UHCDA15 A3: Count

Your Name: Points received: ____ out of 125

In this assignment, you will answer questions about different models specified for a count model.

My dependent variables is pub3, or the number of publications scholars have produced three years after their Ph.D. It's distribution looks as follows:

```
. tab pub3, m

Publication |
s: PhD yr 1 |
      to 3. |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         57       20.50       20.50
          1 |         56       20.14       40.65
          2 |         47       16.91       57.55
          3 |         28       10.07       67.63
          4 |         26        9.35       76.98
          5 |         23        8.27       85.25
          6 |          8        2.88       88.13
          7 |         10        3.60       91.73
          8 |          8        2.88       94.60
          9 |          3        1.08       95.68
         10 |          5        1.80       97.48
         11 |          1        0.36       97.84
         12 |          2        0.72       98.56
         13 |          1        0.36       98.92
         15 |          2        0.72       99.64
         27 |          1        0.36      100.00
------------+-----------------------------------
      Total |        278      100.00
```

I decide to predict this variable using four independent variables: an indicator for gender, an indicator for whether the scholar held a fellowship that allowed for protected research time, a continuous measures of years enrolled in their Ph.D. program, and a measure of Ph.D. program prestige. All five variables are summarized below:

```
. codebook pub3 female fellow enroll phd, compact

Variable   Obs Unique      Mean  Min   Max  Label
-------------------------------------------------------------------------------
pub3       278     16  2.938849    0    27  Publications: PhD yr 1 to 3.
female     278      2  .3453237    0     1  Female? (1=yes)
fellow     278      2  .4064748    0     1  Postdoctoral fellow? (1=yes)
enroll     278      9  5.564748    3    14  Years from BA to PhD.
phd        278     80  3.166331    1  4.66  Prestige of Ph.D. department.
-------------------------------------------------------------------------------

. sum pub3 female fellow enroll phd

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        pub3 |        278    2.938849    3.222193          0         27
      female |        278    .3453237    .4763312          0          1
      fellow |        278    .4064748     .492061          0          1
      enroll |        278    5.564748    1.467253          3         14
         phd |        278    3.166331    .9961592          1       4.66
```

1. ___ of 5: What information might I use from my summary statistics to inform my selection of poisson vs. negative binomial regression model? Why might this also be misleading? (Hint: Relate the second answer to differences in observed heterogeneity.)

Next, I estimate a poisson regression model (PRM) of gender, fellowship, enrollment time Ph.D. prestige on pub3.

```
.  poisson pub3 i.female i.fellow enroll phd, nolog

Poisson regression                              Number of obs     =        278
                                                LR chi2(4)        =      93.51
                                                Prob > chi2       =     0.0000
Log likelihood = -702.03001                     Pseudo R2         =     0.0624

------------------------------------------------------------------------------
        pub3 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |
      1_Yes  |  -.2184532   .0812049    -2.69   0.007    -.3776119   -.0592944
             |
      fellow |
      1_Yes  |   .2983762   .0735101     4.06   0.000     .1542992    .4424533
      enroll |  -.1807045   .0276827    -6.53   0.000    -.2349616   -.1264474
         phd |   .0824519   .0364198     2.26   0.024     .0110704    .1538333
       _cons |   1.716022    .193717     8.86   0.000     1.336344      2.0957
------------------------------------------------------------------------------
```

2. ___ of 5: If overdispersion exists in my data, even after I've accounted for observed heterogeneity with my X variables, which of following would I expect to see in my observed vs. predicted values under the PRM:
    a. MORE/FEWER zeroes
    b. MORE/FEWER cases in the middle of the data (near the mean)
    c. MORE/FEWER cases in the upper tail
3. ___ of 5: If overdispersion exists in my data, even after I've accounted for observed heterogeneity with my X variables, will the z-scores for my PRM model be (choose one) SPURIOUSLY SMALL or SPURIOUSLY LARGE?

Next I decide to check for overdispersion in my model by estimating the negative binomial regression model (NBRM):

```
. nbreg pub3 i.female i.fellow enroll phd, nolog

Negative binomial regression                    Number of obs     =        278
                                                LR chi2(4)        =      32.24
Dispersion     = mean                           Prob > chi2       =     0.0000
Log likelihood = -601.93588                     Pseudo R2         =     0.0261

------------------------------------------------------------------------------
        pub3 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |
      1_Yes  |  -.1982677   .1345457    -1.47   0.141    -.4619725    .0654371
             |
      fellow |
      1_Yes  |   .2942794    .127585     2.31   0.021     .0442174    .5443414
      enroll |  -.1905708   .0449627    -4.24   0.000    -.2786961   -.1024455
```

```
         phd |   .0781415   .0605418     1.29   0.197    -.0405182    .1968012
       _cons |   1.777913   .3199428     5.56   0.000     1.150836    2.404989
-------------+----------------------------------------------------------------
    /lnalpha |  -.4864692   .1453169                     -.7712851   -.2016533
-------------+----------------------------------------------------------------
       alpha |   .6147933   .0893398                      .4624184    .8173782
------------------------------------------------------------------------------
LR test of alpha=0: chibar2(01) = 200.19              Prob >= chibar2 = 0.000
```

4. ___ of 10: Imagine I used my output from my PRM & NBRM models to create comparing the effects and significance of X variables. The table includes the following information: Column 1=Name of variable; Column 2=PRM unstandardized coefficient; Column 3=NBRM unstandardized coefficient; Column 4=ratio of Col2 and Col3; Column 5=PRM z-value; Column 6=NBRM z-value; Column 7=ratio of Col5 and Col6.
    a. Why would you expect the ratio in Column 4 to be close to 1.0? Relate this to the mean structure of the two models.
    b. What do you expect for Column 7? Relate what you find to your answer to question 3 and the effects of ignoring unobserved heterogeneity.

5. ___ of 5: Using the NBRM model results, test the NBRM against the alternative of the PRM. Write up the result as though it were part of a research paper. Include information on statistical significance.

Since my NBRM model output indicates evidence of overdispersion, I opt to use the NBRM instead of the PRM for my continuing work. Next, I compute my factor change coefficients from my NBRM model:

```
. listcoef, help

nbreg (N=278): Factor change in expected count

  Observed SD:  3.2222

------------------------------------------------------------------------------
             |         b        z    P>|z|       e^b   e^bStdX     SDofX
-------------+----------------------------------------------------------------
      female |
      1_Yes  |   -0.1983   -1.474    0.141     0.820     0.910     0.476
             |
      fellow |
      1_Yes  |    0.2943    2.307    0.021     1.342     1.156     0.492
      enroll |   -0.1906   -4.238    0.000     0.826     0.756     1.467
         phd |    0.0781    1.291    0.197     1.081     1.081     0.996
    constant |    1.7779    5.557    0.000         .         .         .
-------------+----------------------------------------------------------------
alpha        |
     lnalpha |   -0.4865        .        .         .         .         .
       alpha |    0.6148        .        .         .         .         .
------------------------------------------------------------------------------
  LR test of alpha=0: 200.19   Prob>=LRX2 = 0.000
       b = raw coefficient
       z = z-score for test of b=0
   P>|z| = p-value for z-test
     e^b = exp(b) = factor change in expected count for unit increase in X
 e^bStdX = exp(b*SD of X) = change in expected count for SD increase in X
   SDofX = standard deviation of X
```

6. ___ of 15: Interpret the <u>standardized</u> factor change coefficient for enroll and the <u>unstandardized</u> factor change coefficients for enroll and female. Include information on statistical significance. This should read as though it were part of a published article.

Next, I decide to run some post-estimation commands. First, I use `mchange` to look at discrete change coefficients:

```
. mchange, atmeans

nbreg: Changes in mu | Number of obs = 278

Expression: Predicted number of pub3, predict()

               |    Change    p-value
---------------+----------------------
female         |
 1 Yes vs 0 No |    -0.532      0.130
fellow         |
 1 Yes vs 0 No |     0.839      0.026
enroll         |
       +1 cntr |    -0.527      0.000
      +SD cntr |    -0.775      0.000
      Marginal |    -0.526      0.000
phd            |
       +1 cntr |     0.216      0.197
      +SD cntr |     0.215      0.197
      Marginal |     0.216      0.197
```

I also use `mtable` to look at the discrete change in fellow:

```
. mtable, dydx(fellow) atmeans pr(0(1)6) stat(est p) roweq(Diff) below

Expression: Marginal effect of Pr(pub), predict(pr(6))

           |        0         1         2         3         4         5         6
 ----------+-------------------------------------------------------------------------
 NoFel     |
         1 |    0.224     0.219     0.173     0.126     0.087     0.059     0.039
 Fel       |
         1 |    0.165     0.180     0.158     0.128     0.099     0.075     0.055
 Diff      |
   d Pr(y) |   -0.059    -0.039    -0.015     0.002     0.012     0.015     0.016
         p |    0.020     0.024     0.040     0.371     0.028     0.023     0.022
```

7. ___ of 10: Using the output from `mchange`, interpret a discrete change coefficient for enroll. Include information about significance.
8. ___ of 10: Using the output from `mtable`, explain the effects of having a fellowship on different predicted probabilities.
9. ___ of 10: How does the information in the `mchange` output differ from the information in the `mtable` output? When might I want to make use of each set of information?

Next, I estimate my ZINB model, to see if a zero-inflated specification fits my data better.

```
. zinb pub3 i.female i.fellow enroll phd, inf(phd) nolog vuong

Zero-inflated negative binomial regression      Number of obs     =        278
                                                Nonzero obs       =        221
                                                Zero obs          =         57

Inflation model = logit                         LR chi2(4)        =      31.27
Log likelihood  = -601.8815                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
        pub3 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
pub3          |
      female  |
      1_Yes   |  -.2004393   .1343632    -1.49   0.136    -.4637865    .0629078
              |
      fellow  |
      1_Yes   |   .2932775   .1274693     2.30   0.021     .0434424    .5431127
      enroll  |  -.1895663    .044968    -4.22   0.000    -.2777019   -.1014306
         phd  |   .0697857   .0652696     1.07   0.285    -.0581405    .1977118
       _cons  |   1.808118   .3321055     5.44   0.000     1.157204    2.459033
-------------+------------------------------------------------------------------
inflate      |
         phd |  -1.384685   2.597276    -0.53   0.594    -6.475252    3.705883
       _cons |  -1.278282   5.016271    -0.25   0.799    -11.10999    8.553428
-------------+------------------------------------------------------------------
    /lnalpha |  -.5115684   .1693917    -3.02   0.003    -.8435701   -.1795667
-------------+------------------------------------------------------------------
       alpha |   .5995545   .1015596                      .430172    .8356322
------------------------------------------------------------------------------
Vuong test of zinb vs. standard negative binomial: z =     0.16  Pr>z = 0.4352
```

10. ___ of 10: The command above indicates that I have included the variable `phd` in the `inflate` portion of my model. What does this mean? In general terms, how should I interpret the coefficient for `phd` in the `inflate` portion of my model? How does this differ from the interpretation of `phd` in the count (upper) portion of my model?
11. ___ of 10: Using the output from the ZINB, test the ZINB against the alternative of the ZIP. Write up the result as though it were part of a research paper. Include information on statistical significance.
12. ___ of 10: Using the output from the ZINB, test the ZINB against the alternative of the NBRM. Write up the result as though it were part of a research paper. Include information on statistical significance.
13. ___ of 10: Comparing across the PRM, NBRM, ZIP & ZINB models, which models seems to be preferred?
14. ___ of 10: My assessment of the overall effectiveness of your answers.