Problem 1 [12 points]

a. [3 points]

```
. tab years severity [freq=pop], all

           | severity
     years |         1          2          3          4          5 |     Total
-----------+-------------------------------------------------------+----------
         1 |         0          1          6         11         12 |        30
         2 |         5         37        114        165        136 |       457
         3 |        29        155        299        268        181 |       932
         4 |        11         35         48         33         28 |       155
         5 |         4         61         41          7          2 |       115
-----------+-------------------------------------------------------+----------
     Total |        49        289        508        484        359 |      1689

         Pearson chi2(16) = 214.0613   Pr = 0.000
likelihood-ratio chi2(16) =        .
               Cramer's V =   0.1780
                    gamma =  -0.3702  ASE = 0.027
          Kendall's tau-b =  -0.2532  ASE = 0.019

. rename row years
. rename col severity

. xi: poisson pop i.years i.severity

i.years              Iyears_1-5   (naturally coded; Iyears_1 omitted)
i.severity           Isever_1-5   (naturally coded; Isever_1 omitted)

Poisson regression                              Number of obs    =      25
Goodness-of-fit chi2(16)    =   210.357         Model chi2(8)    =2060.363
Prob > chi2                 =    0.0000         Prob > chi2      =  0.0000
Log Likelihood              =  -167.489         Pseudo R2        =  0.8602
```

```
------------------------------------------------------------------------------
         pop |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    Iyears_2 |   2.723486   .1884715     14.450   0.000     2.354089    3.092884
    Iyears_3 |   3.436135   .1854893     18.525   0.000     3.072583    3.799688
    Iyears_4 |   1.642228   .1994616      8.233   0.000      1.25129    2.033165
    Iyears_5 |   1.343735   .2050097      6.554   0.000     .9419232    1.745546
    Isever_2 |   1.774606   .1544939     11.487   0.000     1.471804    2.077409
    Isever_3 |   2.338661   .1495883     15.634   0.000     2.045473    2.631849
    Isever_4 |   2.290265   .1499142     15.277   0.000     1.996438    2.584091
    Isever_5 |   1.991502   .1522947     13.077   0.000      1.69301    2.289994
       _cons |  -.1388743   .2305416     -0.602   0.547    -.5907275     .312979
------------------------------------------------------------------------------
```

When the table is treated as a whole, a G^2 statistic of 210.357 is obtained, along with a χ^2 statistic of 214.0613. Both of these values, when compared to the Chi-square distribution with 16 d.f., indicate we should reject the null hypothesis of "no association" between attack severity and number of years since vaccination (among those who have experienced some form of attack). Note, however, that because the ages of the people are not given, there is confounding of age and years-since-vaccination.

b. [3 points]

```
. tab years severity [freq=pop]

           | severity
     years |         1          2          3          4          5 |     Total
-----------+-------------------------------------------------------+----------
         1 |         0          1          6         11         12 |        30
         2 |         5         37        114        165        136 |       457
         3 |        29        155        299        268        181 |       932
         4 |        11         35         48         33         28 |       155
         5 |         4         61         41          7          2 |       115
-----------+-------------------------------------------------------+----------
     Total |        49        289        508        484        359 |      1689

. predict lmhat
. gen mhat=exp(lmhat)
. gen pres=(pop-mhat)/sqrt(mhat)
. table years severity[iw=pop],format(%6.2f)

----------+-----------------------------------------
          |                  severity
    years |      1       2       3       4       5
----------+-----------------------------------------
        1 |              1.00    6.00   11.00   12.00
        2 |   5.00   37.00  114.00  165.00  136.00
        3 |  29.00  155.00  299.00  268.00  181.00
        4 |  11.00   35.00   48.00   33.00   28.00
        5 |   4.00   61.00   41.00    7.00    2.00
----------+-----------------------------------------
```

```
. table years severity[iw=mhat],format(%6.2f)

----------+-----------------------------------------
          |                  severity
    years |      1       2       3       4       5
----------+-----------------------------------------
        1 |   0.87    5.13    9.02    8.60    6.38
        2 |  13.26   78.20  137.45  130.96   97.14
        3 |  27.04  159.47  280.32  267.07  198.10
        4 |   4.50   26.52   46.62   44.42   32.95
        5 |   3.34   19.68   34.59   32.95   24.44
----------+-----------------------------------------

. table years severity[iw=pres],format(%6.2f)

----------+-----------------------------------
          |              severity
    years |     1      2      3      4      5
----------+-----------------------------------
        1 | -0.93  -1.82  -1.01   0.82   2.23
        2 | -2.27  -4.66  -2.00   2.97   3.94
        3 |  0.38  -0.35   1.12   0.06  -1.21
        4 |  3.07   1.65   0.20  -1.71  -0.86
        5 |  0.36   9.32   1.09  -4.52  -4.54
----------+-----------------------------------
```

Note that the residuals in the first 3 columns of rows 1 and 2 are negative (i.e. the actual values are lower than predicted by the model of additive row and column effects), while the actual values are higher than predicted for the last 2 columns in both of these rows. In rows 4 and 5 this pattern is reversed: the last 2 columns have negative residuals while the first 3 columns have positive residuals.

The pattern in the residuals indicates that, when compared to the values expected under a model of "no association", there are relatively few "non-severe" attacks (those categorized as "sparse" and "very sparse") for those who were vaccinated within the last 25 years. Also, it appears there are too many "severe" attacks ("Haemorrhagic", "confluent" and "abundant") in the 0-10 and 10-25 "years-since-vaccination" groups when compared to the values predicted by the model of no association. The deviation from the expected is more extreme for those in the 10-25 "years since vaccination group". The deviation from the expected number of attacks increases with time since vaccination for those within 25 years since vaccination and in general underpredicts for those within 25 years with the 2 categories of least severity and overpredicts for those in the categories of greatest severity. This pattern for those more than 25 years since vaccination is reversed: the "no association" model underpredicts for those of greatest severity and overpredicts for those of least severity. If there really were no association between years since vaccination and severity of attack, we would not expect to see a systematic pattern in the residuals from fitting the model of "no association".

c. [3 points]

From the above table, the Stata's predicted values of cell (2, 2) and cell (2, 4) are 78.20 and 130.96, correspondingly. If we use the coefficients estimated from the Poisson regression, we get the $\hat{m}_{2,2} = e^{(-.1388743+2.723486+1.774606)} = 78.19594$, and $\hat{m}_{2,4} = e^{(-.1388743+2.723486+2.290265)} = 130.958$. The difference only lies in the precision of rounding, since both procedures follow the underlying equation of $log(\hat{m}_{ij}) = \mu + \alpha_i + \beta_j$.

d. [3 points]

```
. tab year severe[freq=pop],all

           | severe
      year |         1          2          3          4          5 |     Total
-----------+-------------------------------------------------------+----------
         1 |         0          1          6         11         12 |        30
         2 |         5         37        114        165        136 |       457
-----------+-------------------------------------------------------+----------
     Total |         5         38        120        176        148 |       487

          Pearson chi2(4) =   2.4001   Pr = 0.663
 likelihood-ratio chi2(4) =        .
               Cramer's V =   0.0702
                    gamma =  -0.2209  ASE = 0.141
          Kendall's tau-b =  -0.0620  ASE = 0.040

. xi:poisson pop i.year i.severe
i.year                Iyear_1-2    (naturally coded; Iyear_1 omitted)
i.severe              Isever_1-5   (naturally coded; Isever_1 omitted)

Poisson regression                                Number of obs    =        10
Goodness-of-fit chi2(4)     =     2.855           Model chi2(5)    =   730.746
Prob > chi2                 =    0.5824           Prob > chi2      =    0.0000
Log Likelihood              =   -23.153           Pseudo R2        =    0.9404

------------------------------------------------------------------------------
         pop |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Iyear_2 |   2.723486   .1884715     14.450   0.000       2.354089    3.092883
    Isever_2 |   2.028148   .4757266      4.263   0.000       1.095741    2.960555
    Isever_3 |   3.178054   .4564355      6.963   0.000       2.283457    4.072651
    Isever_4 |   3.561046   .4535216      7.852   0.000        2.67216    4.449932
    Isever_5 |   3.387774   .4547051      7.450   0.000       2.496569     4.27898
       _cons |  -1.177629   .4809157     -2.449   0.014      -2.120206   -.2350514
------------------------------------------------------------------------------
```

```
. predict lmhat
. gen mhat=exp(lmhat)
. gen pres=(pop-mhat)/sqrt(mhat)
. table year severe [iw=pop], format(%6.2f)

----------+-----------------------------------------
          |                severe
     year |      1       2       3       4       5
----------+-----------------------------------------
        1 |              1.00    6.00   11.00   12.00
        2 |   5.00   37.00  114.00  165.00  136.00
----------+-----------------------------------------

. table year severe [iw=mhat], format(%6.2f)

----------+-----------------------------------------
          |                severe
     year |      1       2       3       4       5
----------+-----------------------------------------
        1 |   0.31    2.34    7.39   10.84    9.12
        2 |   4.69   35.66  112.61  165.16  138.88
----------+-----------------------------------------

. table year severe [iw=pres], format(%6.2f)

----------+---------------------------------
          |              severe
     year |     1      2      3      4      5
----------+---------------------------------
        1 | -0.55  -0.88  -0.51   0.05   0.95
        2 |  0.14   0.22   0.13  -0.01  -0.24
----------+---------------------------------
```

When only the first 2 rows are considered, a G^2 statistic of 2.855 and a χ^2 statistic of 2.4001 are obtained. Neither value is significant when compared to a χ^2 distribution with 4 d.f. Examination of residuals reveals no abnormality of deviance from zero. This implies that we cannot reject the null hypothesis of "no association" between years since vaccination and attack severity. This conclusion does differ from that of part (a). One possible reason for the difference is that the people in these 2 rows are more homogeneous with respect to age. The confounding effect of age and years since vaccination is minimized by excluding people of very different ages.

Exercise 4.2 [12 points]

a. [3 points]

```
. infile wais senility using wais.raw
(54 observations read)
. tab wais senility

           | senility
      wais |         0          1 |     Total
-----------+----------------------+----------
         4 |         1          1 |         2
         5 |         0          1 |         1
         6 |         1          1 |         2
         7 |         1          2 |         3
         8 |         0          2 |         2
         9 |         4          2 |         6
        10 |         5          1 |         6
        11 |         5          1 |         6
        12 |         2          0 |         2
        13 |         5          1 |         6
        14 |         5          2 |         7
        15 |         3          0 |         3
        16 |         4          0 |         4
        17 |         1          0 |         1
        18 |         1          0 |         1
        19 |         1          0 |         1
        20 |         1          0 |         1
-----------+----------------------+----------
     Total |        40         14 |        54
```

```
. logit senility wais

Logit Estimates                                   Number of obs =     54
                                                  chi2(1)       =  10.79
                                                  Prob > chi2   = 0.0010
Log Likelihood =  -25.50869                       Pseudo R2     = 0.1746

------------------------------------------------------------------------------
senility |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    wais |  -.3235304   .1139798     -2.838   0.005      -.5469266   -.1001342
   _cons |   2.404043   1.191835      2.017   0.044       .0680896    4.739997
------------------------------------------------------------------------------

. logistic senility wais

Logit Estimates                                   Number of obs =     54
                                                  chi2(1)       =  10.79
                                                  Prob > chi2   = 0.0010
Log Likelihood =  -25.50869                       Pseudo R2     = 0.1746

------------------------------------------------------------------------------
senility | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    wais |     .72359   .0824746     -2.838   0.005       .5787257     .904716
------------------------------------------------------------------------------
```

```
. logit

Logit Estimates                                   Number of obs =     54
                                                  chi2(1)       =  10.79
                                                  Prob > chi2   = 0.0010
Log Likelihood =  -25.50869                       Pseudo R2     = 0.1746

------------------------------------------------------------------------------
senility |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    wais |  -.3235304   .1139798     -2.838   0.005      -.5469266   -.1001342
   _cons |   2.404043   1.191835      2.017   0.044       .0680896    4.739997
------------------------------------------------------------------------------

. lfit, group(10) table

Logistic model for senility, goodness-of-fit test
(Table collapsed on quantiles of estimated probabilities)

Note: Because of ties, there are only 9 distinct quantiles.

_Group     _Prob     _Obs_1     _Exp_1     _Obs_0     _Exp_0     _Total
     1    0.0588          0        0.4          8        7.6          8
     2    0.0795          0        0.2          3        2.8          3
     3    0.1067          2        0.7          5        6.3          7
     4    0.1416          1        0.8          5        5.2          6
     5    0.2396          1        1.8          7        6.2          8
     7    0.3034          1        1.8          5        4.2          6
     8    0.3757          2        2.3          4        3.7          6
     9    0.5348          4        2.5          1        2.5          5
    10    0.7521          3        3.4          2        1.6          5

       number of observations =        54
             number of groups =         9
      Hosmer-Lemeshow chi2(7) =      5.99
                  Prob > chi2 =    0.5411
```

Using either `logit` or `logistic` command, you get the same estimates for

$$logit(\hat{\pi}) = \alpha + \beta x$$

i.e.

$$logit(\hat{\pi}) = 2.404 - .324x.$$

The command `logistic` gives the odds ratio as default. By typing a following logit you could get the same result as if you had run the `logit` command. Nevertheless, the command `logistic` gives the observed and fitted values for intervals of predictors (something we will need in part(c)), if you type a following `lfit`.

```
. predict pihat
. list wais pihat

         wais      pihat
 1.        20   .0168475
 2.        19   .0231343
 3.        18   .0316915
 4.        17   .0432737
 5.        16   .0588316
 6.        15   .0795181
 7.        14   .1066542
 8.        13   .1416258
 9.        12   .1856811
10.        11   .2396151
11.        10   .3033786
12.         9   .3757258
13.         8   .4540798
14.         7   .5347764
15.         6   .6136926
16.         5   .6870559
17.         4   .7521145
etc...

. plot pihat wais

.752115 +
        | *
        |
        |    *
        |
        |        *
        |
        |            *
        |
   p    |
   i    |               *
   h    |
   a    |                   *
   t    |
        |                       *
        |
        |                           *
        |                               *
        |                                   *
        |                                       *
        |                                           *  *
.016847 +                                                  *  *  *  *
        +------------------------------------------------------------+
             4                          wais                        20
```

From the estimated coefficients, we know $logit(\hat{\pi}) = 2.404 - .324x$. When $\pi = 0.5, log(\frac{\pi}{1-\pi}) = log(\frac{0.5}{1-0.5}) = 0$. Solve $2.404 - .324x = 0, x = 7.420$. That is, when the elderly people scored less than 7.42 points in WAIS, their estimated probability of senility would exceed 0.5. In this data set, it would be those who scored 4, 5, 6, and 7 in WAIS.

b. [3 points]

The odds ratio for $\beta = -.3235304$ is .72359 (from the command `logistic` output). It implies that for one unit change in WAIS scores, there is a corresponding decrease in odds of senility. For testing β=0, we can use the z-test as shown in the output; or equivalently, use the Wald chi-squared test $z^2 = 8.054$ with df=1. Both tests reject the null hypothesis that $\beta = 0$. That is, there is a statistically significant linear effect of the WAIS scores on the odds of senility, i.e. the higher they scored on WAIS, the less likely they would be diagnosed of senility.

c. [3 points]

```
. input wais senility total
         wais   senility      total
  1. 0 2 3
  2. 1 8 19
  3. 2 4 24
  4. 3 0 8
  5. end
. blogit senility total wais

Logit Estimates                                   Number of obs =     54
                                                  chi2(1)       =   9.68
                                                  Prob > chi2   = 0.0019
Log Likelihood = -26.065388                       Pseudo R2     = 0.1565

------------------------------------------------------------------------------
_outcome |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    wais |  -1.400881   .5161109     -2.714   0.007      -2.412439   -.3893219
   _cons |   1.056069   .7820068      1.350   0.177      -.4766364    2.588774
------------------------------------------------------------------------------
. predict lhat
. gen mhat=total*lhat
. gen diff=senility-mhat
. gen res=diff/sqrt(mhat*(1-mhat/total))
. list wais total senility mhat diff res
```

	wais	total	senility	mhat	diff	res
1.	0	3	2	2.225816	-.2258158	-.2979532
2.	1	19	8	7.878181	.1218195	.0567274
3.	2	24	4	3.566191	.4338086	.2489584
4.	3	8	0	.3298121	-.3298121	-.5865098

Using the output shown in part (a), we can see for the approximately equal size of intervals of WAIS, the observed values and fitted values are very close to one another. Again, this is another indication of the model of $logit(\hat{\pi}) = 2.404 - .324x$ fits data adequately. If you regroup data into several intervals (e.g. 4 here), you can also examine the Pearson residuals. We can see the four residuals center around zero and less than 2, thus it indicates this model fits data well.

d. [3 points]

```
. regress senility wais

  Source |       SS       df       MS                  Number of obs =      54
---------+------------------------------               F(  1,    52) =   11.52
   Model |  1.88115078     1  1.88115078               Prob > F      =  0.0013
Residual |   8.4892196    52  .163254223               R-squared     =  0.1814
---------+------------------------------               Adj R-squared =  0.1657
   Total |  10.3703704    53  .195667365               Root MSE      =  .40405

------------------------------------------------------------------------------
senility |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    wais |  -.0507911   .0149626     -3.395   0.001      -.0808158   -.0207664
   _cons |   .8471189   .1816976      4.662   0.000       .4825159    1.211722
------------------------------------------------------------------------------

. glm senility wais

Residual df  =        52                     No. of obs =        54
Pearson X2   =   8.48922                     Deviance   =   8.48922
Dispersion   =  .1632542                     Dispersion =  .1632542

Gaussian (normal) distribution, identity link
------------------------------------------------------------------------------
senility |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    wais |  -.0507911   .0149626     -3.395   0.001      -.0808158   -.0207664
   _cons |   .8471189   .1816976      4.662   0.000       .4825159    1.211722
------------------------------------------------------------------------------
(Model is ordinary regression, use fit or regress instead)

. glm senility wais, f(binomial) l(identity)

Residual df  =        52                     No. of obs =        54
Pearson X2   =  43.77081                     Deviance   =  50.95689
Dispersion   =  .8417463                     Dispersion =  .9799402

Bernoulli distribution, identity link
------------------------------------------------------------------------------
senility |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    wais |  -.0508072   .0115441     -4.401   0.000      -.0734332   -.0281811
   _cons |   .8501443   .1788871      4.752   0.000        .499532    1.200757
------------------------------------------------------------------------------
convergence not achieved.
r(430);

. predict ohat
. list wais pihat ohat

         wais      pihat       ohat
  1.       20   .0168475  -.1659991
  2.       19   .0231343  -.1151919
  3.       18   .0316915  -.0643847
  4.       17   .0432737  -.0135776
  5.       16   .0588316   .0372296
  6.       15   .0795181   .0880368
  7.       14   .1066542   .1388439
  8.       13   .1416258   .1896511
  9.       12   .1856811   .2404583
 10.       11   .2396151   .2912655
 11.       10   .3033786   .3420726
 12.        9   .3757258   .3928798
 13.        8   .4540798    .443687
 14.        7   .5347764   .4944941
 15.        6   .6136926   .5453013
 16.        5   .6870559   .5961084
 17.        4   .7521145   .6469156
etc...

. plot ohat wais

 .646916 +
         | *
         |
         |     *
         |
         |         *
         |
         |             *
         |
         |                 *
         |                     *
    o    |
    h    |                         *
    a    |
    t    |                             *
         |
         |                                 *
         |
         |                                     *
         |
         |                                         *
         |
         |                                             *
         |
         |                                                 *
         |
         |                                                     *
-.165999 +                                                         *
         +-----------------------------------------------------------+
               4                         wais                      20

. plot ohat pihat wais

 .752115 +
         | B
         |
         |     B
         | A       B
         |     A
         |           A   B
         |               A
         |                   *
         |                       *
         |                           A
         |                           B   A
         |                               B   A
         |                                   B   A
         |                                       B   *
         |                                               *   B   B
         |                                                   A       B   B
         |                                                       A
         |                                                           A
-.165999 +                                                             A
         +-----------------------------------------------------------+
               4                         wais                      20
```

I used three different ways to show you how to run a linear probability model. Although we get the same coefficient estimates from these three approaches, you should notice by now that no. of observations is 54 instead of 17, which is the correct one. The data were purposely entered using the raw data format, as shown in Agresti's and usually how your data assistant enters data for your research. By using the layout for weighted linear regression, or `blogit` and `bprobit` format, you will obtain the correct d.f. = 15. It is essential to be aware of how and why the degrees of freedom "evolve" along the course of your analysis.

Nevertheless, the estimated coefficients are unbiased in all cases, and can be used to obtain predicted probabilities. By graph, you can see the linear probability model has poor predictions on the low and high ends of the WAIS scale – lower predicted probabilities on the low end (toward score 4) and the high end (toward score 20), approaching the extreme values (p=0 and p=1) too quickly.

Exercise 4.3 [6 points]

```
. tabi 1 11\13 53\16 42\15 27\7 11

           | col
       row |         1          2 |     Total
-----------+----------------------+----------
         1 |         1         11 |        12
         2 |        13         53 |        66
         3 |        16         42 |        58
         4 |        15         27 |        42
         5 |         7         11 |        18
-----------+----------------------+----------
     Total |        52        144 |       196

          Pearson chi2(4) =   6.8807   Pr = 0.142

. input change infil ntotal

change infil ntotal
  1. 1 1 12
  2. 2 13 66
  3. 3 16 58
  4. 4 15 42
  5. 5 7 18
  6. end

. replace change=change-1
(5 real changes made)

. blogit infil ntotal change
```

```
Logit Estimates                                   Number of obs =    196
                                                  chi2(1)       =   6.65
                                                  Prob > chi2   = 0.0099
Log Likelihood = -110.06697                       Pseudo R2     = 0.0293

------------------------------------------------------------------------------
_outcome |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  change |   .3896544   .1532464      2.543   0.011         .089297    .6900118
   _cons |   -1.81401   .3667985     -4.946   0.000      -2.532922   -1.095098
------------------------------------------------------------------------------

. test change=0

 ( 1)  change = 0.0

           chi2(  1) =    6.47
         Prob > chi2 =    0.0110

. bprobit infil ntotal change
```

```
Probit Estimates                                  Number of obs =    196
                                                  chi2(1)       =   6.73
                                                  Prob > chi2   = 0.0095
Log Likelihood =   -110.029                       Pseudo R2     = 0.0297

------------------------------------------------------------------------------
_outcome |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  change |   .2345331   .0912283      2.571   0.010        .0557289    .4133372
   _cons |  -1.101694   .2121698     -5.193   0.000       -1.517539   -.6858484
------------------------------------------------------------------------------

. test change=0

 ( 1)  change = 0.0

           chi2(  1) =    6.61
         Prob > chi2 =    0.0101
```

In the logit model, the Wald test reveals a χ^2 statistic 6.47, with 1 degree of freedom. We reject the null hypothesis $\beta = 0$. Using the log likelihood ratio test, we found the LRT $\chi^2 = 6.65$, with 1 degree of freedom. This statistic is obtained by 2 times the likelihood ratio difference between the current model and the constant model. We reject the null hypothesis that the constant model is a better model, i.e. the current model (with the β coefficient) fits data better. The results are similar to the trend test z^2=6.67 from Agresti's (p.102). It confirms that Pearson's χ^2 Goodness of Fit test, though useful, is a conservative index of testing association. The trend test, model fitting, and more detailed tests (here, likelihood ratio test and Wald χ^2 test) are usually needed. Similarly, you could fit a probit model to reach the same conclusion. (In the probit model, the Wald test reveals a χ^2 statistic 6.61, with 1 degree of freedom. We reject the null hypothesis $\beta = 0$. Using the log likelihood ratio test, we found the LRT $\chi^2 = 6.73$, with 1 degree of freedom.)

Exercise 4.6 [10 points]

```
. tabi 400 1380\416 1823\188 1168, all

           | col
       row |         1          2 |     Total
-----------+----------------------+----------
         1 |       400       1380 |      1780
         2 |       416       1823 |      2239
         3 |       188       1168 |      1356
-----------+----------------------+----------
     Total |      1004       4371 |      5375

          Pearson chi2(2) =  37.5663   Pr = 0.000
 likelihood-ratio chi2(2) =  38.3658   Pr = 0.000
                Cramer's V =   0.0836
                     gamma =   0.1770  ASE = 0.028
          Kendall's tau-b =   0.0786  ASE = 0.013

. input parent yes no

       parent        yes         no
  1. 2 400 1380
  2. 1 416 1823
  3. 0 188 1168
  4. end

. gen ntotal=yes+no
. gen p=yes/ntotal
. glm yes parent, f(binomial ntotal)
```

```
Residual df  =         1                          No. of obs =         3
Pearson X2   =   .569279                          Deviance   =  .5686519
Dispersion   =   .569279                          Dispersion =  .5686519

Binomial (N=ntotal) distribution, logit link
------------------------------------------------------------------------------
     yes |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  parent |   .2866273   .0470443      6.093   0.000         .194422    .3788325
   _cons |  -1.795024   .0657553    -27.299   0.000      -1.923902   -1.666146
------------------------------------------------------------------------------

. test parent=0

 ( 1)  parent = 0.0

           chi2(  1) =   37.12
         Prob > chi2 =    0.0000

. glmpred mu_log, mu
. glmpred xb_log, xb
. glmpred res_log, pearson
. gen pi_hat=mu_log/ntotal
. list parent yes ntotal p mu_log xb_log res_log pi_log

   parent  yes  ntotal         p    mu_log     xb_log    res_log    pi_log
1.      2  400    1380  .2247191  405.1729  -1.221769   -.292414  .2276252
2.      1  416    1823  .1857972  405.6542  -1.508397   .5676613  .1811765
3.      0  188    1168  .1386431  193.1729  -1.795024  -.4019126  .1424579
```

From Pearson's GOF test ($\chi^2 = 37.5663$), we reject the null hypothesis that these 6 cells are independent Poisson counts. From the Likelihood-ratio test ($\chi^2 = 38.3658$), we reject the null hypothesis that the independence model can explain as well as the saturated model. Therefore, there is evident information for us to model that the number of smoking parents can explain the smoking habits of these Arizona high school students.

Most of you have done logit models successfully by using the commands `logistic` or `logit`. Here I tried to show you how to reach the same conclusion by using `glm`. In addition, you can also fit the probit and complementary log-log models. They have similar estimates and results as the logit model. Due to the parsimony in interpretation, the logit model is preferred.

By fitting the logit model, we found that the additive model $logit(\hat{\pi}) = \alpha + \beta x$ has little deviance ($\chi^2 = .5686519$) from the saturated model (i.e. if you had run `xi:glm yes i.parent`). It is one of the indications that this model fits data well. The Wald test ($\chi^2 = 37.12$) allows us to reject the null hypothesis that $\beta = 0$. By looking at the observed counts vs. the fitted counts, the observed probabilities vs. the fitted probabilities, and the Pearson residuals (centered around zero and less than 2), data have shown strong evidence that the number of parents who smoke has a linear effect on the odds of whether their teens smoke. With a positive β, the model shows that as the number of parents who smoke increases, the odds of their teens have a smoking habit increases as well. For both-parent-smoke households, their teens have a probability of .225 to smoke; for one-parent-smoke households, their teens have a probability of .186 to smoke; and for neither-parent-smoke households, their teens have a probability of .139 to smoke.