Lab #11 Discussion

Question 1. *What do the results from* `tabodds` *indicate about the associations between these three variables and esophageal cancer*?

```
. tabodds case alcgp, or
-------------------------------------------------------------------------------
      alcgp |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
------------+------------------------------------------------------------------
   0-39 gm~y |    1.000000          .          .               .           .
       40-79 |    3.556034      32.55       0.0000       2.232193    5.665003
      80-119 |    7.693966      73.98       0.0000       4.439351   13.334630
        120+ |   27.155172     159.99       0.0000      12.476332   59.104180
-------------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(3)  =   157.99
                                  Pr>chi2  =   0.0000
Score test for trend of odds:     chi2(1)  =   151.98
                                  Pr>chi2  =   0.0000
. tabodds case tobgp, or
-------------------------------------------------------------------------------
      tobgp |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
------------+------------------------------------------------------------------
   0-9 gm/~y |    1.000000          .          .               .           .
       10-19 |    1.871507      10.54       0.0012       1.274043    2.749151
       20-29 |    1.934066       8.01       0.0047       1.214629    3.079632
         30+ |    3.491202      25.41       0.0000       2.078915    5.862908
-------------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(3)  =    29.58
                                  Pr>chi2  =   0.0000
Score test for trend of odds:     chi2(1)  =    27.30
                                  Pr>chi2  =   0.0000
. tabodds case agegp, or
-------------------------------------------------------------------------------
      agegp |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
------------+------------------------------------------------------------------
      25-34 |    1.000000          .          .               .           .
      35-44 |    5.447368       3.18       0.0743       0.671840   44.167984
      45-54 |   31.676647      26.29       0.0000       3.943092  254.472873
      55-64 |   52.650602      43.21       0.0000       6.304213  439.719592
      65-74 |   59.669811      46.18       0.0000       6.674741  533.426917
        75+ |   48.225806      32.67       0.0000       4.682406  496.695189
-------------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(5)  =    96.94
                                  Pr>chi2  =   0.0000
Score test for trend of odds:     chi2(1)  =    83.37
                                  Pr>chi2  =   0.0000
```

The results show increasing esophageal cancer risk with increasing levels of the categories of each of the three categorized predictors. Lower confidence bounds for each estimated OR (for all three) indicate significantly elevated disease odds (above the baseline category) for all levels except for the age group 35-44. Further, the global test of homogeneity and trend are significant for all three predictors. The former test evaluates the hypothesis that odds ratios are constant across predictor levels. (Note that the interpretation of this test differs from the analogous test used to investigate differences in odds ratios for one predictor compared across levels of a second predictor.) Rejection indicates unspecified differences across levels. The trend test is based on the more specific alternative that the log odds increase linearly with increasing levels of the

predictors (assuming the levels are equally spaced). We can conclude that we have fairly strong evidence for increased risk (as measured by the odds) associated with increasing levels of all three predictors.

Question 2. *Do the above results indicate that age has a confounding influence on the measured associations between tobacco and/or alcohol consumption and the outcome*?

```
. tabodds case tobgp, or adjust(agegp)
Mantel-Haenszel odds ratios adjusted for agegp
------------------------------------------------------------------------------
       tobgp |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   0-9 gm/~y |    1.000000          .          .               .           .
       10-19 |    1.830268       8.92       0.0028      1.223417    2.738134
       20-29 |    1.980508       7.50       0.0062      1.202969    3.260607
         30+ |    6.528439      39.26       0.0000      3.318106   12.844832
------------------------------------------------------------------------------
Score test for trend of odds: chi2(1)  =   34.17
                              Pr>chi2  =  0.0000

. tabodds case alcgp, or adjust(agegp)
Mantel-Haenszel odds ratios adjusted for agegp
------------------------------------------------------------------------------
       alcgp |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   0-39 gm~y |    1.000000          .          .               .           .
       40-79 |    4.286018      37.49       0.0000      2.578709    7.123699
      80-119 |    7.829443      58.20       0.0000      4.182181   14.657465
        120+ |   28.533878     139.50       0.0000     12.130672   67.117648
------------------------------------------------------------------------------
Score test for trend of odds: chi2(1)  =  134.09
                              Pr>chi2  =  0.0000
```

The odds ratios estimated in the above `tabodds` commands are adjusted for the presence of age (as measured by the categorical predictor `agegp`). This means that the estimates are obtained by separately estimating the odds ratios for each level of the exposure variables (`alcgp` and `tobgp`) in each of the six strata defined by `agegp`, and combining the estimates into a single weighted estimate using the Mantel-Haenszel (M-H) procedure. These age-adjusted estimates correspond very closely to those obtained in a logistic regression model including both predictors. For example, the following model for `tobgp` and `agegp` yields age-adjusted odds ratio estimates for tobacco consumption:

```
. logistic case i.tobgp i.agegp
Logistic regression                               Number of obs   =        975
                                                  LR chi2(8)      =     157.69
                                                  Prob > chi2     =     0.0000
Log likelihood = -415.89796                       Pseudo R2       =     0.1594
------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Itobgp_2 |   1.835637   .3782121     2.95   0.003     1.225763    2.748952
   _Itobgp_3 |   1.946139   .4880221     2.66   0.008     1.190482    3.181447
   _Itobgp_4 |   5.706004   1.725556     5.76   0.000     3.154422    10.32154
   _Iagegp_2 |   6.009024   6.405328     1.68   0.093      .743814    48.54489
   _Iagegp_3 |   36.04627   36.94488     3.50   0.000     4.83547    268.7088
   _Iagegp_4 |   61.51665   62.82635     4.03   0.000      8.3113    455.3197
   _Iagegp_5 |   83.19723    85.4861     4.30   0.000     11.10426    623.3445
   _Iagegp_6 |   60.18631   64.24026     3.84   0.000     7.429544    487.5659
------------------------------------------------------------------------------
```

The differences between the adjusted and unadjusted estimates (obtained in Question #1) are not large, but the adjusted results yield larger estimated odds ratios for the highest levels of tobacco and alcohol consumption, indicating some confounding influence. You may have noticed that we haven't ruled out the possibility of interaction between age and these two exposures in their influence on cancer risk. Evaluation using frequency table methods would be awkward given the large number of categories. Logistic regression (possibly using appropriately modeled continuous versions of these predictors) would be simpler and more interpretable.

Question 3. *Interpret the results of the likelihood ratio tests. Which risk factor results in the biggest change in log-likelihood?*

```
. quietly logistic case i.tobgp i.alcgp i.agegp
. estimates store M0
. quietly xi:logistic case i.tobgp i.agegp
. lrtest M0
likelihood-ratio test                                 LR chi2(3)  =    127.52
(Assumption: . nested in M0)                          Prob > chi2 =    0.0000
. quietly logistic case i.alcgp i.agegp
. lrtest M0
likelihood-ratio test                                 LR chi2(3)  =     23.69
(Assumption: . nested in M0)                          Prob > chi2 =    0.0000
```

Clearly alcohol makes a bigger likelihood contribution than tobacco in a model including all three predictors. This would be more difficult to evaluate using the individual coefficients, associated Z statistics and Wald test results.

Question 4. *Create a dichotomous indicator of heavy alcohol use with the statements below, and evaluate it as a predictor in the model including* `tobgp` *and* `agegp`. *Use the likelihood ratio test to compare this model to the model including the four-level categorical predictor* `tobgp` *(fitted above). Do we lose anything from modeling alcohol consumption as a binary predictor?*

```
. quietly logistic case alcb i.agegp i.tobgp
. lrtest M0
likelihood-ratio test                                 LR chi2(2)  =     57.59
(Assumption: . nested in M0)                          Prob > chi2 =    0.0000
```

The two-level version of alcohol is derived from the four-level version by collapsing categories of, so the model requested is nested within the full model fitted above. The LR test results indicate a significant change from the model incorporating four levels of alcohol consumption. We conclude that the model based on the reduced variable fits less well than the full model, and that some information is lost in the simplification. Based on these results, we would prefer to retain more categories of alcohol consumption (or use the continuous version) in further analyses. Of course, this doesn't preclude investigating specific hypothesis about heavy drinkers if this category of individuals is of interest.

Question 5. *Compare the results of this model to the model fitted above including only the indicator* `est` *of estrogen use and comment on the apparent effects of including the additional risk factors on the association between the outcome and estrogen use.*

Assuming that estrogen use is the predictor of primary interest, comparison of the two regression models (see below) indicates potential confounding influence of the other predictors on the association between `est` and the outcome.

```
. clogit case est hyp non gall obes, group(set) or
Conditional (fixed-effects) logistic regression   Number of obs   =        315
                                                  LR chi2(5)      =      49.05
                                                  Prob > chi2     =     0.0000
Log likelihood = -76.868906                       Pseudo R2       =     0.2419
------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         est |   6.782407   3.086647     4.21   0.000     2.779747    16.54864
         hyp |   .8248199   .2903746    -0.55   0.584      .413711    1.644452
         non |   1.951244   .9958351     1.31   0.190      .717618    5.305545
        gall |   3.633828   1.503164     3.12   0.002     1.615316    8.174687
        obes |   1.610575   .5871015     1.31   0.191     .7883029    3.290552
------------------------------------------------------------------------------

. clogit case est, group(set) or
Conditional (fixed-effects) logistic regression   Number of obs   =        315
                                                  LR chi2(1)      =      35.35
                                                  Prob > chi2     =     0.0000
Log likelihood =  -83.72159                       Pseudo R2       =     0.1743
------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         est |   7.954681   3.347525     4.93   0.000     3.486714    18.14802
------------------------------------------------------------------------------

. gen galest = gall*est
. clogit case est hyp non gall obes galest, group(set) or
Conditional (fixed-effects) logistic regression   Number of obs   =        315
                                                  LR chi2(6)      =      53.38
                                                  Prob > chi2     =     0.0000
Log likelihood = -74.704446                       Pseudo R2       =     0.2632
------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         est |   12.26034   7.651924     4.02   0.000     3.607865    41.66338
         hyp |   .8183825   .2849189    -0.58   0.565     .4136296    1.619202
         non |   2.104605   1.125999     1.39   0.164     .7375024    6.005895
        gall |   18.75607   16.72336     3.29   0.001     3.267312    107.6696
        obes |   1.564086   .5779477     1.21   0.226     .7581096    3.226929
      galest |   .1254424   .1255502    -2.07   0.038     .0176405    .8920255
```

However, further analysis reveals a significant negative interaction between the indicator of gall bladder disease (`gall`) estrogen use (`est`) (shown below). Note the use of `lincom` (below) giving the estimated odds ratios for the association between `est` and case status for women with and without gall bladder disease. Note the large standard error and upper 95% confidence bound for the coefficients for `gall` and `est` in the model above. The two tables below show the reason for the uncertainty: the relatively small number of cases (7) reporting no estrogen use.

```
. lincom est, or
 ( 1)  est = 0
------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   12.26034   7.651924     4.02   0.000     3.607865    41.66338
------------------------------------------------------------------------------
```

```
. lincom est + galest, or
 ( 1)  est + galest = 0
------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.537967   1.243321     0.53   0.594     .3153693    7.500227
------------------------------------------------------------------------------
```

```
. tab case gall if est==0
      Case |
    status |
  (1=case, |  Gallbladder disease
0=control) |        No        Yes |     Total
-----------+----------------------+----------
         0 |       117          8 |       125
         1 |         3          4 |         7
-----------+----------------------+----------
     Total |       120         12 |       132
```

```
. tab case gall if est==1
      Case |
    status |
  (1=case, |  Gallbladder disease
0=control) |        No        Yes |     Total
-----------+----------------------+----------
         0 |       111         16 |       127
         1 |        43         13 |        56
-----------+----------------------+----------
     Total |       154         29 |       183
```