1.) Estimate models for y_i assuming that the model is traditional linear regression model.

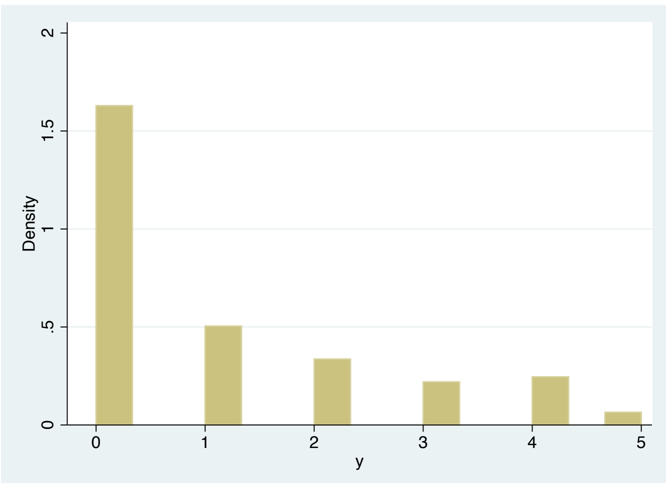
Interpret your estimated result.

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

Interpretation:
1.sign&meaning: x1, x2, x4 have positive correlation with y but x3 negative correlate with y
2.overall test: the model is significant in overall test at 95% Confidence since p-value is less than 0.05
3.GOF: R-square is quite low
4.individual test: x3 coefficient is insignificant at 95% CI

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)
```



From the histogram we can see that the data is counted data which is not normal distribution Which making the OLS result to be biased.

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 *R2* , marginal effects).

Estimate the poisson model

```
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
------------------------------------------------------------------------------
```

GOF test

```
. estat gof

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

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

Sign & meaning

```
. lincom x1, ir

 ( 1)  [y]x1 = 0

------------------------------------------------------------------------------
           y |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.102023   .0338059     3.17   0.002     1.037717    1.170314
------------------------------------------------------------------------------

. lincom x2, ir

 ( 1)  [y]x2 = 0

------------------------------------------------------------------------------
           y |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.138034   .0376594     3.91   0.000     1.066566    1.214291
------------------------------------------------------------------------------

. lincom x3, ir

 ( 1)  [y]x3 = 0

------------------------------------------------------------------------------
           y |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .9309467   .0318548    -2.09   0.037     .8705599    .9955222
------------------------------------------------------------------------------
```

```
. lincom x4, ir

 ( 1)  [y]x4 = 0

------------------------------------------------------------------------------
           y |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.189399   .0603866     3.42   0.001     1.076742    1.313844
------------------------------------------------------------------------------
```

Marginal effect

```
. mfx

Marginal effects after poisson
      y  = Predicted number of events (predict)
         =  .95621703
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
      x1 |    .092894      .02882    3.22   0.001   .036403  .149385  -.317697
      x2 |   .1236411       .0308    4.01   0.000   .063277  .184005   .812709
      x3 |  -.0684205      .03246   -2.11   0.035   -.13204 -.004801  -.818103
      x4 |   .1658541      .04736    3.50   0.000   .073022  .258686   -.28275
------------------------------------------------------------------------------
```

Appropriateness: from the GOF test we reject the null hypothesis which mean that the Poisson model is not an appropriated one.

Interpretation:
1.sign&meaning: from IRR x1, x2, x4 have positive correlation with y but x3 negative correlate with y
2.overall test: the model is significant in overall test at 95% Confidence since p-value is less than 0.05
3.GOF: Pseudo R-square is quite low
4.individual test: x3 coefficient is insignificant at 95% CI

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 *R2* , marginal effects).

Estimate the negative binomial model

```
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
```

Sign & meaning (get IRR)

```
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
------------------------------------------------------------------------------
LR test of alpha=0: chibar2(01) = 50.78              Prob >= chibar2 = 0.000
```

Marginal effect

```
. mfx

Marginal effects after nbreg
      y  = Predicted number of events (predict)
         =  .94607122
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
      x1 |   .1216207      .04796    2.54   0.011    .02763  .215611  -.317697
      x2 |   .1428671      .04796    2.98   0.003   .048862  .236872   .812709
      x3 |  -.0636573      .04556   -1.40   0.162  -.152956  .025642  -.818103
      x4 |   .1633214      .06686    2.44   0.015   .032269  .294374   -.28275
------------------------------------------------------------------------------
```

Appropriateness: from the LR test we reject the null hypothesis which mean that the negative binomial is appropriated.

Interpretation:
1.sign&meaning: from IRR x1, x2, x4 have positive correlation with y but x3 negative correlate with y
2.overall test: the model is significant in overall test at 95% Confidence since p-value is less than 0.05
3.GOF: Pseudo R-square is quite low
4.individual test: x3 coefficient is insignificant at 95% CI

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)

Estimate the ZIP model

```
zip y x1 x2 x3, inflate(x4) vuong

Fitting constant-only model:

Iteration 0:   log likelihood = -355.48956
Iteration 1:   log likelihood =  -321.8304
Iteration 2:   log likelihood = -317.80147
Iteration 3:   log likelihood = -317.79274
Iteration 4:   log likelihood = -317.79274

Fitting full model:

Iteration 0:   log likelihood = -317.79274
Iteration 1:   log likelihood = -312.73621
Iteration 2:   log likelihood =  -312.6159
Iteration 3:   log likelihood =  -312.6158
Iteration 4:   log likelihood =  -312.6158

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
```

Appropriateness: from the LR test we reject the null hypothesis which mean that the ZIP model is appropriated.

Interpretation:
overall test: the model is significant in overall test at 95% Confidence since p-value is less than 0.05
individual test: x3 coefficient is insignificant at 95% CI but significant at 90%

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

5&6 answer: the most appropriated one is ZIP model since it's the only one model that appropriate from the appropriateness test and significant in individual test at the same time.