IGP 304 (Spring 2006) Homework 4 Keys:

1. In a 1985 study of the relationship between contraceptive use and infertility, 89 out of 283 infertile women, compared with 640 out of 3833 control women, had used an IUD at some time in their lives.
    - Carry out a Pearson's chi-squared test of association between contraceptive use and infertility.

The two-way table of the data (with expected counts in parentheses) is

	IUD use	No IUD use	Total
Infertile	89 (50.1)	194 (232.9)	283
Control	640 (678.9)	3193 (3154.1)	3833
Total	729	3387	4116

The Pearson's chi-squared statistic is 39.35 with one degree of freedom. The corresponding p-value is 3.5×10^{-10}. Therefore, there is strong evidence of association of IUD use with infertility. [Stata command: `tabi 89 194 \ 640 3193, exp cchi chi2`]

    - Compute a 95% CI for the difference in the proportion of women who have ever used IUDs between the case and control groups.

The proportions are $p_1 = 89/283 = 0.3145$ for the infertility group and $p_0 = 640/3833 = 0.1670$ for the control group. The difference in proportion is $p_1 - p_0 = 0.1475$, with standard error $\sqrt{p_1(1-p_1)/n_1 + p_0(1-p_0)/n_0} = \sqrt{0.3145\cdot(1-0.3145)/283 + 0.1670\cdot(1-0.1670)/3833} = 0.0283$. Then the 95% CI for the difference in proportion is from $0.1475 - 1.96\times0.283 = 0.092$ to $0.1475 + 1.96\times0.283 = 0.203$. [Stata command: `csi 89 640 194 3193`]

    - Compute the odds ratio in favor of ever using an IUD for infertile women versus control women. Provide a 95% CI for the true odds ratio corresponding to your answer.

The odds ratio is OR = (89/194) / (640/3193) = 2.289. The log-odds ratio is log(OR) = 0.828, with standard error $\sqrt{1/d_1 + 1/h_1 + 1/d_0 + 1/h_0} = \sqrt{1/89 + 1/194 + 1/640 + 1/3193} = 0.135$. Then the 95% CI for log(OR) is from $0.828 - 1.96\times0.135 = 0.563$ to $0.828 + 1.96\times0.135 = 1.093$, and the 95% CI for OR is from $e^{0.563} = 1.756$ to $e^{1.093} = 2.983$. [Stata command: `csi 89 640 194 3193, or woolf`]

Alternatively, you could calculate the error factor $EF = e^{1.96\times0.135} = 1.303$, and the 95% CI for OR would be from $2.289/1.303 = 1.757$ to $2.289\times1.303 = 2.982$.

2. The presence of bacteria in the urine (bacteriuria) has been associated with kidney disease. Conflicting results have been reported from several studies concerning the possible role of oral contraceptives (OC) on bacteriuria. The following data were

collected in a population-based study of non-pregnant pre-menopausal women below the age of 50. The data are presented on an age-specific basis:

Age group	OC users		Non-OC users	
	with bacteriuria	total	with bacteriuria	total
16-19	1	84	9	281
20-29	16	284	22	552
30-39	6	96	34	623
40-49	4	18	13	482

- Input the data into your statistical software package. Use the software to create tables to make sure the data entered by you reflect the data collected in the study. (You may learn how to input data into Stata from the examples in my Stata Notes.)

[I had this question to make sure you are not misled to treat 84, 281, etc. as the counts for those without bacteriuria.] In Stata, you may input data this way:

```
clear
input bacteriuria ocuse agegroup freq
1 1 1 1
0 1 1 83
1 0 1 9
0 0 1 272
1 1 2 16
0 1 2 268
1 0 2 22
0 0 2 530
1 1 3 6
0 1 3 90
1 0 3 34
0 0 3 589
1 1 4 4
0 1 4 14
1 0 4 13
0 0 4 469
end
bysort agegroup: tabulate bacteriuria ocuse [freq=freq]
```

- Perform a significance test to examine the association between OC use and bacteriuria after controlling for age.

There are three options: (1) Use Mantel-Haenszel method. Assume there is a common odds ratio of the odds of bacteriuria in the OC users compared with that in the non-OC users across all the age groups. We can use Mantel-Haenszel method to test if this assumed common odds ratio is one (EMS page 184). The test statistic is $\chi^2_{MH} = 2.30$ with one degree of freedom. The corresponding p-value is 0.1295. Therefore, there is no strong evidence for overall association between OC use and bacteriuria. A look at the odds ratio estimates for the four age groups also will tell. [Stata command `cc bacteriuria ocuse [weight=freq], by(agegroup)`]

(2) Use logistic regression including age group as a categorical variable. The z-test for OC use variable has p-value 0.131, and we have the same conclusion as above. In fact, (1) and (2) treat data in similar ways and should give similar results. [Stata command `xi: logistic bacteriuria ocuse i.agegroup [weight=freq]`]

(3) Use logistic regression with age group as a continuous variable. The z-test for OC use has p-value 0.089, and we end up with the same conclusion. [Stata command `logistic bacteriuria ocuse agegroup [weight=freq]`]

- Estimate the odds ratio in favor of bacteriuria for OC users versus non-OC users after controlling for age. Provide a 95% CI for the odds ratio.

Again, there are three options, all giving similar results. (1) Mantel-Haenszel method will give an OR estimate of 1.425 with 95% CI as [0.894, 2.273]. (2) Logistic regression with age group as a categorical variable will give an OR estimate of 1.442 with 95% CI as [0.897, 2.317]. (3) Logistic regression with age group as a numerical variable will give an OR estimate of 1.502 with 95% CI as [0.940, 2.398].

- Is the association between bacteriuria and OC use comparable among different age groups? Why or why not?

Again, there are three options: (1) Mantel-Haenszel test of homogeneity will give a test statistic 11.64 with 3 degrees of freedom, with corresponding p-value 0.0087. (2) Testing for interaction effect between OC use and age group in a logistic regression with age group as a categorical variable will give a likelihood-ratio statistic 9.95 with 3 degrees of freedom, with corresponding p-value 0.019. (3) Testing for interaction effect between OC use and age group in a logistic regression with age group as a numerical variable will give a likelihood-ratio statistic 6.80 with 1 degree of freedom, with corresponding p-value 0.009. All lead to the same conclusion: the association between bacteriuria and OC use is not comparable among different age groups. You can see apparent heterogeneity by comparing the odds ratios across age groups.

Stata commands:

```
** Mantel-Haenszel test of homogeneity, odds ratio estimates for each group
cc bacteriuria ocuse [weight=freq], by(agegroup)
** Logistic regression with age group as a categorical variable
xi: logit bacteriuria ocuse i.agegroup [weight=freq]
est store A
xi: logit bacteriuria i.agegroup*ocuse [weight=freq]
lrtest A, stat
** Logistic regression with age group as a numerical variable
logit bacteriuria ocuse agegroup [weight=freq]
est store B
gen inter = ocuse * agegroup
logit bacteriuria ocuse agegroup inter [weight=freq]
lrtest B, stat
```

- Suppose you did not control for age in the preceding analyses. Calculate the crude (unadjusted for age) odds ratio in favor of bacteriuria for OC users versus non_OC users.

The crude OR is (27/455) / (78/1860) = 1.415.

- How do your answers between adjusting and not adjusting for age relate to each other? Explain any differences found.

The numbers happen to be similar. The OR estimate after adjusting for age is based on the assumption that the effect of OC use on bacteriuria risk as measured by odds ratio is the same for all age groups. The OR estimate without adjusting for age reflects the marginal effect of OC use on bacteriuria risk. Given the fact that there is heterogeneity of odds ratios across age groups, both estimates are based on overly simplified assumptions.

3. In a study on esophageal cancer, researchers collected data on 975 subjects. The variables collected were age, alcohol and tobacco usage, and esophageal cancer status. The data were tallied into a four-way table with 6 age groups, 4 alcohol usage levels, 4 tobacco usage levels, and 2 cancer status. In the data, the "patients" column has the counts for all four-way combinations. The "heavy" variable is an indicator variable for heavy alcohol consumption. The values and their meanings are:
   age: 1 (25-34), 2 (35-44), 3 (45-54), 4 (55-64), 5 (65-74), 6 (>= 75)
   alcohol: 1 (0-39), 2 (40-79), 3 (80-119), 4 (>= 120)
   tobacco: 1 (0-9), 2 (10-19), 3 (20-29), 4 (>= 30)
   cancer: 0 (No), 1 (Yes)
   heavy: 0 (< 80 gm), 1 (>= 80 gm)
   Cancer is the outcome variable and age group, alcohol usage, and tobacco usage are input variables.
   - Understanding the relationships among the input variables is an important part of statistical analysis. It allows you to gain insight into how the variables correlate with each other and if the results on some variables could be influenced by inclusion of some other variables in the analysis. Explore the relationship among age group, alcohol usage, and tobacco usage.

You can generate two-way tables for every pair of the variables, and compare the observed counts with expected counts under independence. Displaying row or column percentages may help. [Stata command `tab2 age alcohol tobacco [weight=patients], exp chi2 row`]

For age and alcohol usage, the Pearson's test for overall independence gives a p-value <0.001. A further examination of the two-way table suggests young and old people in the data set (age groups 1, 2, 5, 6) tend to drink less alcohol than expected under independence and people in age groups 3 and 4 tend to drink more alcohol than expected under independence.

For age and tobacco usage, the Pearson's test for overall independence gives a p-value 0.05 (marginal significance). A further examination of the two-way table suggests similar patterns as above, although not as strong.

For alcohol usage and tobacco usage, the Pearson's test for overall independence gives a p-value <0.001. A further examination of the two-way table suggests high alcohol usage tends to associate with high tobacco usage.

- For each input variable, create a two-way table between the input variable and the outcome variable, and carry out logistic regression analysis using the input variable as the only regression variable. Summarize results.

Stata output for analyzing only <u>age group</u>:

```
. tab cancer age [weight=patients], exp chi2 col
(frequency weights assumed)

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
| column percentage  |
+--------------------+
           |                                    age
    cancer |         1          2          3          4          5          6 |     Total
-----------+------------------------------------------------------------------+----------
         0 |       115        190        167        166        106         31 |       775
           |      92.2      158.2      169.3      192.4      128.0       35.0 |     775.0
           |     99.14      95.48      78.40      68.60      65.84      70.45 |     79.49
-----------+------------------------------------------------------------------+----------
         1 |         1          9         46         76         55         13 |       200
           |      23.8       40.8       43.7       49.6       33.0        9.0 |     200.0
           |      0.86       4.52      21.60      31.40      34.16      29.55 |     20.51
-----------+------------------------------------------------------------------+----------
     Total |       116        199        213        242        161         44 |       975
           |     116.0      199.0      213.0      242.0      161.0       44.0 |     975.0
           |    100.00     100.00     100.00     100.00     100.00     100.00 |    100.00

          Pearson chi2(5) =  97.0360   Pr = 0.000

. logistic cancer age [weight=patients]
(frequency weights assumed)

Logistic regression                               Number of obs   =        975
                                                  LR chi2(1)      =      87.29
                                                  Prob > chi2     =     0.0000
Log likelihood = -451.09778                       Pseudo R2       =     0.0882

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.775724   .1172932     8.69   0.000     1.560092    2.021159
------------------------------------------------------------------------------

. gen age2=age*age
. logistic cancer age age2 [weight=patients]
(frequency weights assumed)

Logistic regression                               Number of obs   =        975
                                                  LR chi2(2)      =     118.79
                                                  Prob > chi2     =     0.0000
Log likelihood = -435.34809                       Pseudo R2       =     0.1201

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   16.84312    7.91089     6.01   0.000     6.708561    42.28784
        age2 |   .7418488   .0442782    -5.00   0.000      .659949    .8339124
------------------------------------------------------------------------------
```

As expected, there is a strong association between age and esophageal cancer risk: As age increases, the proportion of cancer patients increases. If we treat age group as a continuous variable with a linear effect, the ratio of the odds of having cancer for an age group compared to the odds for an age group one level younger is about 1.78 (95% CI [1.56, 2.02]). If we compare the raw cancer rates with the predicted rates, we see over-prediction for younger and older age groups and under-prediction for middle age groups (see "predicted rate 1" in the table below, generated through cut-and-paste). This indicates a higher-order effect of age group. Thus we add a quadratic term into the regression analysis, resulting in a very good fit (see "predicted rate 2" below). In fact, analyses with higher-order effects won't fit the data significantly better.

Age group	1	2	3	4	5	6
Raw rate	0.86	4.52	21.60	31.40	34.16	29.55
Predicted rate 1	5.54	9.44	15.61	24.73	36.84	50.88
Predicted rate 2	0.88	5.77	18.80	25.86	32.54	35.60

Stata output for analyzing only <u>alcohol usage</u>:

```
 . tab cancer alcohol [weight=patients], exp chi2 col
(frequency weights assumed)

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
| column percentage  |
+--------------------+

           |                    alcohol
    cancer |         1          2          3          4 |     Total
-----------+--------------------------------------------+----------
         0 |       386        280         87         22 |       775
           |     329.9      282.2      109.7       53.3 |     775.0
           |     93.01      78.87      63.04      32.84 |     79.49
-----------+--------------------------------------------+----------
         1 |        29         75         51         45 |       200
           |      85.1       72.8       28.3       13.7 |     200.0
           |      6.99      21.13      36.96      67.16 |     20.51
-----------+--------------------------------------------+----------
     Total |       415        355        138         67 |       975
           |     415.0      355.0      138.0       67.0 |     975.0
           |    100.00     100.00     100.00     100.00 |    100.00

          Pearson chi2(3) = 158.9546   Pr = 0.000

. logistic cancer alcohol [weight=patients]
(frequency weights assumed)

Logistic regression                               Number of obs   =        975
                                                  LR chi2(1)      =     144.64
                                                  Prob > chi2     =     0.0000
Log likelihood =  -422.4246                       Pseudo R2       =     0.1462

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     alcohol |   2.848443    .266343    11.19   0.000     2.371461    3.421361
------------------------------------------------------------------------------
```

There is a strong association between alcohol usage and esophageal cancer risk: As alcohol usage increases, the proportion of cancer patients increases. If we treat alcohol usage as a continuous variable with a linear effect, the ratio of the odds of having cancer for an alcohol

usage level compared to the odds for one level lower is about 2.85 (95% CI [2.37, 3.42]). If we compare the raw cancer rates with the predicted rates, we see a good fit (see "predicted rate" in the table below). In fact, analyses with higher-order effects won't fit the data significantly better.

alcohol	1	2	3	4
raw rate	6.99	21.13	36.96	67.16
predicted rate	7.70	19.21	40.38	65.86

Stata output for analyzing only tobacco usage:

```
. tab cancer tobacco [weight=patients], exp chi2 col
(frequency weights assumed)

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
| column percentage  |
+--------------------+

           |                    tobacco
    cancer |         1          2          3          4 |     Total
-----------+--------------------------------------------+----------
         0 |       447        178         99         51 |       775
           |     417.3      187.6      104.9       65.2 |     775.0
           |     85.14      75.42      75.00      62.20 |     79.49
-----------+--------------------------------------------+----------
         1 |        78         58         33         31 |       200
           |     107.7       48.4       27.1       16.8 |     200.0
           |     14.86      24.58      25.00      37.80 |     20.51
-----------+--------------------------------------------+----------
     Total |       525        236        132         82 |       975
           |     525.0      236.0      132.0       82.0 |     975.0
           |    100.00     100.00     100.00     100.00 |    100.00

          Pearson chi2(3) =  29.3570   Pr = 0.000

. logistic cancer tobacco [weight=patients]
(frequency weights assumed)

Logistic regression                               Number of obs   =        975
                                                  LR chi2(1)      =      25.37
                                                  Prob > chi2     =     0.0000
Log likelihood = -482.05896                       Pseudo R2       =     0.0256

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tobacco |   1.474687   .1123596     5.10   0.000     1.270121    1.712201
------------------------------------------------------------------------------
```

Similarly, there is a strong association between tobacco usage and esophageal cancer risk: As tobacco usage increases, the proportion of cancer patients increases. If we treat tobacco usage as a continuous variable with a linear effect, the ratio of the odds of having cancer for a tobacco usage level compared to the odds for one level lower is about 1.47 (95% CI [1.27, 1.71]). If we compare the raw cancer rates with the predicted rates, we see a good fit (see "predicted rate" in the table below). In fact, analyses with higher-order effects won't fit the data significantly better.

tobacco	1	2	3	4
raw rate	14.86	24.58	25.00	37.80
predicted rate	15.53	21.33	28.56	37.09

- Carry out logistic regression analysis using all three input variables. Are the results similar or different from those in the last question? If they are different, why?

Logistic regression analysis using all three variables (linear effects only):

```
. logistic cancer age alcohol tobacco [weight=patients]
(frequency weights assumed)

Logistic regression                               Number of obs   =        975
                                                  LR chi2(3)      =     259.17
                                                  Prob > chi2     =     0.0000
Log likelihood = -365.15675                       Pseudo R2       =     0.2619

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   2.103813   .1720625     9.09   0.000     1.792218    2.469582
     alcohol |   3.011851   .3107276    10.69   0.000      2.46046    3.686808
     tobacco |   1.538566   .1445272     4.59   0.000     1.279845    1.849587
------------------------------------------------------------------------------
```

We saw earlier that all three input variables are associated with each other. But each of them is still significant even after adjusting for the linear effects of the other two variables. So, the basic conclusions are the same as above: all three variables influence esophageal cancer risk significantly. In fact, in a simple logistic regression considering only a single input variable, even if we had fitted a saturated model (i.e. treating the variable as categorical), the highest log-likelihood we could have reached would be -421 (for alcohol usage), much lower than -365 from above model fit with only 4 parameters.

We knew that age effect is more than linear. Thus, we fit a logistic regression with a quadratic term for age group and found the log-likelihood increased to -357 (significant if you do a likelihood-ratio test). The Stata output is:

```
. logistic cancer age age2 alcohol tobacco [weight=patients]
(frequency weights assumed)

Logistic regression                               Number of obs   =        975
                                                  LR chi2(4)      =     274.78
                                                  Prob > chi2     =     0.0000
Log likelihood = -357.35324                       Pseudo R2       =     0.2777

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   12.25292   6.149473     4.99   0.000     4.581845    32.76718
        age2 |   .7912292   .0506571    -3.66   0.000     .6979198    .8970137
     alcohol |   2.901154   .3033955    10.18   0.000     2.363493    3.561126
     tobacco |   1.551943   .1483482     4.60   0.000     1.286799    1.871721
------------------------------------------------------------------------------
```

- Explore the interaction effects among the input variables on the risk of esophageal cancer.

We first explore product interaction effects while treating all input variables as numerical variables (Stata output below). The log-likelihood only increased by 1.879, with corresponding

likelihood-ratio test statistic 3.758 and 4 more degrees of freedom (p-value 0.44). Thus, there is no evidence of product interaction when the input variables are treated as numerical ones.

```
. gen v1 = age*alcohol
. gen v2 = age*tobacco
. gen v3 = alcohol*tobacco
. gen v4 = v1*tobacco
. logistic cancer age age2 alcohol tobacco v1 v2 v3 v4 [weight=patients]
(frequency weights assumed)

Logistic regression                               Number of obs   =        975
                                                  LR chi2(8)      =     278.54
                                                  Prob > chi2     =     0.0000
Log likelihood = -355.47413                       Pseudo R2       =     0.2815

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   28.38399   22.74389     4.18   0.000      5.90217    136.5008
        age2 |   .7621933   .0537292    -3.85   0.000     .6638371    .8751224
     alcohol |   8.167358   6.810693     2.52   0.012     1.593205    41.86891
     tobacco |   3.259001   2.783776     1.38   0.167     .6109481    17.38459
          v1 |   .8269482   .1627261    -0.97   0.334      .562315    1.216121
          v2 |   .8976182    .180739    -0.54   0.592     .6049212    1.331939
          v3 |   .8005576   .2705387    -0.66   0.510     .4127997    1.552551
          v4 |   1.019672   .0864459     0.23   0.818     .8635691    1.203992
------------------------------------------------------------------------------
```

We may continue to explore the interaction effects while treating the input variables as categorical variables. But watch out on two issues: (1) When considering all combinations between two variables each with multiple categories, we may face the problem of sparseness. That is, the cell counts for some combinations may be very small or zero, making the model fit for those combinations fragile. (2) Another issue is interpretability. It often is hard to interpret results of full interaction between variables with more than two categories. [I did explore multiple ways of interaction. All failed to improve the model fit significantly.]