IGP 304 (Spring 2007) Homework 4 Keys

1. A study was performed looking at the risks of fractures in three rural Iowa communities according to whether their drinking water was "higher calcium" or "control" as determined by water samples. The following table presents the data comparing the rate of fractures over 5 years between the higher calcium versus the control communities for women ages 20-35 and 55-80, respectively.

Age 20-35	Total number of women	Number of women with fractures	Age 55-80	Total number of women	Number of women with fractures
Control	37	3	Control	121	11
Higher calcium	33	1	Higher calcium	148	21

   a. What test can be used to compare the fracture rates in these two communities while controlling for age?

The Mantel-Haenszel test for rate data is the best approach. But the Mantel-Haenszel test for odds ratios may be okay too, given the rate is relatively low. (Both lead to the same conclusion.)

   b. Implement the test and report a p-value (two-sided).

You can use either the formula in the textbook or Stata. I use Stata here.
(1) Rate based test.

```
clear
// the variable 'outcome' is used for Poisson regression
input calcium agegroup total fracture outcome
0 0 37 3 1
1 0 33 1 1
0 1 121 11 1
1 1 148 21 1
end
gen unit = total/fracture
stset unit [freq=fracture]
stmh calcium, compare(1,0) by(agegroup)
```

The Stata output is

```
Overall estimate controlling for agegroup

------------------------------------------------------------
     RR       chi2        P>chi2      [95% Conf. Interval]
------------------------------------------------------------
  1.336       0.71        0.4009         0.678      2.632
------------------------------------------------------------

Approx test for unequal RRs (effect modification): chi2(1) =     1.57
                                                   Pr>chi2 =   0.2108
```

We test if the rate ratio comparing high calcium and control groups with respect to risk of fracture is different from 1. Since the p-value for this test is 0.40, we cannot claim there is a difference in fracture rate between the two calcium groups.

(2) Odds ratio based test. Note that we need to calculate the number of women without fractures.

```
clear
```

```
input calcium fracture agegroup freq
0 0 0 34
0 1 0 3
1 0 0 32
1 1 0 1
0 0 1 110
0 1 1 11
1 0 1 127
1 1 1 21
end
mhodds fracture calcium agegroup [weight=freq], compare(1,0)
```

The Stata output is:

```
Mantel-Haenszel estimate of the odds ratio
Comparing calcium==1 vs. calcium==0, controlling for agegroup

    ------------------------------------------------------------------
     Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
    ------------------------------------------------------------------
       1.382093       0.79        0.3739         .675191   2.829095
  ------------------------------------------------------------------
```

We test if the odds ratio comparing high calcium and control groups with respect to risk of fracture is different from 1. Since the p-value for this test is 0.37, we cannot claim there is a difference in fracture rate between the two calcium groups.

c. Estimate the odds ratio relating higher calcium and fractures while controlling for age.

This can be obtained from the results in b(2). The estimated odds ratio comparing high calcium and control groups with respect to risk of fracture is 1.38.

d. Provide a 95% confidence interval for the estimate obtained.

This can be obtained from the results in b(2). The 95% confidence interval for the odds ratio comparing high calcium and control groups with respect to risk of fracture is [0.68, 2.83].

e. Is there an alternative test?

Two tests have been covered in b. We could carry out the regression analyses on rate or on odds.
(1) For rate, the Poisson regression Stata command is `poisson outcome calcium agegroup [freq=fracture], e(unit) irr`, with Stata output

```
Poisson regression                                Number of obs   =         36
                                                  LR chi2(2)      =       3.02
                                                  Prob > chi2     =     0.2209
Log likelihood =  -36.80247                       Pseudo R2       =     0.0394

------------------------------------------------------------------------------
     outcome |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     calcium |   1.331852   .4558934     0.84   0.402     .6809092    2.605088
    agegroup |   2.035776   1.080955     1.34   0.181     .7190504    5.763689
        unit | (exposure)
------------------------------------------------------------------------------
```

The p-value 0.402 that is associated with the variable calcium indicates the effect of this variable is not significant after adjusting for the effect of age group.

(2) For odds, the logistic regression Stata command is `logit fracture calcium agegroup [weight=freq]`, with Stata output

```
Logistic regression                                     Number of obs   =        339
                                                        LR chi2(2)      =       3.34
                                                        Prob > chi2     =     0.1879
Log likelihood = -113.07468                             Pseudo R2       =     0.0146

------------------------------------------------------------------------------
    fracture |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     calcium |   .3218858   .3621852     0.89   0.374    -.3879842    1.031756
    agegroup |   .7774358   .5492447     1.42   0.157    -.2990641    1.853936
       _cons |   -2.96656   .5514725    -5.38   0.000    -4.047427   -1.885694
------------------------------------------------------------------------------
```

The p-value 0.374 that is associated with the variable calcium indicates the effect of this variable is not significant after adjusting for the effect of age group.

2. A longitudinal study was conducted among children in the Greater Boston Otitis Media Study. Based on all doctor visits during the first year of life, children were classified as having ≥1 episodes versus 0 episodes of otitis media (OTM). A separate classification was performed for the right and left ears. Several risk factors were studies as possible predictors of OTM. One such risk factor was a sibling history of ear infection, with relevant data displayed in the following table.

Sib ear infection = yes			Sib ear infection = no		
Right ear	Left ear	n	Right ear	Left ear	n
−	−	76	−	−	115
+	−	21	+	−	20
−	+	20	−	+	18
+	+	77	+	+	91
Total		194	Total		244

   a. Assess whether a sib history of ear infection is associated with OTM incidence in the first year of life.

We can do a test on whether rate difference is different from zero, or a test on whether rate ratio is different from one. Here is the result for rate difference. Here, $d_1 = 21 + 20 + 77 = 118$, $T_1 = 194$, $d_2 = 20 + 18 + 91 = 129$, and $T_2 = 244$. The rate difference is estimated to be $118/194 – 129/244 = .61 – .53 = .08$. The corresponding standard error is estimated to be $\sqrt{(118/194^2 + 129/244^2)} = 0.0728$. Thus the test statistic is $(.08 – 0)/0.0728 = 1.099$. Comparing with the standard normal distribution, the p-value is 0.27. So, there is not strong evidence for association between a sib history of ear infection with OTM incidence in the first year of life.

   b. Provide a 95% confidence interval for the true difference in incidence rates for those children with sibs between those children with and without a sib history of ear infection.

So, the 95% confidence interval for the true difference in incidence rates is $0.08 \pm 1.96 \times 0.0728 =$ [-0.06, 0.22]. [The Stata command is `iri 118 129 194 244` will give the same results.]

3. Improving control of blood glucose levels in an important motivation for the use of insulin pumps by diabetic patients. However, certain side effects have been reported with pump therapy. The following table provides data on the occurrence of diabetic ketoacidosis (DKA) in patients before and after start of pump therapy.

	Before pump therapy	
After pump therapy	No DKA	DKA
No DKA	128	7
DKA	19	7

a. What is the appropriate procedure to test if the rate of DKA is different before and after start of pump therapy?

This is paired data. McNemar's test is appropriate.

b. Perform the significance test and report a p-value.

The McNemar's test statistic is $(7 - 19)^2/(7 + 19) = 5.54$. Comparing to the chi-squared distribution with one degree of freedom, the p-value is 0.019. Thus, the rate of DKA is significantly different before and after pump therapy.

NOTE: Below is not part of the homework. I leave it here just for those who are interested.

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 "esophageal.csv", 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. Age group, alcohol usage, and tobacco usage are input variables.

a. 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.

b. 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 |
```

c. 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
------------------------------------------------------------------------------
```

d. 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.]