------------------------------
lnalpha      |
         sex |  -1.557369   .1884906    -8.26   0.000    -1.926804   -1.187934
     married |  -4.256861    .819715    -5.19   0.000    -5.863473   -2.650249
        sibs |  -.1051836   .0405024    -2.60   0.009    -.1845669   -.0258003
        born |   .1353893   .3910783     0.35   0.729      -.63111    .9018887
        educ |   .1619184   .0358938     4.51   0.000     .0915678     .232269
       _cons |   .3279141   .7155448     0.46   0.647    -1.074528    1.730356
------------------------------------------------------------------------------
```

Looks like overdispersion parameter varies by sex, marital status, number of siblings, and education, so the contagion process operates differently for different people.

Zero-Inflated Count Data Models

The problem that our negative binomial model still has – underpredicting zeros, overpredicting ones -- is very common and sometimes this problem can be very severe when there are a lot of zeros in the distribution. Example – Sarkisian and Gerstel 2004 article. We can use zero-inflated count models to correct for that – they model two different processes. They assume two latent groups – one is capable of having positive counts, the other one is not – it will always have zero count. For example, some are capable of having children, and the number that they can have might vary, but others cannot have children and their count will always remain zero. But these two groups are latent – no information on actual fertility situation. We can also have zeros in the first group. We can distinguish structural zeros (this behavior is not in this person's repertoire at all) vs chance zeros (this behavior is in this person's repertoire, but did not occur during the specified period). E.g.: "How many times last week did you smoke marijuana?" Some zeros mean the person never smokes it; other zeros mean the person does smoke but did not smoke last week.

Therefore, this model is a two-step process – first, have to predict the membership in two groups – "always zero" and "not always zero" and second, predict the count in the "not always zero" group.

```
. zip childs sex married sibs  born educ, inflate(sex married sibs born educ)

Zero-inflated poisson regression                  Number of obs   =       2745
                                                  Nonzero obs     =       1951
                                                  Zero obs        =        794

Inflation model = logit                           LR chi2(5)      =     130.65
Log likelihood  = -4524.192                       Prob > chi2     =     0.0000
------------------------------------------------------------------------------
      childs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
childs       |
         sex |   .0014908   .0320997     0.05   0.963    -.0614234     .064405
     married |   .0307475   .0333411     0.92   0.356    -.0345999    .0960949
        sibs |   .0292838   .0045691     6.41   0.000     .0203286     .038239
        born |  -.1728303   .0563097    -3.07   0.002    -.2831953   -.0624654
        educ |  -.0382489   .0052824    -7.24   0.000    -.0486021   -.0278956
       _cons |   1.363043   .1094042    12.46   0.000     1.148615    1.577472
-------------+----------------------------------------------------------------
inflate      |
         sex |  -1.267402   .1427508    -8.88   0.000    -1.547189    -.987616
     married |  -3.867796   .6722317    -5.75   0.000    -5.185346   -2.550246
        sibs |  -.0907598   .0284525    -3.19   0.001    -.1465256    -.034994
        born |   .3182067   .2733966     1.16   0.244    -.2176408    .8540542
        educ |   .1671403   .0267744     6.24   0.000     .1146635    .2196171
       _cons |  -.9103566   .5168716    -1.76   0.078    -1.923406     .102693
------------------------------------------------------------------------------
```

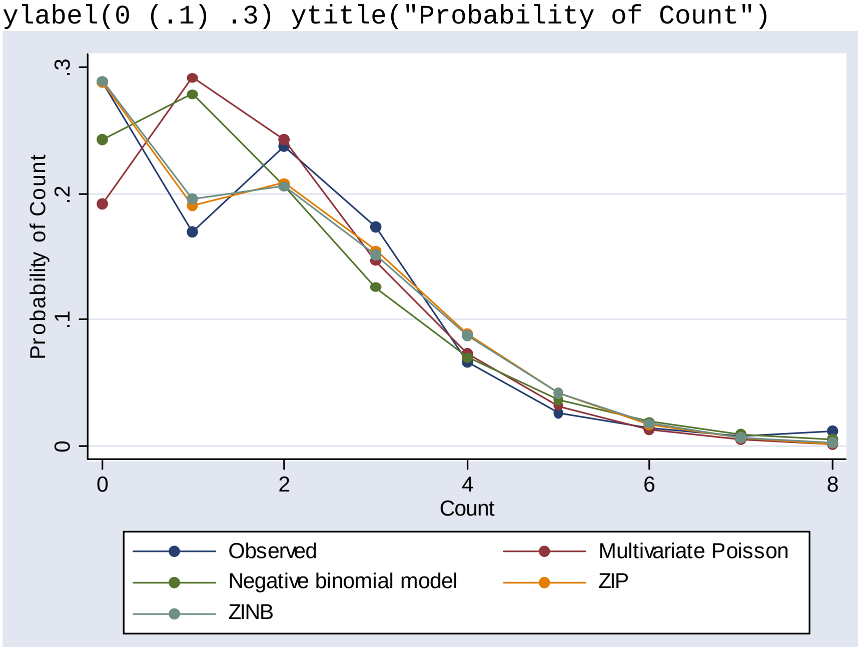
Note the inflate option we specified – we have to specify that option, it tells Stata what variables to use to predict the membership in "Always Zero" group. In this case, we used the same variables but we could have used a smaller subset of the variables or even different variables altogether. We'll return to interpreting this output. But let's prepare to graphically examine the fit:

```
. prcounts zip, plot max(8)
(19 missing values generated)
. lab var zippreq "ZIP"
```

```
. zinb childs sex married sibs  born educ, inflate(sex married sibs born educ)
Zero-inflated negative binomial regression        Number of obs    =       2745
                                                  Nonzero obs      =       1951
                                                  Zero obs         =        794
Inflation model = logit                           LR chi2(5)       =     124.23
Log likelihood  =  -4522.91                       Prob > chi2      =     0.0000
------------------------------------------------------------------------------
      childs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
childs       |
         sex |   .0060583   .0331917     0.18   0.855    -.0589961    .0711128
     married |   .0346028   .0344018     1.01   0.314    -.0328234     .102029
        sibs |   .0297016    .004743     6.26   0.000     .0204055    .0389977
        born |  -.1730859   .0572733    -3.02   0.003    -.2853394   -.0608324
        educ |  -.0384851   .0054302    -7.09   0.000    -.0491281   -.0278422
       _cons |   1.347192   .1125643    11.97   0.000      1.12657    1.567814
-------------+----------------------------------------------------------------
inflate      |
         sex |  -1.290154   .1468538    -8.79   0.000    -1.577982   -1.002326
     married |  -4.405718   1.215488    -3.62   0.000     -6.78803   -2.023406
        sibs |  -.0911606     .02947    -3.09   0.002    -.1489207   -.0334006
        born |   .3417874   .2818703     1.21   0.225    -.2106681     .894243
        educ |   .1715742   .0277136     6.19   0.000     .1172565    .2258919
       _cons |  -.9919407   .5360101    -1.85   0.064    -2.042501    .0586197
-------------+----------------------------------------------------------------
    /lnalpha |  -3.718083   .6593754    -5.64   0.000    -5.010435   -2.425731
-------------+----------------------------------------------------------------
       alpha |   .0242805   .0160099                       .006668    .0884134
------------------------------------------------------------------------------
. prcounts zinb, plot max(8)
(19 missing values generated)
. lab var zinbpreq "ZINB"
```

Before interpreting the results, let’s figure out which model fits best.

```
. gr twoway connected poisobeq prmpreq nbpreq zippreq zinbpreq poisval,
ylabel(0 (.1) .3) ytitle("Probability of Count")
```



Both ZIP and ZINB approximate the observed distribution much better than regular Poisson and NB models. We could also plot deviations from observed counts rather than actual counts and get comparisons of fit:

```
. countfit childs sex married sibs  born educ, inflate(sex married sibs born
educ)
--------------------------------------------------------------------------
                        Variable |    PRM        NBRM         ZIP       ZINB
---------------------------------+----------------------------------------
childs                           |
                 respondents sex |     1.216       1.232       1.001       1.006
                                 |      6.73        6.02        0.05        0.18
                    R is married |     1.566       1.602       1.031       1.035
                                 |     15.54       13.59        0.92        1.01
   number of brothers and sisters |     1.039       1.041       1.030       1.030
                                 |      9.14        7.32        6.41        6.26
       was r born in this country |     0.802       0.800       0.841       0.841
                                 |     -4.23       -3.62       -3.07       -3.02
highest year of school completed |     0.940       0.940       0.962       0.962
                                 |    -12.81      -10.58       -7.24       -7.09
                        Constant |     2.598       2.509       3.908       3.847
                                 |      9.45        7.59       12.46       11.97
---------------------------------+----------------------------------------
lnalpha                          |
                        Constant |                 0.218                   0.024
                                 |                -14.03                   -5.64
---------------------------------+----------------------------------------
inflate                          |
                 respondents sex |                             0.282       0.275
                                 |                             -8.88       -8.79
                    R is married |                             0.021       0.012
                                 |                             -5.75       -3.62
   number of brothers and sisters |                             0.913       0.913
                                 |                             -3.19       -3.09
       was r born in this country |                             1.375       1.407
                                 |                              1.16        1.21
highest year of school completed |                             1.182       1.187
                                 |                              6.24        6.19
                        Constant |                             0.402       0.371
                                 |                             -1.76       -1.85
---------------------------------+----------------------------------------
Statistics                       |
                           alpha |                 0.218
                               N |      2745        2745        2745        2745
                              ll | -4784.508   -4711.679   -4524.192   -4522.910
                             bic |  9616.521    9478.781    9143.394    9148.749
                             aic |  9581.016    9437.358    9072.383    9071.821
--------------------------------------------------------------------------
                                                                 legend: b/t
Comparison of Mean Observed and Predicted Count
            Maximum       At       Mean
Model     Difference    Value    |Diff|
---------------------------------------------
PRM        -0.122          1      0.028
NBRM       -0.109          1      0.027
ZIP         0.030          2      0.012
ZINB        0.032          2      0.013

PRM: Predicted and actual probabilities
Count   Actual    Predicted    |Diff|   Pearson
------------------------------------------------
```

0	0.289	0.192	0.097	135.055
1	0.170	0.292	0.122	139.312
2	0.238	0.242	0.005	0.231
3	0.174	0.147	0.027	13.674
4	0.067	0.073	0.006	1.361
5	0.026	0.032	0.006	3.069
6	0.015	0.013	0.002	0.526
7	0.008	0.005	0.003	5.097
8	0.012	0.002	0.011	163.156
9	0.000	0.001	0.001	1.924
Sum	1.000	1.000	0.278	463.405

NBRM: Predicted and actual probabilities

Count	Actual	Predicted	|Diff|	Pearson
0	0.289	0.242	0.047	24.952
1	0.170	0.279	0.109	116.103
2	0.238	0.206	0.032	13.512
3	0.174	0.126	0.048	50.004
4	0.067	0.070	0.003	0.315
5	0.026	0.037	0.011	8.820
6	0.015	0.019	0.005	3.010
7	0.008	0.010	0.002	0.867
8	0.012	0.005	0.007	30.214
9	0.000	0.003	0.003	7.016
Sum	1.000	0.997	0.265	254.813

ZIP: Predicted and actual probabilities

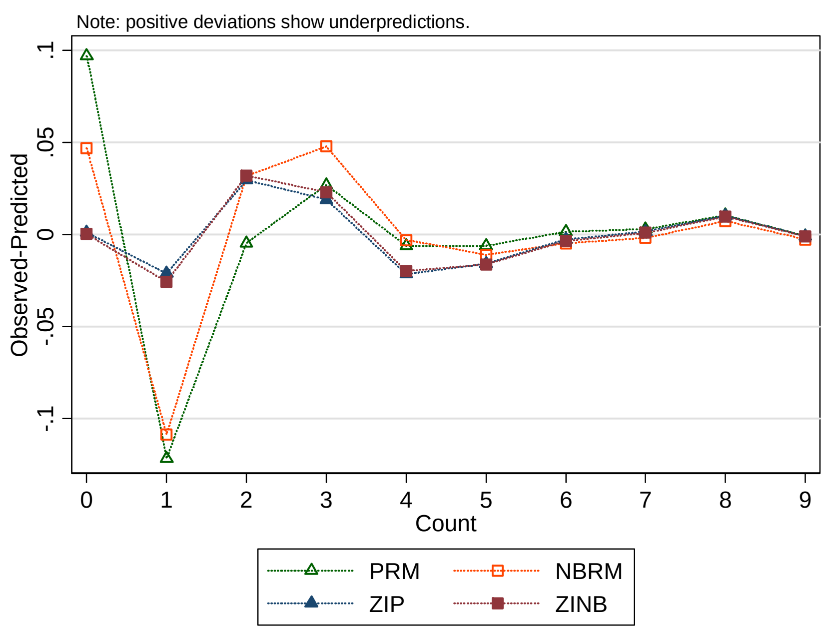
Count	Actual	Predicted	|Diff|	Pearson
0	0.289	0.288	0.001	0.014
1	0.170	0.191	0.021	6.403
2	0.238	0.208	0.030	11.561
3	0.174	0.155	0.019	6.512
4	0.067	0.089	0.021	14.210
5	0.026	0.042	0.016	16.286
6	0.015	0.017	0.003	1.083
7	0.008	0.006	0.002	1.298
8	0.012	0.002	0.010	135.546
9	0.000	0.001	0.001	1.886
Sum	1.000	1.000	0.124	194.798

ZINB: Predicted and actual probabilities

Count	Actual	Predicted	|Diff|	Pearson
0	0.289	0.289	0.000	0.001
1	0.170	0.196	0.026	9.202
2	0.238	0.206	0.032	13.730
3	0.174	0.151	0.023	9.695
4	0.067	0.087	0.020	12.320
5	0.026	0.042	0.016	16.787
6	0.015	0.018	0.003	1.855
7	0.008	0.007	0.001	0.389
8	0.012	0.003	0.010	104.052

```
9         0.000        0.001       0.001     2.445
------------------------------------------------
Sum       1.000        1.000       0.132   170.477

Tests and Fit Statistics
PRM             BIC=-12117.116  AIC=      3.490  Prefer  Over  Evidence
-------------------------------------------------------------------------
  vs NBRM       BIC=-12254.857  dif=    137.740  NBRM    PRM   Very strong
                AIC=      3.438  dif=      0.052  NBRM    PRM
                LRX2=   145.658  prob=     0.000  NBRM    PRM   p=0.000
-------------------------------------------------------------------------
  vs ZIP        BIC=-12590.244  dif=    473.127  ZIP     PRM   Very strong
                AIC=      3.305  dif=      0.185  ZIP     PRM
                Vuong=   11.165  prob=     0.000  ZIP     PRM   p=0.000
-------------------------------------------------------------------------
  vs ZINB       BIC=-12584.889  dif=    467.772  ZINB    PRM   Very strong
                AIC=      3.305  dif=      0.185  ZINB    PRM
-------------------------------------------------------------------------
NBRM            BIC=-12254.857  AIC=      3.438  Prefer  Over  Evidence
-------------------------------------------------------------------------
  vs ZIP        BIC=-12590.244  dif=    335.387  ZIP     NBRM  Very strong
                AIC=      3.305  dif=      0.133  ZIP     NBRM
-------------------------------------------------------------------------
  vs ZINB       BIC=-12584.889  dif=    330.032  ZINB    NBRM  Very strong
                AIC=      3.305  dif=      0.133  ZINB    NBRM
                Vuong=   10.441  prob=     0.000  ZINB    NBRM  p=0.000
-------------------------------------------------------------------------
ZIP             BIC=-12590.244  AIC=      3.305  Prefer  Over  Evidence
-------------------------------------------------------------------------
  vs ZINB       BIC=-12584.889  dif=     -5.355  ZIP     ZINB  Positive
                AIC=      3.305  dif=      0.000  ZINB    ZIP
                LRX2=     2.563  prob=     0.055  ZINB    ZIP   p=0.000
-------------------------------------------------------------------------
```



So now let’s interpret this final model:

```
. zip childs sex married sibs  born educ, inflate(sex married sibs born educ)
Zero-inflated poisson regression                  Number of obs   =      2745
                                                  Nonzero obs     =      1951
                                                  Zero obs        =       794
Inflation model = logit                           LR chi2(5)      =    130.65
Log likelihood  = -4524.192                       Prob > chi2     =    0.0000
------------------------------------------------------------------------------
      childs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
childs       |
         sex |   .0014908   .0320997     0.05   0.963    -.0614234     .064405
     married |   .0307475   .0333411     0.92   0.356    -.0345999    .0960949
        sibs |   .0292838   .0045691     6.41   0.000     .0203286     .038239
        born |  -.1728303   .0563097    -3.07   0.002    -.2831953   -.0624654
        educ |  -.0382489   .0052824    -7.24   0.000    -.0486021   -.0278956
       _cons |   1.363043   .1094042    12.46   0.000     1.148615    1.577472
-------------+----------------------------------------------------------------
inflate      |
         sex |  -1.267402   .1427508    -8.88   0.000    -1.547189    -.987616
     married |  -3.867796   .6722317    -5.75   0.000    -5.185346   -2.550246
        sibs |  -.0907598   .0284525    -3.19   0.001    -.1465256    -.034994
        born |   .3182067   .2733966     1.16   0.244    -.2176408    .8540542
        educ |   .1671403   .0267744     6.24   0.000     .1146635    .2196171
       _cons |  -.9103566   .5168716    -1.76   0.078    -1.923406     .102693
------------------------------------------------------------------------------
```

The first set of coefficients is from the equation predicting counts for the “Not Always Zero” group. These show that number of siblings increases number of children and being foreign born and having more education decreases it. These coefficients can be interpreted the same way as regular Poisson coefficients.

The second set of coefficients is from the equation that predicts membership in “Always Zero” group. These can be interpreted as logit coefficients. Note that they predict zeros – so their sign will usually be the opposite to that of the coefficients in the upper half of the output. These show that women are less likely than men to be in “Always zero” group, married are less likely than single people to be in it, those with more siblings are also likely to be in it, and those with more education are more likely to be in “Always zero” group.

To be able to interpret the size of these effects, let’s use listcoef:

```
. listcoef
zip (N=2745): Factor Change in Expected Count
 Observed SD: 1.6887584
Count Equation: Factor Change in Expected Count for Those Not Always 0
----------------------------------------------------------------------
      childs |      b         z     P>|z|    e^b    e^bStdX      SDofX
-------------+--------------------------------------------------------
         sex |   0.00149    0.046   0.963   1.0015   1.0007     0.4970
     married |   0.03075    0.922   0.356   1.0312   1.0154     0.4985
        sibs |   0.02928    6.409   0.000   1.0297   1.0919     3.0008
        born |  -0.17283   -3.069   0.002   0.8413   0.9512     0.2893
        educ |  -0.03825   -7.241   0.000   0.9625   0.8925     2.9741
----------------------------------------------------------------------
Binary Equation: Factor Change in Odds of Always 0
----------------------------------------------------------------------
```

Always0	b	z	P>|z|	e^b	e^bStdX	SDofX
sex	-1.26740	-8.878	0.000	0.2816	0.5326	0.4970
married	-3.86780	-5.754	0.000	0.0209	0.1454	0.4985
sibs	-0.09076	-3.190	0.001	0.9132	0.7616	3.0008
born	0.31821	1.164	0.244	1.3747	1.0964	0.2893
educ	0.16714	6.243	0.000	1.1819	1.6439	2.9741

Or better yet with percentages:

```
. listcoef, percent
zip (N=2745): Percentage Change in Expected Count
 Observed SD: 1.6887584
Count Equation: Percentage Change in Expected Count for Those Not Always 0
```

childs	b	z	P>|z|	%	%StdX	SDofX
sex	0.00149	0.046	0.963	0.1	0.1	0.4970
married	0.03075	0.922	0.356	3.1	1.5	0.4985
sibs	0.02928	6.409	0.000	3.0	9.2	3.0008
born	-0.17283	-3.069	0.002	-15.9	-4.9	0.2893
educ	-0.03825	-7.241	0.000	-3.8	-10.8	2.9741

Binary Equation: Factor Change in Odds of Always 0

Always0	b	z	P>|z|	%	%StdX	SDofX
sex	-1.26740	-8.878	0.000	-71.8	-46.7	0.4970
married	-3.86780	-5.754	0.000	-97.9	-85.5	0.4985
sibs	-0.09076	-3.190	0.001	-8.7	-23.8	3.0008
born	0.31821	1.164	0.244	37.5	9.6	0.2893
educ	0.16714	6.243	0.000	18.2	64.4	2.9741

Each additional sibling increases one’s count by 3%, each year of education decreases it by 3.8%, and being foreign born decreases it by 16%. At the same time, women’s odds of having no kids (being in always zero group) are 71.8% lower than men’s, and the odds for married to be in always zero group are 97.9% lower than for single people. Further, each additional sibling decreases one’s odds of not having kids by 8.7% and each additional year of education increases those odds by 18.2%.

Further, as for regular Poisson we can interpret predicted rates and predicted probabilities. Predicted rates for native-born:

```
. prtab sex married, x(born=1)
zip: Predicted rates for childs
```

responden ts sex	married 0	1
male	1.0721	2.2151
female	1.6977	2.2531

```
base x values for count equation:
          sex    married       sibs       born       educ
x=  1.5555556  .45974499  3.6018215          1  13.358834
base z values for binary equation:
          sex    married       sibs       born       educ
z=  1.5555556  .45974499  3.6018215          1  13.358834
```

Note that we could have separately specified the values of independent variables for the two equations – we would only used that if we used different variables in the two equations.

For foreign-born:

```
. prtab sex married, x(born=2)
zip: Predicted rates for childs
--------------------------
responden |    married
ts sex    |      0       1
----------+---------------
     male | 0.7569  1.8487
   female | 1.3159  1.8912
--------------------------
base x values for count equation:
          sex    married       sibs       born       educ
x=  1.5555556  .45974499  3.6018215          2  13.358834

base z values for binary equation:
          sex    married       sibs       born       educ
z=  1.5555556  .45974499  3.6018215          2  13.358834
```

We can also examine changes in predicted rates as well as marginal effects.

```
. prchange
zip: Changes in Predicted Rate for childs
         min->max      0->1     -+1/2    -+sd/2
    sex    0.2339    0.5252    0.2212    0.1072
married    0.7951    0.7951    0.8680    0.3761
   sibs    2.4221    0.0697    0.0740    0.2221
   born   -0.3756   -0.4412   -0.4010   -0.1159
   educ   -2.2847   -0.1419   -0.1047   -0.3117
exp(xb):   2.0117
base x values for count equation:
            sex  married     sibs     born     educ
    x=  1.55556  .459745  3.60182  1.09217  13.3588
sd(x)=  .496995  .498468  3.00084  .289315  2.97411

base z values for binary equation:
            sex  married     sibs     born     educ
    z=  1.55556  .459745  3.60182  1.09217  13.3588
sd(z)=  .496995  .498468  3.00084  .289315  2.97411
```

We interpret these results the same way as for regular Poisson model. Note that here prchange does not compute marginal effects. But we can obtain them using mfx compute (this calculation will take a long time – takes a while to calculate standard errors).

```
. mfx compute
Marginal effects after zip
      y  = predicted number of events (predict)
         =  2.0116755
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
     sex |   .2137696      .07513    2.85   0.004   .066517  .361022   1.55556
 married*|   .7950725      .06097   13.04   0.000   .675569  .914576   .459745
    sibs |    .074003       .0096    7.71   0.000   .055192  .092814   3.60182
```

```
    born |  -.4005967      .11142   -3.60   0.000  -.618976 -.182218   1.09217
    educ |  -.1047399      .01113   -9.41   0.000  -.126553 -.082927   13.3588
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

Note that all marginal effects are significant – this is because some of the variables had significant coefficients in the count model, and others in “Always zero” model, and marginal effects combined the two to calculate the overall impact of each variable on the expected count. It is evaluated at the mean of each variable with other variables also held at their means; for dummy variables it is evaluated as discrete change in the predicted rate. Unfortunately, because our sex and born variables are not 0-1 variables, mfx compute does not realize they are dummy variables. Therefore, always try to code all dummies as 0-1. An example of using marginal effects can be found in Sarkisian and Gerstel 2004.

We can also examine predicted probabilities using prvalue and prgen. The only difference in using these is that now we will get two probabilities for zero: One is the total probability – either because one is in “Always Zero” group or because they just didn’t have their first kid yet. The other one is probability of being in “Always zero” group only. Let’s examine these:

```
. prvalue, x(married=0 sex=1 born=1)
 zip: Predictions for childs
Predicted rate: 1.07
Predicted probabilities:
  Pr(y=0|x,z): 0.6788  Pr(y=1|x):   0.1792
  Pr(y=2|x):   0.0961  Pr(y=3|x):   0.0343
  Pr(y=4|x):   0.0092  Pr(y=5|x):   0.0020
  Pr(y=6|x):   0.0004  Pr(y=7|x):   0.0001
  Pr(y=8|x):   0.0000  Pr(y=9|x):   0.0000
Pr(Always0|z): 0.5116
x values for count equation
          sex    married       sibs       born       educ
x=          1          0  3.6018215          1  13.358834
z values for binary equation
          sex    married       sibs       born       educ
z=          1          0  3.6018215          1  13.358834
```

These were predicted probabilities (and the predicted rate!) for average single native-born men. We can see that according to our model 68% of these men don’t have kids and most of these men are in always zero group – the probability of being in that group is .51. So the remaining 17% we assume just didn’t start having children yet. No let’s look at married men:

```
. prvalue, x(married=1 sex=1 born=1)
zip: Predictions for childs
Predicted rate: 2.22
Predicted probabilities:
  Pr(y=0|x,z): 0.1282  Pr(y=1|x):   0.2366
  Pr(y=2|x):   0.2620  Pr(y=3|x):   0.1935
  Pr(y=4|x):   0.1071  Pr(y=5|x):   0.0475
  Pr(y=6|x):   0.0175  Pr(y=7|x):   0.0055
  Pr(y=8|x):   0.0015  Pr(y=9|x):   0.0004
Pr(Always0|z): 0.0214
x values for count equation
          sex    married       sibs       born       educ
x=          1          1  3.6018215          1  13.358834
z values for binary equation
```

```
          sex    married       sibs       born       educ
z=          1          1  3.6018215          1  13.358834
```

Only 13% of these men are expected to have no kids, and only 2% of them are in always zero group – the remaining 11% just didn't start having kids yet. We can do a similar analysis for women – let's put their results next to each other:

```
. quietly prvalue, x(married=0 sex=2 born=1) save
. prvalue, x(married=1 sex=2 born=1) dif
zip: Change in Predictions for  childs
Predicted rate: 2.25           Saved: 1.7
    Difference: .555
Predicted probabilities:
                     Current       Saved  Difference
  Pr(y=0|x,z):        0.1106      0.3692     -0.2586
  Pr(y=1|x):          0.2353      0.2401     -0.0048
  Pr(y=2|x):          0.2651      0.2038      0.0613
  Pr(y=3|x):          0.1991      0.1153      0.0838
  Pr(y=4|x):          0.1121      0.0489      0.0632
  Pr(y=5|x):          0.0505      0.0166      0.0339
  Pr(y=6|x):          0.0190      0.0047      0.0143
  Pr(y=7|x):          0.0061      0.0011      0.0050
  Pr(y=8|x):          0.0017      0.0002      0.0015
  Pr(y=9|x):          0.0004      0.0000      0.0004
Pr(Always0|z):        0.0061      0.2278     -0.2216
x values for count equation
                sex    married       sibs       born       educ
Current=          2          1  3.6018215          1  13.358834
  Saved=          2          0  3.6018215          1  13.358834
   Diff=          0          1          0          0          0
z values for binary equation
                sex    married       sibs       born       educ
Current=          2          1  3.6018215          1  13.358834
  Saved=          2          0  3.6018215          1  13.358834
   Diff=          0          1          0          0          0
```

According to our model, 36% of single women don't have kids and 23% never will, while only 11% of married women don't have kids and only 0.6% never will.

We can also use prgen to make graphs like we did for Poisson model – but here again we will have two sets of probabilities for zero counts –total probability of zero and probability of "Always zero." E.g., see Long and Freese p. 282.

We can also adjust our final, best-fitting model to exposure time:

```
. zip childs sex married sibs  born educ, inflate(sex married sibs born educ)
exposure(reprage)
(31 missing values generated)

Zero-inflated poisson regression                  Number of obs   =       2734
                                                  Nonzero obs     =       1946
                                                  Zero obs        =        788
Inflation model = logit                           LR chi2(5)      =     119.40
Log likelihood  = -4334.455                       Prob > chi2     =     0.0000
------------------------------------------------------------------------------
      childs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
childs       |
```

```
         sex |   .0673734   .0319959     2.11   0.035     .0046625    .1300842
     married |   .0372361   .0329312     1.13   0.258    -.0273079      .10178
        sibs |   .0213414    .004529     4.71   0.000     .0124647    .0302181
        born |   -.099738   .0548672    -1.82   0.069    -.2072757    .0077996
        educ |    -.04122   .0051174    -8.05   0.000    -.0512498   -.0311901
       _cons |  -1.996286   .1081046   -18.47   0.000    -2.208167   -1.784405
     reprage | (exposure)
-------------+----------------------------------------------------------------
inflate      |
         sex |  -1.258563   .1789565    -7.03   0.000    -1.609311   -.9078144
     married |   -7.69451   37.75966    -0.20   0.839    -81.70207    66.31305
        sibs |  -.0533748   .0340675    -1.57   0.117    -.1201459    .0133964
        born |   .3318979   .3383992     0.98   0.327    -.3313523    .9951481
        educ |   .1963433   .0342241     5.74   0.000     .1292652    .2634213
       _cons |  -1.914812   .6732486    -2.84   0.004    -3.234355   -.5952693
------------------------------------------------------------------------------
```

Note that the model changed – marriage that seemed so important is no longer significant! Looks like that was just function of age. Sex, siblings, and education predict the count, and sex and education predict the membership in always zero group.

Let’s use fitstat to see whether this model with exposure performs better than the model without:

```
. quietly fitstat, save
. quietly zip childs sex married sibs  born educ if reprage~=., inflate(sex
married sibs born educ)
```

Note: Here we limit the model without exposure only to those who don’t miss data on reprage variable.

```
. fitstat, dif
Measures of Fit for zip of childs
                               Current            Saved        Difference
Model:                             zip              zip
N:                                2734             2734                 0
Log-Lik Intercept Only       -4825.719        -4825.719             0.000
Log-Lik Full Model           -4509.577        -4334.455          -175.121
D                             9019.153(2722)   8668.911(2722)     350.243(0)
LR                             632.285(10)      982.528(10)       350.243(0)
Prob > LR                        0.000            0.000                 .
McFadden's R2                    0.066            0.102            -0.036
McFadden's Adj R2                0.063            0.099            -0.036
ML (Cox-Snell) R2                0.206            0.302            -0.095
Cragg-Uhler(Nagelkerke) R2       0.213            0.311            -0.098
AIC                              3.308            3.180             0.128
AIC*n                         9043.153         8692.911           350.243
BIC                         -12521.451       -12871.693           350.243
BIC'                          -553.150         -903.393           350.243
BIC used by Stata             9114.116         8763.873           350.243
AIC used by Stata             9043.153         8692.911           350.243
Difference of  350.243 in BIC' provides very strong support for saved model.
Note: p-value for difference in LR is only valid if models are nested.
```

We can see very strong support for the model with exposure.

The issue of diagnostics for zero-inflated models:
Unfortunately, many tests and work-around solutions that worked for nbreg and poisson don’t work for zip and zinb. One big problem is that zip and zinb cannot be modeled using GLM. We can still test for multicollinearity and use

robust option, but linearity diagnostics and those used to identify outliers and leverage points are not available here. One could test for those using regular poisson or nbreg and then see if suggested fixes (e.g., a transformation or omitted leverage points) appear to improve the corresponding zero-inflated model.

Zero-truncated models

Sometimes we have count data that have no zeros at all, because we only start accumulating data once at least one count was observed. For example, the length of hospital stay cannot be 0 because we only start observing counts once a person is admitted. In such cases, zero-truncated models, implemented by ztp and ztnb commands, are useful. E.g. say we only have data on the number of children after the person has their first one:

```
. gen childs0=childs
(5 missing values generated)
. replace childs0=. if childs==0
(799 real changes made, 799 to missing)
. ztp childs0 sex married sibs  born educ
Zero-truncated Poisson regression                 Number of obs   =       1951
                                                  LR chi2(5)      =     168.39
                                                  Prob > chi2     =     0.0000
Log likelihood = -3129.8812                       Pseudo R2       =     0.0262
------------------------------------------------------------------------------
     childs0 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .0050533   .0341538     0.15   0.882     -.061887    .0719936
     married |   .0439347   .0344268     1.28   0.202    -.0235405      .11141
        sibs |   .0283134   .0047432     5.97   0.000      .019017    .0376098
        born |  -.1934924   .0631899    -3.06   0.002    -.3173423   -.0696426
        educ |  -.0403873   .0055964    -7.22   0.000    -.0513561   -.0294186
       _cons |   1.406071   .1183233    11.88   0.000     1.174161     1.63798
------------------------------------------------------------------------------

. ztnb childs0 sex married sibs  born educ
Zero-truncated negative binomial regression       Number of obs   =       1951
                                                  LR chi2(5)      =     114.29
Dispersion     = mean                             Prob > chi2     =     0.0000
Log likelihood = -3128.9162                       Pseudo R2       =     0.0179
------------------------------------------------------------------------------
     childs0 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .0043327   .0352032     0.12   0.902    -.0646644    .0733297
     married |   .0440371   .0354945     1.24   0.215    -.0255309    .1136051
        sibs |   .0285975   .0049392     5.79   0.000     .0189169    .0382781
        born |  -.1951289   .0649357    -3.00   0.003    -.3224005   -.0678573
        educ |  -.0403866   .0057732    -7.00   0.000    -.0517018   -.0290714
       _cons |   1.398945   .1221116    11.46   0.000      1.15961    1.638279
-------------+----------------------------------------------------------------
    /lnalpha |  -3.811634   .7616972                     -5.304533   -2.318735
-------------+----------------------------------------------------------------
       alpha |    .022112   .0168427                      .004969     .098398
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) =    1.93 Prob>=chibar2 = 0.082
```

Note that the results of these models look very similar to those from the count equations of zero-inflated Poisson and NB models.

Examples of count data models:

Van der Burg, Brigitte, Jacques Siegers, and Rudolf Winter-Ebmer. 1998. Gender and Promotion in the Academic Labour Market. *Labour,* 12: 701-713.

Questions to answer about the article:
1. What are the dependent and the independent variables in this analysis?
2. What is reported in Table 1? How can we interpret these results? How do the authors discuss these results in the text?
3. What is presented in Table 2? How can we interpret these results?
4. In addition to what the authors chose to present, how else could they have presented their results?
5. What measures of model fit and model diagnostics are presented? What diagnostics and potential problems did the authors not address?

Sarkisian, Natalia and Naomi Gerstel. 2004. "Explaining the Gender Gap in Help to Parents: The Importance of Employment." *Journal of Marriage and the Family, 66*: 431-451.

Questions to answer about the article:
1. What are the dependent and the independent variables in this analysis?
2. What is reported in Table 1? How can we interpret these results? How do the authors discuss these results in the text?
3. In addition to what the authors chose to present, how else could they have presented their results?
4. What measures of model fit and model diagnostics are presented? What diagnostics and potential problems did the authors not address?