SOCY7704: Regression Models for Categorical Data
Instructor: Natasha Sarkisian

Count Data Models

<u>Negative Binomial Model</u>

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 model rarely fits due to overdispersion. One key process that often creates overdispersion 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 the conditional variance of our count variable. 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:

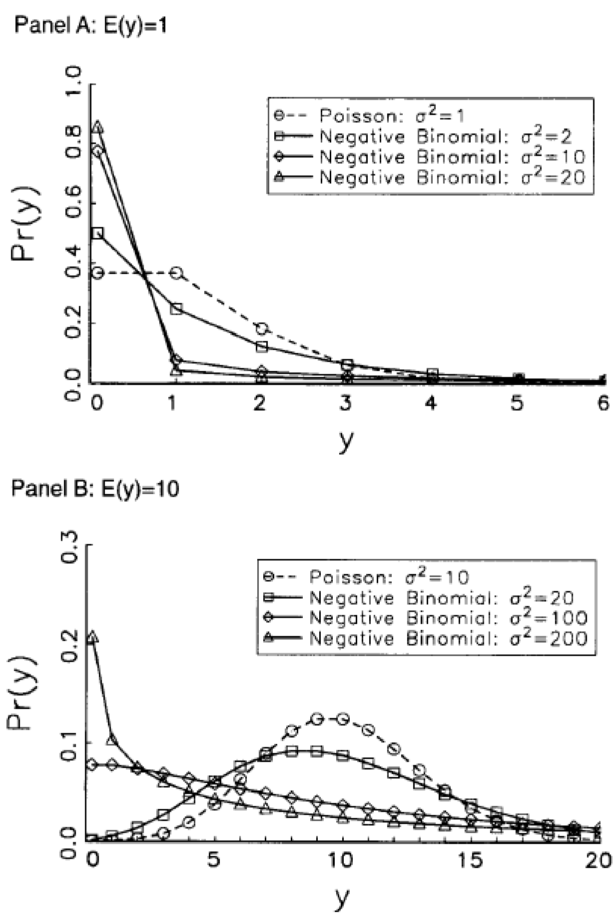


Figure 8.6. Comparisons of the Negative Binomial and Poisson Distributions

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

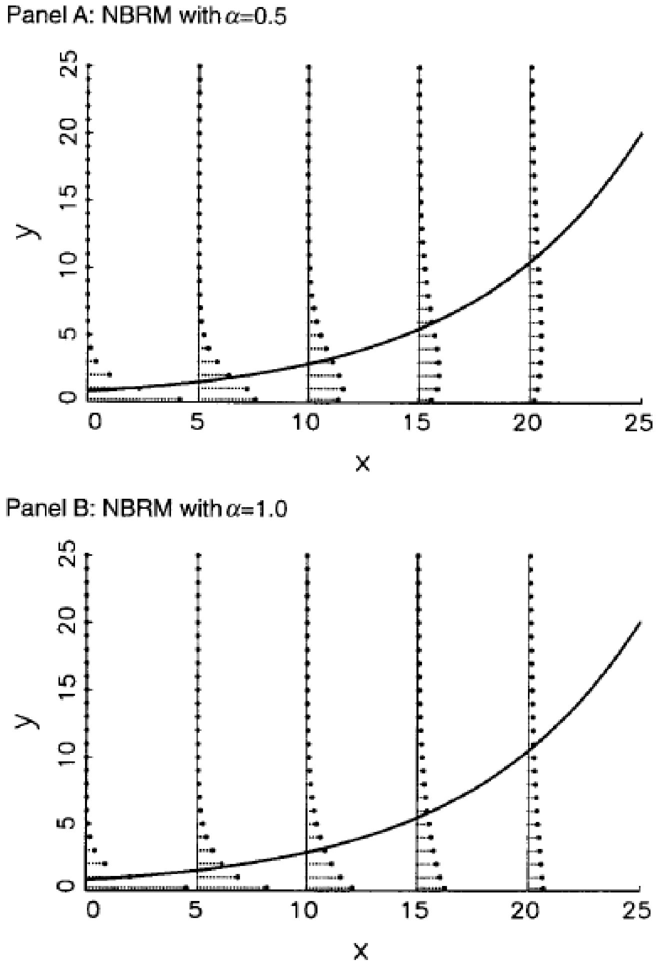


Figure 8.7. Distribution of Counts for the Negative Binomial Regression Model

Now let's run NB model for our data:

```
. nbreg childs sex married sibs  born educ
Fitting Poisson model:
Iteration 0:   log likelihood = -4784.5123
Iteration 1:   log likelihood = -4784.5079
Iteration 2:   log likelihood = -4784.5079
Fitting constant-only model:
Iteration 0:   log likelihood = -5023.5027
Iteration 1:   log likelihood = -4901.9594
Iteration 2:   log likelihood = -4901.9154
Iteration 3:   log likelihood = -4901.9154
Fitting full model:
Iteration 0:   log likelihood = -4732.0308
Iteration 1:   log likelihood =  -4712.421
Iteration 2:   log likelihood = -4711.6797
Iteration 3:   log likelihood = -4711.6789
Iteration 4:   log likelihood = -4711.6789

Negative binomial regression                      Number of obs   =       2745
                                                  LR chi2(5)      =     380.47
Dispersion     = mean                             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
```

Or better yet, we will estimate this model with robust standard errors – it is recommended that we use them with negative binomial model in case the variance is misspecified.

```
. nbreg childs sex married sibs  born educ, robust
Negative binomial regression                      Number of obs   =       2745
Dispersion           = mean                       Wald chi2(5)    =     386.44
Log pseudolikelihood = -4711.6789                 Prob > chi2     =     0.0000
------------------------------------------------------------------------------
             |               Robust
      childs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .2086278    .035025     5.96   0.000     .1399801    .2772755
     married |    .471206   .0348392    13.53   0.000     .4029225    .5394895
        sibs |   .0397041    .005216     7.61   0.000      .029481    .0499272
        born |  -.2231164   .0585515    -3.81   0.000    -.3378753   -.1083576
        educ |  -.0616831   .0060308   -10.23   0.000    -.0735032    -.049863
       _cons |   .9198597   .1225929     7.50   0.000     .6795821    1.160137
-------------+----------------------------------------------------------------
    /lnalpha |  -1.523939   .1167233                     -1.752712   -1.295165
-------------+----------------------------------------------------------------
       alpha |   .2178522   .0254284                      .1733033    .2738526
------------------------------------------------------------------------------
```

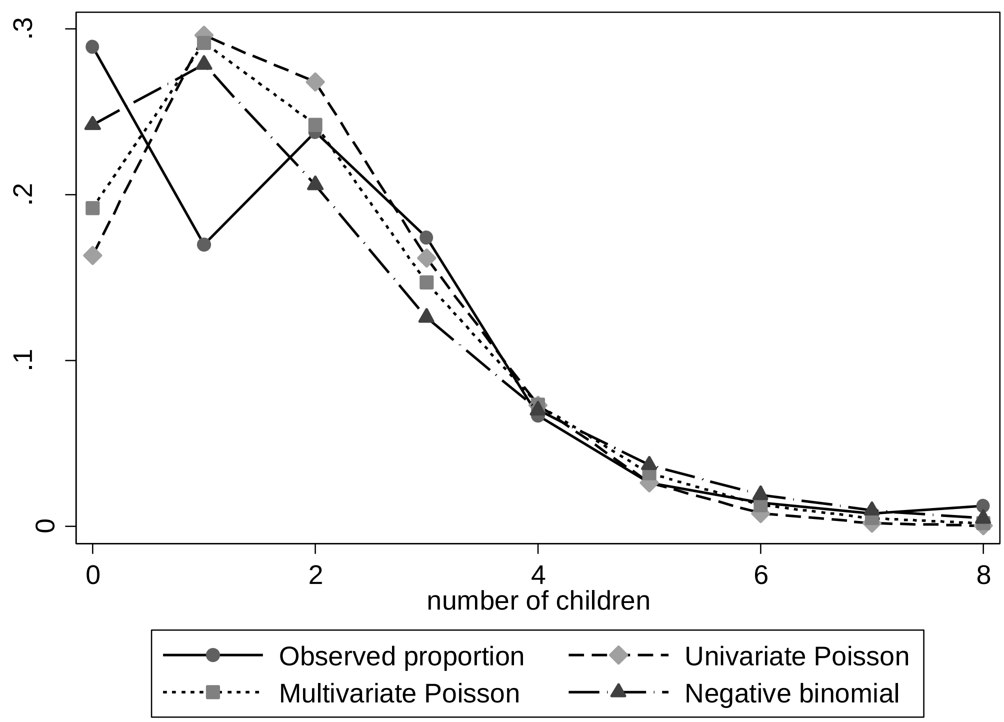
Interpretation of the results for negative binomial model is exactly the same as for Poisson model. But we have an extra line of output to interpret – the likelihood-ratio test. This allows us to see whether NB model should be used in place of regular Poisson. If probability is below the cutoff, it means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

```
. est store nbreg
. qui poisson childs sex married sibs  born educ
. est store poisson
. est table poisson nbreg, star b(%4.3f)
----------------------------------------
    Variable |  poisson       nbreg     
-------------+--------------------------
childs       |
         sex |   0.195***     0.209***
     married |   0.449***     0.471***
        sibs |   0.039***     0.040***
        born |  -0.221***    -0.223***
        educ |  -0.062***    -0.062***
       _cons |   0.955***     0.920***
-------------+--------------------------
lnalpha      |
       _cons |               -1.524***
----------------------------------------
legend: * p<0.05; ** p<0.01; *** p<0.001
```

Now let's compare their performance graphically:

```
. mgen, pr(0/8) meanpred stub(nb_)
Predictions from:
Variable   Obs Unique      Mean       Min       Max  Label
------------------------------------------------------------------------------------------
nb_val       9      9         4         0         8  number of children
nb_obeq      9      9  .1111111  .0080146  .2892532  Observed proportion
nb_oble      9      9  .7987047  .2892532         1  Observed cum. proportion
nb_preq      9      9  .1105054  .0049814  .2786995  Avg predicted Pr(y=#)
nb_prle      9      9  .7990764  .2423203  .9945486  Avg predicted cum. Pr(y=#)
nb_ob_pr     9      9  .0006057  -.108572  .0479451  Observed - Avg Pr(y=#)
------------------------------------------------------------------------------------------

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



The graph confirms the results of the alpha significance test: NB model does better than regular multivariate Poisson, especially with regard to dealing with 0s. But it still underpredicts zeros and overpredicts ones, and it underpredicts 2s and 3s (while Poisson was more on target). 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 negative binomial model performs better than Poisson. We can also compare them using BIC:

```
. qui poisson childs sex married sibs  born educ
. qui fitstat, save
. qui nbreg childs sex married sibs  born educ
. fitstat, diff
                         |     Current        Saved   Difference
-------------------------+---------------------------------------
Log-likelihood           |
                   Model |   -4711.679    -4784.508       72.829
          Intercept-only |   -4901.915    -5070.839      168.924
```

```
-------------------------+--------------------------------------
Chi-square               |
     D (df=2738/2739/-1) |    9423.358     9569.016     -145.658
         Wald (df=5/5/0) |     386.441            .            .
                 p-value |       0.000        0.000            .
-------------------------+--------------------------------------
R2                       |
                McFadden |       0.039        0.056       -0.018
     McFadden (adjusted) |       0.037        0.055       -0.018
            Cox-Snell/ML |       0.129        0.188       -0.059
  Cragg-Uhler/Nagelkerke |       0.133        0.193       -0.060
-------------------------+--------------------------------------
IC                       |
                     AIC |    9437.358     9581.016     -143.658
      AIC divided by N   |       3.438        3.490       -0.052
         BIC (df=7/6/1)  |    9478.781     9616.521     -137.740
Note: Some measures based on pseudolikelihoods.
Difference of  137.740 in BIC provides very strong support for current model.
```

The interpretation tools for nbreg are the same as for Poisson; we can get IRR and use mtable, mchange, and mgen commands. 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 (i.e., allow for different degree of overdispersion for different groups):

```
. 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 (it is especially pronounced for men, unmarried people, those with fewer siblings, and those with more education).

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. 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 will have children eventually, but others do not have kids and cannot have them anymore or do not want to, so their count will always remain zero. But these two groups are latent – no information on their fertility situation or preferences. 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, you 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
------------------------------------------------------------------------------
```

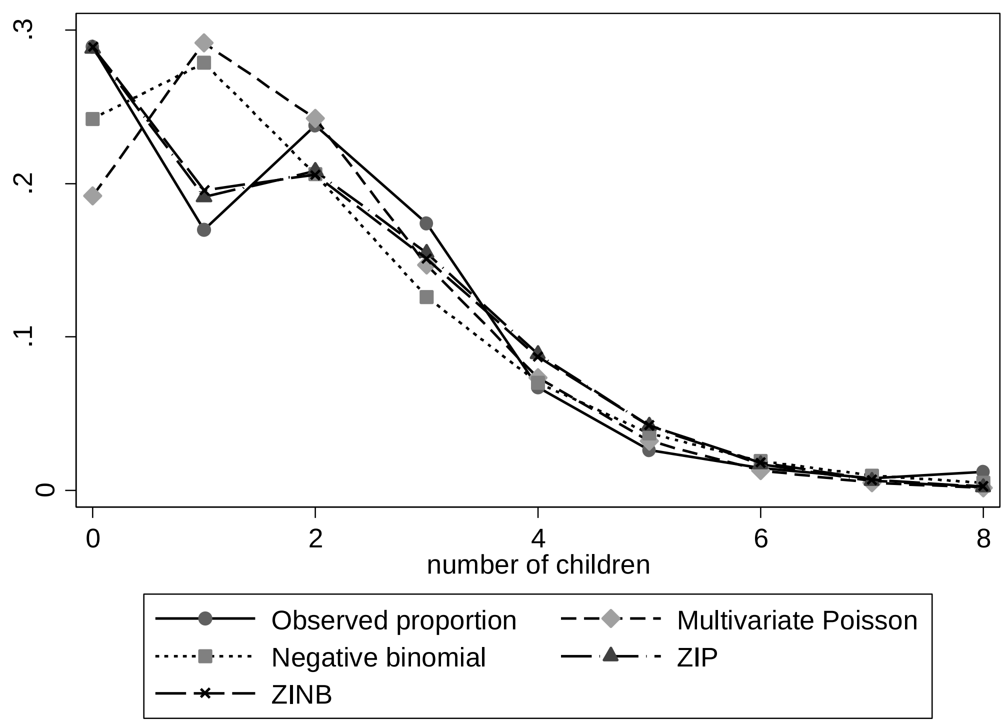
Note the inflate option we specified – we have to specify that option, it tells Stata what variables to use to predict the membership in "Always Zero" group. In this case, we used the same variables but we could have used a smaller subset of the variables or even different variables altogether. We'll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

```
. mgen, pr(0/8) meanpred stub(zip_)
Predictions from:
Variable   Obs Unique      Mean       Min       Max  Label
------------------------------------------------------------------------------------
zip_val      9      9         4         0         8  number of children
zip_obeq     9      9  .1111111  .0080146  .2892532  Observed proportion
zip_oble     9      9  .7987047  .2892532         1  Observed cum. proportion
zip_preq     9      9  .1109995  .0021302  .2880608  Avg predicted Pr(y=#)
zip_prle     9      9  .7987461  .2880608  .9989958  Avg predicted cum. Pr(y=#)
zip_ob_pr    9      9  .0001116  -.021445  .0296168  Observed - Avg Pr(y=#)
------------------------------------------------------------------------------------
. lab var zip_preq "ZIP"
```

We will also estimate a zero-inflated negative binomial model and then compare all of them.

```
. 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
------------------------------------------------------------------------------
```

```
. mgen, pr(0/8) meanpred stub(zinb_)
Predictions from:
Variable    Obs Unique      Mean       Min       Max  Label
---------------------------------------------------------------------------------------
zinb_val      9      9         4         0         8  number of children
zinb_obeq     9      9  .1111111  .0080146  .2892532  Observed proportion
zinb_oble     9      9  .7987047  .2892532         1  Observed cum. proportion
zinb_preq     9      9  .1109602  .0025516   .288929  Avg predicted Pr(y=#)
zinb_prle     9      9   .798788   .288929  .9986414  Avg predicted cum. Pr(y=#)
zinb_ob_pr    9      9   .000151 -.0256162  .0320836  Observed - Avg Pr(y=#)
---------------------------------------------------------------------------------------
. lab var zinb_preq "ZINB"
. graph twoway connected poi_obeq mpoi_preq nb_preq zip_preq zinb_preq poi_val, 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
                        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 compl~d  |     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
                              |
                      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 compl~d |                            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
```

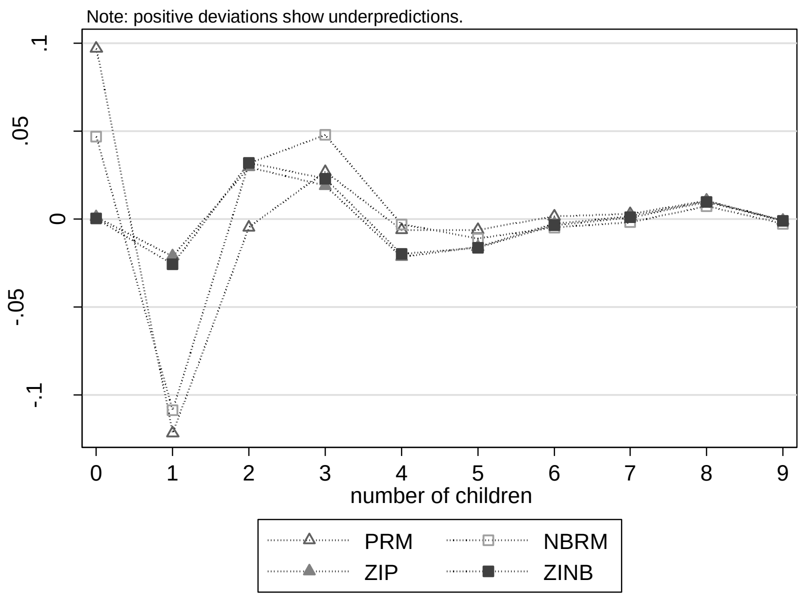
```
5        0.026       0.037      0.011     8.820
6        0.015       0.019      0.005     3.010
7        0.008       0.010      0.002     0.867
8        0.012       0.005      0.007    30.214
9        0.000       0.003      0.003     7.016
------------------------------------------------
Sum      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=  9616.521  AIC=  9581.016  Prefer  Over  Evidence
-------------------------------------------------------------------------
  vs NBRM      BIC=  9478.781  dif=   137.740  NBRM    PRM   Very strong
               AIC=  9437.358  dif=   143.658  NBRM    PRM
               LRX2=  145.658  prob=    0.000  NBRM    PRM   p=0.000
-------------------------------------------------------------------------
  vs ZIP       BIC=  9143.394  dif=   473.127  ZIP     PRM   Very strong
               AIC=  9072.383  dif=   508.632  ZIP     PRM
               Vuong=  11.165  prob=    0.000  ZIP     PRM   p=0.000
-------------------------------------------------------------------------
  vs ZINB      BIC=  9148.749  dif=   467.772  ZINB    PRM   Very strong
               AIC=  9071.821  dif=   509.195  ZINB    PRM
-------------------------------------------------------------------------
NBRM           BIC=  9478.781  AIC=  9437.358  Prefer  Over  Evidence
-------------------------------------------------------------------------
  vs ZIP       BIC=  9143.394  dif=   335.387  ZIP     NBRM  Very strong
               AIC=  9072.383  dif=   364.974  ZIP     NBRM
-------------------------------------------------------------------------
  vs ZINB      BIC=  9148.749  dif=   330.032  ZINB    NBRM  Very strong
               AIC=  9071.821  dif=   365.537  ZINB    NBRM
               Vuong=  10.441  prob=    0.000  ZINB    NBRM  p=0.000
-------------------------------------------------------------------------
```

```
ZIP            BIC=  9143.394  AIC=  9072.383  Prefer  Over  Evidence
-------------------------------------------------------------------------
  vs ZINB      BIC=  9148.749  dif=    -5.355  ZIP     ZINB  Positive
               AIC=  9071.821  dif=     0.563  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 less 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 to see IRR (but irr option is also available for zip and zinb commands themselves):

```
. 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 number of kids 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, predicted probabilities of specific counts, and changes in both rates and probabilities using mtable, mchange, and mgen. Predicted rates for by born and sex for married people:

```
. zip childs i.sex i.married sibs  i.born educ, inflate(i.sex i.married sibs i.born
educ)
. mtable, at(sex=(1 2) born=(1 2) married==1) atmeans stat(ci)
Expression: Predicted number of childs, predict()
          |      sex      born        mu        ll        ul
----------+--------------------------------------------------
        1 |        1         1     2.215     2.102     2.328
        2 |        1         2     1.849     1.645     2.053
        3 |        2         1     2.253     2.142     2.364
        4 |        2         2     1.891     1.684     2.099
Specified values of covariates
          |  married      sibs      educ
----------+------------------------------
  Current |        1       3.6      13.4
```

Changes in predicted rates as well as marginal effects:

```
. mchange, amount(all)
zip: Changes in mu | Number of obs = 2745
Expression: Predicted number of childs, predict()
                |    Change    p-value
----------------+----------------------
sex             |
 female vs male |     0.332      0.000
married         |
         1 vs 0 |     0.801      0.000
sibs            |
         0 to 1 |     0.068      0.000
             +1 |     0.076      0.000
            +SD |     0.235      0.000
          Range |     2.547      0.000
       Marginal |     0.075      0.000
born            |
      no vs yes |    -0.361      0.000
educ            |
         0 to 1 |    -0.153      0.000
             +1 |    -0.108      0.000
            +SD |    -0.310      0.000
          Range |    -2.411      0.000
       Marginal |    -0.110      0.000

Average prediction
    1.812
```

We interpret these results the same way as for regular Poisson model. Discrete changes and marginal effects are particularly useful in zero-inflated models because they combine the two equations to calculate the overall impact of each variable on the expected count. I would recommend presenting marginal effects (average ones or at means) along with two sets of exponentiated coefficients (IRR and OR) when reporting the results of zero-inflated models.

We can also examine predicted probabilities of counts:

```
. mtable, at(sex=(1 2) born=(1 2) married==1) atmeans pr(0/4)
Expression: Pr(childs), predict(pr())
           |      sex      born      none       one       two     three      four
 ----------+----------------------------------------------------------------------
         1 |        1         1     0.123     0.230     0.261     0.197     0.111
         2 |        1         2     0.174     0.275     0.262     0.166     0.079
         3 |        2         1     0.109     0.233     0.265     0.200     0.113
         4 |        2         2     0.156     0.281     0.268     0.170     0.081
Specified values of covariates
           |  married      sibs      educ
 ----------+-----------------------------
   Current |        1       3.6      13.4
```

And changes in probabilities of counts:

```
. mchange, amount(all) pr(0/4)
zip: Changes in PrAny0 | Number of obs = 2745
Expression: Pr(childs = any 0), predict(pr(0))
                |          0          1          2          3          4
----------------+-------------------------------------------------------
sex             |
 female vs male |     -0.135      0.038      0.040      0.029      0.016
        p-value |      0.000      0.000      0.000      0.000      0.000
married         |
         1 vs 0 |     -0.314      0.084      0.092      0.069      0.040
        p-value |      0.000      0.000      0.000      0.000      0.000
sibs            |
         0 to 1 |     -0.016     -0.003      0.003      0.006      0.005
        p-value |      0.000      0.046      0.006      0.000      0.000
             +1 |     -0.014     -0.004      0.001      0.005      0.005
        p-value |      0.000      0.003      0.249      0.000      0.000
            +SD |     -0.042     -0.013      0.002      0.014      0.016
        p-value |      0.000      0.001      0.529      0.000      0.000
          Range |     -0.282     -0.145     -0.079      0.035      0.111
        p-value |      0.000      0.000      0.007      0.094      0.000
       Marginal |     -0.015     -0.004      0.001      0.005      0.005
        p-value |      0.000      0.005      0.160      0.000      0.000
born            |
      no vs yes |      0.067      0.026     -0.007     -0.028     -0.027
        p-value |      0.014      0.100      0.444      0.001      0.000
educ            |
         0 to 1 |      0.009      0.009      0.009      0.003     -0.004
        p-value |      0.000      0.000      0.000      0.141      0.001
             +1 |      0.024      0.003     -0.004     -0.008     -0.007
        p-value |      0.000      0.019      0.000      0.000      0.000
            +SD |      0.074      0.008     -0.013     -0.024     -0.021
        p-value |      0.000      0.066      0.000      0.000      0.000
          Range |      0.399      0.109      0.007     -0.100     -0.139
        p-value |      0.000      0.000      0.728      0.000      0.000
       Marginal |      0.024      0.004     -0.003     -0.008     -0.007
        p-value |      0.000      0.009      0.000      0.000      0.000

Average predictions

             |          0          1          2          3          4
-------------+-------------------------------------------------------
  Pr(y|base) |      0.288      0.191      0.208      0.155      0.089
```

We can also use mgen to make all kinds of graphs for predicted rates and probabilities of counts and changes in these, like we did for regular Poisson.

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, and neither is foreign born status! Looks like the effects of those were just function of age. Gender, siblings, and education predict the count, and gender 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, diff
                         |     Current        Saved   Difference
-------------------------+---------------------------------------
Log-likelihood           |
                   Model |   -4509.577    -4334.455     -175.121
          Intercept-only |   -4825.719    -4825.719        0.000
-------------------------+---------------------------------------
Chi-square               |
      D (df=2722/2722/0) |    9019.153     8668.911      350.243
         LR (df=10/10/0) |     632.285      982.528     -350.243
                 p-value |       0.000        0.000            .
-------------------------+---------------------------------------
R2                       |
```

```
                McFadden |       0.066        0.102       -0.036
     McFadden (adjusted) |       0.063        0.099       -0.036
            Cox-Snell/ML |       0.206        0.302       -0.095
  Cragg-Uhler/Nagelkerke |       0.213        0.311       -0.098
-------------------------+-------------------------------------
IC                       |
                     AIC |    9043.153     8692.911      350.243
       AIC divided by N  |       3.308        3.180        0.128
       BIC (df=12/12/0)  |    9114.116     8763.873      350.243

Difference of  350.243 in BIC provides very strong support for saved model.
```

We can see very strong support for the model with exposure, so we would select it as our final one.

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 for robust SE, but linearity diagnostics and those used to identify outliers and leverage points are not available here. So the strategy to use is:

1. Do the diagnostics 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.
2. Generate a dichotomy for 0 vs non-zero, run logit for that, and do diagnostics for logit as well (that would approximate the "Always zero" equation of ZIP and ZINB, and it is possible, for example, for a nonlinear relationship to exist in predicting counts but not predicting zeroes, or other way around).

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 zero-inflated NB models.