**Methods in Epidemiologic Research**

**Sample Problems**

**Chapter 18 - Modelling Count Data**

## Preparation

As we indicated in the sample problems for Chapter 15, we are going to carry on with the -mi- dataset but now we will use Poisson and negative binomial models to evaluate how various factors influence the length of hospital stay (-los-).

The variables we will use in this exercise are listed below. The outcome will be -los-. Most of the variables listed have already been defined. -age_inv- and -bmi_ct- are transformed versions of -age- and -bmi-, respectively (see below).

```
Contains data from C:\mer\data\mi.dta
  obs:         2,965
 vars:            13                          28 Feb 2012 10:53
 size:       112,670
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
id              float   %9.0g                 patient id
los             int     %8.0g                 length of hospital stay
sex             byte    %8.0g                 gender
age             float   %9.0g                 age at admission
age_inv         float   %9.0g                 transformed and centred age (1/age)-0.015
white           float   %9.0g                 race=White
mar_c2          float   %9.0g                 married Y/N
bmi             float   %9.0g                 body mass index
bmi_ct          float   %9.0g                 centred bmi (bmi - 28)
prmi            byte    %8.0g                 previous MI
card            byte    %8.0g                 cardiac arrest during hospitalization
cabg            byte    %8.0g                 coronary artery bypass surgery
died_hosp       float   %9.0g                 died in hospital
-------------------------------------------------------------------------------
Sorted by:
```

**Note:** It is important to note that, for the purpose of this exercise, we are ignoring the possible clustering of lengths of stay within hospital (*ie* some hospitals may have, on average, longer stays than others). We will evaluate the impact of this in the exercises for Chapter 22.

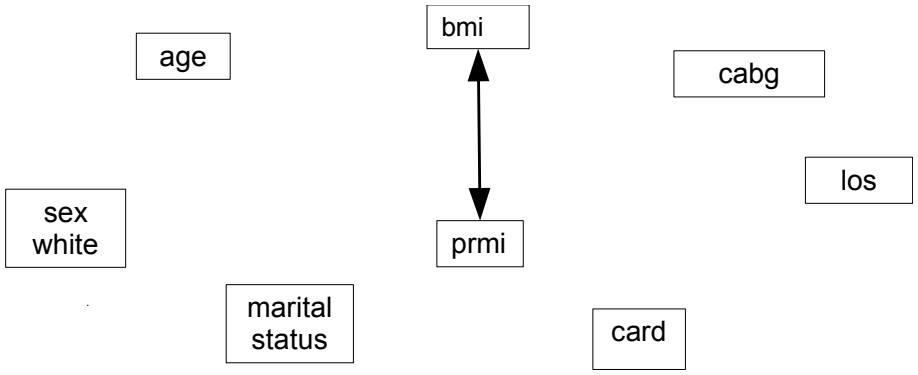
## Questions

Your primary interest is how marital status (married vs no-married) (-mar_c2-) and body mass index (-bmi-) influence the length of stay. However, we also investigate the role of other factors.

1. Draw a causal diagram incorporating all of the predictors listed above.

   *In the following diagram, lines have been left off for clarity (this is the same diagram used in the problems for Chapter 15). However it was assumed that factors would influence all factors to the right of them (*eg *age would influence marital status, bmi, prmi, card, cabg and los). The exception is the relationship between -bmi- and -prmi- in that either may have influenced the other (*ie *bmi might have influenced the risk of an MI, but a previous MI might also influence bmi) ... so a*

*bidirectional arrow was included.*




2. Is there any evidence that the effect of age is not linear? (You should revisit what was done in Chapter 15 to address this question).

   (a) create quadratic terms and add them to the model

   *The output from this exercise is not shown, but it is clear that the quadratic term for age is significant, suggesting that the relationship is not linear. However, in order to avoid having to use 2 terms for age in the model, fractional polynomials were used to determine but the best single term transformation. This turns out to be 1/age, which we also centred. This new variable -age_inv- is provided in the data set -mi_cnt-*

3. Poisson model

   (a) Rather than going through a full model building exercise, we will start with the final model from the sample problems for Chapter 15. (-los- as outcome and the following as predictors: -sex-, -age_inv-, -white-, -mar_c2-, -bmi_ct-, -prmi-)

```
. poisson los sex age_inv white mar_c2 bmi_ct prmi

Iteration 0:   log likelihood = -8750.2674
Iteration 1:   log likelihood = -8750.2673
Poisson regression                                Number of obs   =       2629
                                                  LR chi2(6)      =     260.42
                                                  Prob > chi2     =     0.0000
Log likelihood = -8750.2673                       Pseudo R2       =     0.0147
------------------------------------------------------------------------------
         los |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .0117129   .0181274     0.65   0.518    -.0238163     .047242
     age_inv |  -37.03328    2.72905   -13.57   0.000    -42.38212   -31.68444
       white |   -.092017   .0251536    -3.66   0.000    -.1413171   -.0427169
      mar_c2 |  -.0653547   .0176532    -3.70   0.000    -.0999543   -.0307551
      bmi_ct |   .0093158   .0013842     6.73   0.000     .0066027    .0120288
        prmi |   .0316093   .0185608     1.70   0.089    -.0047692    .0679878
       _cons |   1.778887   .0261279    68.08   0.000     1.727677    1.830097
------------------------------------------------------------------------------
```

   *While -sex- is not significant, we will retain it as a potential confounder.*

   (b) Compute the expected number of days in hospital for: a "baseline" individual (female, age=67, non-white, not married, bmi=28, no previous mi); the same except married; a "long stay"

individual (age~=85 (age_inv=-0.003), bmi=48 (bmi_ct=20) and had a previous mi)

```
.* baseline (female, age=67, non-white, not married, bmi=28, no prev. mi)
. display exp(_b[_cons])
5.923261

.* baseline but married
. display exp(_b[_cons] + _b[mar])
5.5485266

. * longest stay individuals (age~=85 (age_inv=-0.003), bmi=48 (bmi_ct=20), and
had a prev. mi)
. display exp(_b[_cons] - 0.003*_b[age_inv] + 20*_b[bmi_ct] +_b
> [prmi])
8.2310322
```

*The expected length of stay for these three types of individual were 5.9, 5.5 and 8.2 days*

(c) Fit the same Poisson regression model in the GLM framework

*The original (maximum likelihood) and GLM models will not be shown, but here are the commands for fitting the models (and saving the results), and then a table comparing the two sets of results*

```
. poisson los sex age_inv white mar_c2 bmi_ct prmi
.   estimates store pois_ml
. glm los sex age_inv white mar_c2 bmi_ct prmi, fam(poisson) link(log)
.   estimates store pois_glm
. estimates table pois_ml pois_glm

----------------------------------------
    Variable |  pois_ml      pois_glm
-------------+--------------------------
         sex |  .01171289    .01171289
     age_inv | -37.033283   -37.033285
       white | -.09201696   -.09201696
      mar_c2 | -.06535473   -.06535473
      bmi_ct |  .00931576    .00931576
        prmi |  .03160932    .03160932
       _cons |  1.7788871    1.7788871
----------------------------------------
```

*As you can see, the results are identical (with the exception of one small change in the 6th decimal place)*

(d) Express the same model .... but in terms of count ratios (called incidence rate ratios (IRR) in the statistical output, because the data are quite often incidence data)

```
.glm los sex age_inv white mar_c2 bmi_ct prmi, fam(poisson) link(log) eform

Generalized linear models                         No. of obs      =      2629
Optimization     : ML                             Residual df     =      2622
                                                  Scale parameter =         1
Deviance         =  8872.443582                   (1/df) Deviance =  3.383846
Pearson          =  14201.29271                   (1/df) Pearson  =  5.416206

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]
```

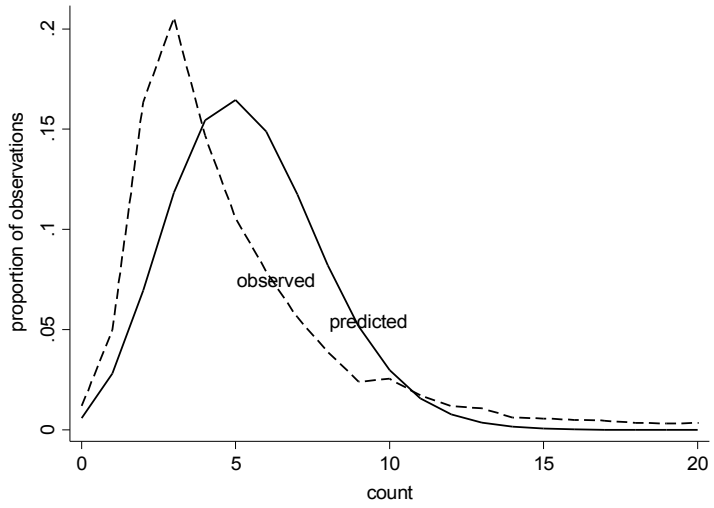
```
                                                AIC             =  6.662052
Log likelihood   = -8750.267314                 BIC             = -11774.13
------------------------------------------------------------------------------
             |                 OIM
         los |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   1.011782    .018341     0.65   0.518     .9764651    1.048376
     age_inv |   8.25e-17   2.25e-16   -13.57   0.000     3.92e-19    1.74e-14
       white |   .9120897   .0229423    -3.66   0.000      .868214    .9581826
      mar_c2 |   .9367351   .0165364    -3.70   0.000     .9048787     .969713
      bmi_ct |   1.009359   .0013972     6.73   0.000     1.006625    1.012101
        prmi |   1.032114   .0191569     1.70   0.089     .9952422    1.070352
       _cons |   5.923261   .1547623    68.08   0.000     5.627568     6.23449
------------------------------------------------------------------------------
```

*Most of the count ratios are fairly close to 1, but remember that these represent the expected change with a one unit change in the predictor. For bmi_ct, this is just a single point change in bmi. The coefficient for age_inv is very close to zero, but the range of values for age_inv is only from -0.005 to +0.021 (much less than a 1 unit change).(Note: it might have been better to rescale age (by dividing by 10 before transforming) so that we were dealing with 10 year changes in age rather than 1 year changes).*

4. Poisson model diagnostics - overall model fit

(a) Compare the observed and predicted counts of days in hospital.

*Generating this graph was done in Stata, but required an "add-on" program called -spost-.*




*There were clearly more low values of -los- and more high values of -los- than were predicted by the model. This suggests limited predictive ability for the model (ie unable to predict which stays were going to be short and which ones long ... so it put most of them in the middle).*

(b) Compute both deviance and Pearson residual based goodness-of-fit statistics (dispersion tests)

```
. estat gof
        Deviance goodness-of-fit =  8872.445
        Prob > chi2(2622)        =    0.0000

        Pearson goodness-of-fit  =  14201.29
```

```
        Prob > chi2(2622)          =    0.0000
```

*Both tests had very large (and highly significant) $\chi^2$ statistics indicating lack of fit.*

(c) Given that there is clear evidence of overdispersion, refit the model but compute scaled SE (scaled by the Pearson dispersion parameter)

```
. glm los sex age_inv white mar_c2 bmi_ct prmi, fam(poisson) link(log) scale(x2)

Generalized linear models                         No. of obs      =      2629
Optimization     : ML                             Residual df     =      2622
                                                  Scale parameter =         1
Deviance         =  8872.443582                   (1/df) Deviance =  3.383846
Pearson          =  14201.29271                   (1/df) Pearson  =  5.416206

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

                                                  AIC             =  6.662052
Log likelihood   = -8750.267314                   BIC             = -11774.13

------------------------------------------------------------------------------
             |                 OIM
         los |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .0117129   .0421875     0.28   0.781    -.0709732     .094399
     age_inv |  -37.03329   6.351249    -5.83   0.000     -49.4815   -24.58507
       white |   -.092017   .0585393    -1.57   0.116    -.2067518    .0227179
      mar_c2 |  -.0653547   .0410838    -1.59   0.112    -.1458775    .0151681
      bmi_ct |   .0093158   .0032215     2.89   0.004     .0030017    .0156298
        prmi |   .0316093    .043196     0.73   0.464    -.0530534     .116272
       _cons |   1.778887   .0608067    29.25   0.000     1.659708    1.898066
------------------------------------------------------------------------------
(Standard errors scaled using square root of Pearson X2-based dispersion.)
```

*This substantially changes the levels of statistical significance for parameters in the model. The predictors (-white-, -mar_c2- and -prmi-) which were significant at P<0.1 are no longer significant at that level. This suggests that the original model was substantially overestimating the significance of the parameters.*

5. Detailed diagnostics

(a) Compute residuals (deviance, Pearson and Anscombe) as well as Cook's distance values for each observation. As a first step, determine if there are many Pearson residuals which you would consider extreme? The determine if any individuals had particularly large Cook's D.

*There were no Pearson residuals <0, but there was a huge number (n=502, 19% of observations) of values >3. This suggests that the model is doing a very poor job of predicting long hospital stays. The 3 smallest and 10 largest residuals are shown here.*

```
. list id los died_hosp sex age white mar_c2 bmi prmi mu res_p in 1/3, clean noobs ab(5)

     id   los   die~p   sex   age   white   mar~2        bmi   prmi     mu   res_p
    789     0       0     0    83       1       0   31.73985      0   6.24   -2.50
    302     0       1     0    89       1       1   33.23273      0   6.11   -2.47
    231     0       1     1    88       1       1   28.24324      1   6.06   -2.47

. list id los died_hosp sex age white mar_c2 bmi prmi mu res_p in -10/-1, clean noobs
```

```
ab(5)
```

| id | los | die~p | sex | age | white | mar~2 | bmi | prmi | mu | res_p |
|---|---|---|---|---|---|---|---|---|---|---|
| 107 | 34 | 0 | 1 | 44 | 1 | 0 | 36.91451 | 1 | 4.60 | 13.73 |
| 181 | 35 | 0 | 0 | 64 | 1 | 1 | 21.21436 | 1 | 4.79 | 13.82 |
| 90 | 36 | 1 | 1 | 43 | 1 | 0 | 48.78378 | 1 | 5.04 | 13.84 |
| 2234 | 40 | 1 | 1 | 49 | 1 | 0 | 53.05947 | 0 | 5.65 | 14.51 |
| 2827 | 42 | 1 | 0 | 59 | 0 | 1 | 37.75879 | 0 | 5.65 | 15.33 |
| 1334 | 43 | 1 | 1 | 71 | 1 | 1 | 20.08571 | 1 | 5.08 | 16.85 |
| 91 | 51 | 0 | 1 | 59 | 1 | 0 | 33.76678 | 1 | 5.54 | 19.35 |
| 2271 | 61 | 0 | 1 | 75 | 1 | 1 | 26.25952 | 0 | 5.36 | 24.05 |
| 2880 | 68 | 1 | 0 | 90 | 1 | 0 | 18.35611 | 0 | 5.70 | 26.12 |
| 366 | 102 | 0 | 1 | 85 | 1 | 0 | 27.82202 | 1 | 6.35 | 38.02 |

*There are no obvious patterns in either set of residuals.*

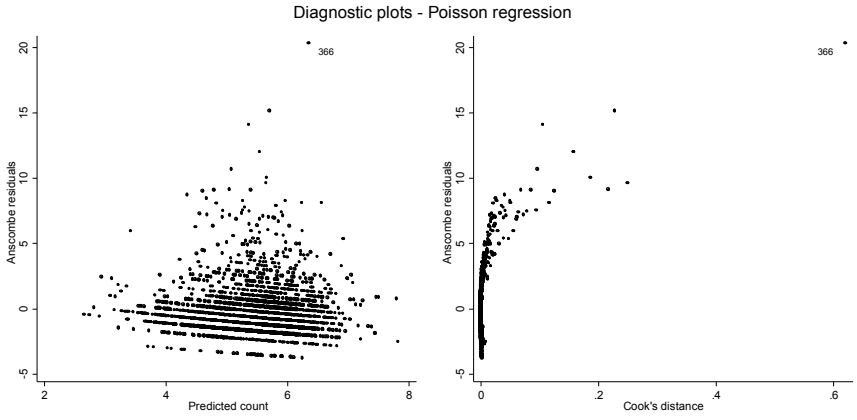
*The 10 largest Cook's D values were:*

```
. list  id res_d res_p res_a cooks_d in -10/-1, clean noobs
```

| id | res_d | res_p | res_a | cooks_d |
|---|---|---|---|---|
| 1334 | 10.39832 | 16.85 | 10.67493 | .0962329 |
| 2271 | 13.62624 | 24.05 | 14.10777 | .1052394 |
| 176 | 7.973344 | 11.55 | 8.106568 | .1167735 |
| 107 | 8.804462 | 13.73 | 9.007199 | .1248038 |
| 91 | 11.65881 | 19.35 | 11.99619 | .1582765 |
| 2827 | 9.812962 | 15.33 | 10.04043 | .1868115 |
| 90 | 8.957086 | 13.84 | 9.15582 | .2166497 |
| 2880 | 14.59363 | 26.12 | 15.13116 | .2278622 |
| 2234 | 9.412844 | 14.51 | 9.619774 | .2499195 |
| 366 | 19.39733 | 38.02 | 20.31488 | .6198722 |

*Clearly, observation 366 stands out as having a very large Cook's D, indicating that it has a large impact on the predicted values in the model.*

(b) Diagnostic plots - generate plots of Anscombe residuals against predicted values and Cook's distances.




*Both plots highlight observation 366 as an extreme outlier. Although it did not have a very large residual, it did have a large Cook's D indicating that removal of this observation would substantially improve the fit of the model.*

(c) Refit the model without obs. # 366.

*The revised model is not shown, but there was a 27% drop in the coefficient for -mar_c2- (married), (from -.065 to -.047). The estimate of the effect of being married was quite sensitive to this observation. There appeared to be large changes in the coefficients for -sex- and -prmi-, but these predictors were completely non-significant, so these were just very small coefficients that were "bouncing around zero". There were no substantive changes in any other coefficients.*

6. Negative binomial regression

(a) Fit the same model as developed above using negative binomial regression. Is there evidence that an NB model would be preferred to a Poisson model?

```
. nbreg los sex age_inv white mar_c2 bmi_ct prmi

... some output omitted

Negative binomial regression                    Number of obs   =       2629
                                                LR chi2(6)      =      89.29
Dispersion     = mean                           Prob > chi2     =     0.0000
Log likelihood = -6949.1536                     Pseudo R2       =     0.0064

------------------------------------------------------------------------------
         los |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .0087693   .0314031     0.28   0.780    -.0527796    .0703183
     age_inv |  -38.22552   4.566919    -8.37   0.000    -47.17652   -29.27453
       white |  -.0971096   .0447498    -2.17   0.030    -.1848176   -.0094016
      mar_c2 |  -.0687895   .0304447    -2.26   0.024    -.1284599    -.009119
      bmi_ct |   .0090554    .002423     3.74   0.000     .0043065    .0138043
        prmi |   .0322752   .0327668     0.98   0.325    -.0319467     .096497
       _cons |   1.786068   .0463574    38.53   0.000     1.695209    1.876927
-------------+----------------------------------------------------------------
    /lnalpha |  -.9757421   .0384263                     -1.051056   -.9004279
-------------+----------------------------------------------------------------
       alpha |   .3769125   .0144834                      .3495683    .4063957
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) = 3602.23 Prob>=chibar2 = 0.000
```

*Alpha equals 0.377 and is highly significant, indicating that there is more variability in the outcome (after adjustment for the fixed effects), than would be expected with a Poisson distribution. The estimated variance of the NB distribution fit to these data is:*

$$\text{var} = (1+\alpha)\mu = 1.37\mu$$

(b) Are the estimates of the fixed effects similar from a Poisson and NB model?

```
 . estimates table pois_ml nb_ml

----------------------------------------
    Variable |  pois_ml       nb_ml
-------------+--------------------------
los          |
         sex |  .01171289    .00876935
     age_inv | -37.033283   -38.225523
       white | -.09201696   -.09710957
      mar_c2 | -.06535473   -.06878946
```

```
      bmi_ct |  .00931576     .0090554
        prmi |  .03160932    .03227516
       _cons |  1.7788871    1.7860681
-------------+----------------------------
lnalpha      |
       _cons |              -.97574214
-------------------------------------------
```

*The estimates of the fixed effects are quite close.*

(c) Are NB models fit by maximum likelihood and by GLM comparable?

*To 3 decimal places, the results are identical (results not shown)*

7. NB diagnostics

(a) Obtain both deviance and Pearson $\chi^2$ statistics. Do they provide evidence of lack of fit?

*The deviance $\chi^2$ is 2569 on 2622 df (P=0.77) which shows no evidence of lack of fit. However, the Pearson $\chi^2$ is 4634 (P<0.001) which suggest substantial overdispersion remains. Faced with this conflicting evidence, it is wise to investigate possible reasons for lack of fit as thoroughly as possible.*

(b) Compute Pearson residuals for all observations. Are there an excess of observations <-3 or >3?

*There were no observations with residuals <-3, but there were 413 with >3 (the largest is 20.6!). These are individuals with unexpectedly long hospital stays. Clearly, some important predictors of length of stay are missing from the model.*

(c) Compute Cook's D for all observations. Do any observations stand out as having very large values? If so, refit the model with this/these observation(s).

*Once again, observation 366 stands out. It has a Cook's D that is more than twice as large as the next largest. The effect of deleting this observation on the fixed effects were similar to those discussed in 5(c). Deleting this observation reduced the Pearson $\chi^2$ from 4633 to 4347, but there is still strong evidence of lack of fit as a result of these unexpectedly long hospital stays.*

8. Zero-inflated models

(a) Fit a zero-inflated NB model. Is there any evidence that there are more values of zero (*ie* patients discharged on the same day as admission) than would be expected?

```
. zinb los sex age_inv white mar_c2 bmi_ct prmi,  ///
      inflate(sex age_inv white mar_c2 bmi_ct prmi) nolog vuong

Zero-inflated negative binomial regression        Number of obs   =       2629
                                                  Nonzero obs     =       2598
                                                  Zero obs        =         31

Inflation model = logit                           LR chi2(6)      =      89.29
Log likelihood  = -6949.154                       Prob > chi2     =     0.0000

------------------------------------------------------------------------------
         los |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
los          |
         sex |   .0087694   .0314031     0.28   0.780    -.0527796    .0703184
```

```
     age_inv |  -38.22552   4.566919    -8.37   0.000    -47.17652   -29.27452
       white |  -.0971096   .0447498    -2.17   0.030    -.1848175   -.0094016
      mar_c2 |  -.0687894   .0304447    -2.26   0.024    -.1284599    -.009119
      bmi_ct |   .0090554    .002423     3.74   0.000     .0043065    .0138043
        prmi |   .0322751   .0327668     0.98   0.325    -.0319467     .096497
       _cons |   1.786068   .0463574    38.53   0.000     1.695209    1.876927
-------------+----------------------------------------------------------------
inflate      |
         sex |   .0189264    8522.52     0.00   1.000    -16703.81    16703.85
     age_inv |  -9.301194    1177277    -0.00   1.000     -2307429     2307411
       white |   .1627786   11705.04     0.00   1.000    -22941.29    22941.61
      mar_c2 |   .0400816   8271.041     0.00   1.000     -16210.9    16210.98
      bmi_ct |  -.0007988   637.8099    -0.00   1.000    -1250.085    1250.084
        prmi |    .064025   8823.299     0.00   1.000    -17293.29    17293.41
       _cons |  -24.34123   12289.37    -0.00   0.998    -24111.07    24062.39
-------------+----------------------------------------------------------------
    /lnalpha |  -.9757422   .0384263   -25.39   0.000    -1.051056    -.900428
-------------+----------------------------------------------------------------
       alpha |   .3769125   .0144834                      .3495683    .4063957
------------------------------------------------------------------------------
Vuong test of zinb vs. standard negative binomial: z =    -0.02  Pr>z = 0.5071
```

*A simple tabulation of the variable -los- shows that only 55 of the 2963 observations had a value of zero, so it seems unlikely that there is going to be any evidence of zero inflation. This is confirmed by Vuong statistic which is very close to zero (P=0.51). There is absolutely no evidence of excess zeros.*