## **The model**

The generalized linear regression model can be stated as:

$$I_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + u_i \qquad (1)$$

and

$$\Pr(Y_i = y_i) = f(I_i)$$

where: $I_i$ is index variables.
$y_i$ is counted number 0, 1, 2,...
$x_{ki}$ is independent variable $k$.
$f(.)$ is either Poisson or Negative Binomial probability distribution function.
$u_i$ is disturbance term.

**Requirements (using Data set file – `assign12.dta`):**

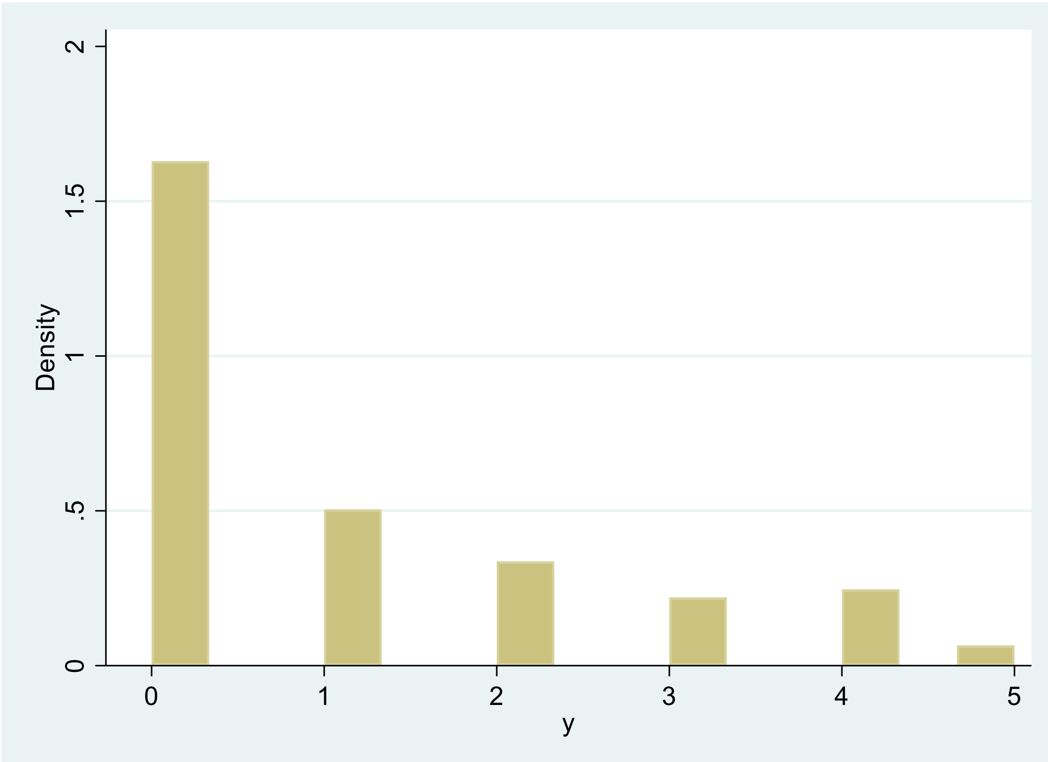
1 Estimate models for $y_i$ assuming that the model is traditional linear regression model. Interpret your estimated result.

```
. use "C:\Users\Jilllin\OneDrive\Desktop\Thammasat\EE426\Data\assign12.dta"

. reg y x1 x2 x3 x4

      Source |       SS           df       MS      Number of obs   =       232
-------------+----------------------------------   F(4, 227)       =      5.96
       Model |  44.7298499         4  11.1824625   Prob > F        =    0.0001
    Residual |  425.748598       227  1.87554449   R-squared       =    0.0951
-------------+----------------------------------   Adj R-squared   =    0.0791
       Total |  470.478448       231  2.03670324   Root MSE        =    1.3695

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .1016201   .0435073     2.34   0.020     .0158904    .1873499
          x2 |   .1345044   .0462142     2.91   0.004     .0434407     .225568
          x3 |  -.0748194   .0480457    -1.56   0.121    -.1694919    .0198531
          x4 |   .1684563   .0688243     2.45   0.015     .0328401    .3040725
       _cons |   .9568064    .107007     8.94   0.000     .7459523     1.16766
------------------------------------------------------------------------------

. est store ols
```

From estimated result, overall significance of both model and individual variables is significant except x3. However, the R-square is quite low. This means that the model has low explanatory power.

2 Create histogram for $y_i$. Determine whether there is limitation of dependent variable in this case. If yes, what type of limitation is it?

```
. histogram y

(bin=15, start=0, width=.33333333)
```




We can see that the distribution of y is highly skewed. Hence, normal OLS should not be applied here. Furthermore, the distribution looks like poisson distribution.

3 Estimate models for $y_i$ assuming that the probability functions follow Poisson probability distribution. Perform GOF test and determine whether Poisson is appropriated in this case. Interpret the estimated result (sign and meaning (in term of incidence-rate ratios), overall test, individual test, pseudo $R^2$, marginal effects).

```
. poisson y x1 x2 x3 x4 ,nolog

Poisson regression                                Number of obs     =        232
                                                  LR chi2(4)        =      43.33
                                                  Prob > chi2       =     0.0000
Log likelihood = -342.88107                       Pseudo R2         =     0.0594

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .0971474   .0306762     3.17   0.002     .0370231    .1572717
          x2 |   .1293024   .0330916     3.91   0.000      .064444    .1941607
          x3 |  -.0715533   .0342177    -2.09   0.037    -.1386187   -.0044879
          x4 |   .1734482   .0507707     3.42   0.001     .0739395    .2729569
       _cons |  -.1284876   .0849064    -1.51   0.130     -.294901    .0379259
------------------------------------------------------------------------------
```

```
. estat gof

         Deviance goodness-of-fit =  409.4921
         Prob > chi2(227)         =    0.0000

         Pearson goodness-of-fit  =  423.3541
         Prob > chi2(227)         =    0.0000

. poisson y x1 x2 x3 x4 ,ir nolog

Poisson regression                              Number of obs     =        232
                                                LR chi2(4)        =      43.33
                                                Prob > chi2       =     0.0000
Log likelihood = -342.88107                     Pseudo R2         =     0.0594

------------------------------------------------------------------------------
           y |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   1.102023   .0338059     3.17   0.002     1.037717    1.170314
          x2 |   1.138034   .0376594     3.91   0.000     1.066566    1.214291
          x3 |   .9309467   .0318548    -2.09   0.037     .8705599    .9955222
          x4 |   1.189399   .0603866     3.42   0.001     1.076742    1.313844
       _cons |   .8794245   .0746687    -1.51   0.130     .7446053    1.038654
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.

. est store poisson
```

From GOF test, we can see that H0 is rejected. This indicates that poisson model should not be employed here. However, after using poisson model, we can see higher significance of all variables.

**4** Estimate models for $y_i$ assuming that the probability functions follow Negative Binomial probability distribution. Determine whether Negative Binomial regression model is appropriated in this case. Interpret your estimated result (sign and meaning (in term of incidence-rate ratios), overall test, individual test, pseudo $R^2$, marginal effects).

```
. nbreg y x1 x2 x3 x4, nolog

Negative binomial regression                    Number of obs     =        232
                                                LR chi2(4)        =      21.24
Dispersion     = mean                           Prob > chi2       =     0.0003
Log likelihood = -317.49278                     Pseudo R2         =     0.0324

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .1285534   .0506934     2.54   0.011     .0291962    .2279106
          x2 |    .151011   .0506477     2.98   0.003     .0517434    .2502785
          x3 |  -.0672859   .0481376    -1.40   0.162    -.1616339    .0270621
          x4 |   .1726312   .0707035     2.44   0.015      .034055    .3112075
       _cons |  -.1435596   .1177204    -1.22   0.223    -.3742874    .0871682
-------------+----------------------------------------------------------------
    /lnalpha |   .0479945   .2389531                     -.4203449    .5163339
```

```
-------------+----------------------------------------------------------------
       alpha |   1.049165   .2507012                      .6568202    1.675872
------------------------------------------------------------------------------
LR test of alpha=0: chibar2(01) = 50.78                Prob >= chibar2 = 0.000

. nbreg y x1 x2 x3 x4,ir nolog

Negative binomial regression                      Number of obs     =        232
                                                  LR chi2(4)        =      21.24
Dispersion     = mean                             Prob > chi2       =     0.0003
Log likelihood = -317.49278                       Pseudo R2         =     0.0324

------------------------------------------------------------------------------
           y |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   1.137182   .0576476     2.54   0.011     1.029627    1.255973
          x2 |   1.163009   .0589037     2.98   0.003     1.053105    1.284383
          x3 |   .9349279   .0450052    -1.40   0.162     .8507526    1.027432
          x4 |   1.188428    .084026     2.44   0.015     1.034641    1.365072
       _cons |   .8662692   .1019776    -1.22   0.223     .6877792     1.09108
-------------+----------------------------------------------------------------
    /lnalpha |   .0479945   .2389531                     -.4203449    .5163339
-------------+----------------------------------------------------------------
       alpha |   1.049165   .2507012                      .6568202    1.675872
------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation.
Note: _cons estimates baseline incidence rate.
LR test of alpha=0: chibar2(01) = 50.78                Prob >= chibar2 = 0.000

.
```

From LR test, Ho is rejected. This suggests that Negative Binomial model is more appropriated in this case. Overall significance is good except x3. However, pseudo R2 is quite low.

5 Estimate models for $y_i$ assuming that the model is Zero Inflated Poisson ($x_{1i}$, $x_{2i}$, and $x_{3i}$ are independent variables in Poisson model and $x_{4i}$ is independent variable in Inflated (Logit) model). Interpret your estimated result. Determine which model (Linear regression model, Poisson, Negative Binomial, or ZIP) is the most appropriated model in this case? Why? (provide the tests)

```
. zip y x1 x2 x3, inflate(x4) forcevuong nolog

Zero-inflated Poisson regression                  Number of obs     =        232
                                                  Nonzero obs       =        106
                                                  Zero obs          =        126

Inflation model = logit                           LR chi2(3)        =      10.35
Log likelihood  = -312.6158                       Prob > chi2       =     0.0158

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
y            |
          x1 |   .0805446   .0398159     2.02   0.043     .0025068    .1585824
          x2 |   .0857883   .0372107     2.31   0.021     .0128567    .1587199
          x3 |  -.0672468   .0357098    -1.88   0.060    -.1372367     .002743
       _cons |   .4589728   .1106031     4.15   0.000     .2421947    .6757508
```

```
-------------+----------------------------------------------------------------
inflate      |
          x4 |  -.2738532   .1212311    -2.26   0.024    -.5114618   -.0362446
       _cons |  -.3379298   .1908217    -1.77   0.077    -.7119334    .0360738
------------------------------------------------------------------------------
Vuong test of zip vs. standard Poisson:            z =     3.92  Pr>z = 0.0000
Warning: The Vuong test is not appropriate for testing zero-inflation.

.
```

From Vuong test, Ho is rejected. This suggests that ZIP model is more appropriated than poisson model. Overall significance is good except x3.

## 6 According to the above (1-5), determine the most appropriated model for this case. Give explanation why?

```
est tab ols poisson nbreg zip, star(.1 .05 .01)

------------------------------------------------------------------------------
    Variable |      ols           poisson          nbreg             zip
-------------+----------------------------------------------------------------
_            |
          x1 |  .10162013**
          x2 |  .13450436***
          x3 | -.07481941
          x4 |  .16845626**
       _cons |   .9568064***
-------------+----------------------------------------------------------------
y            |
          x1 |                 .09714739***    .12855344**     .08054462**
          x2 |                 .12930238***    .15101096***    .08578827**
          x3 |                 -.0715533**    -.06728591      -.06724685*
          x4 |                 .17344816***    .17263123**
       _cons |                -.12848756      -.14355961       .45897275***
-------------+----------------------------------------------------------------
    /lnalpha |                                  .0479945
-------------+----------------------------------------------------------------
inflate      |
          x4 |                                                -.27385319**
       _cons |                                                -.33792981*
------------------------------------------------------------------------------
                                         legend: * p<.1; ** p<.05; *** p<.01
```

From the table, we can see that ZIP has the most significance overall. Furthermore, Voung test also suggest that ZIP model is more appropriated than poisson model.