Models In Epidemiology And Biostatistics
Gordon Hilton Fick

Esophageal Cancer Study

Data from a case-control study of esophageal cancer and tobacco consumption (Breslow & Day, 1980; Tuyns, 1977) are available in bd1.dta. We begin with interest in the relation between tobacco consumption (tobhigh: 1 = 20+ g/day, 0 = less than 20 g/day) and esophageal cancer (case: 1 = case, 0 = control) while considering alcohol consumption (alchigh: 1 = 80+ g /day, 0 = < 80 g/day) as a potential confounder/modifier.

```
. cc case tobhigh,by(alchigh)

         alchigh |       OR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
               0 |   2.061417      1.266494   3.305202     11.54545 (exact)
               1 |   1.699605      .8631744   3.362881     7.404878 (exact)
-----------------+-------------------------------------------------
           Crude |   1.960784      1.361333   2.806287              (exact)
    M-H combined |   1.920038      1.326402   2.779359
-----------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     0.24  Pr>chi2 = 0.6256

                  Test that combined OR = 1:
                              Mantel-Haenszel chi2(1) =     12.29
                                                Pr>chi2 =    0.0005

. cc case tobhigh if alchigh==0
                                                              Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |        34          70  |        104       0.3269
        Controls |       127         539  |        666       0.1907
-----------------+------------------------+------------------------
           Total |       161         609  |        770       0.2091
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         2.061417       |    1.266494    3.305202 (exact)
 Attr. frac. ex. |         .5148969       |    .2104186    .6974466 (exact)
 Attr. frac. pop |         .1683317       |
                 +-------------------------------------------------
                             chi2(1) =    10.10  Pr>chi2 = 0.0015

. cc case tobhigh if alchigh==1
                                                              Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |        30          66  |         96       0.3125
        Controls |        23          86  |        109       0.2110
-----------------+------------------------+------------------------
           Total |        53         152  |        205       0.2585
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         1.699605       |    .8631744    3.362881 (exact)
 Attr. frac. ex. |         .4116279       |   -.1585145    .7026359 (exact)
 Attr. frac. pop |         .1286337       |
                 +-------------------------------------------------
                             chi2(1) =     2.74  Pr>chi2 = 0.0977
```

Breslow and Day [Volume 1] provide considerably more analyses based grouped data: alc and tob

each with 4 levels:

```
. table alc case tob

-------------------------------------------------------------------------------
          |                                tob and case
          | ------ 0-9 -----    ----- 10-19 ----    ----- 20-29 ----    ------ 30+ -----
      alc | control     case    control     case    control     case    control     case
----------+--------------------------------------------------------------------------
     0-39 |     252        9         74       10         37        5         23        5
    40-79 |     145       34         68       17         47       15         20        9
   80-119 |      42       19         30       19         10        6          5        7
     120+ |       8       16          6       12          5        7          3       10
-------------------------------------------------------------------------------
```

First, simultaneous stratification on alc and tob:

```
. egen ta=group(tob alc)

. tabodds case ta,or

-------------------------------------------------------------------------------
          ta |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+-----------------------------------------------------------------
           1 |    1.000000          .          .              .          .
           2 |    6.565517      29.04       0.0000      2.979617   14.466964
           3 |   12.666667      47.63       0.0000      4.993105   32.133201
           4 |   56.000000     109.38       0.0000     14.125754  222.005855
           5 |    3.783784       8.71       0.0032      1.462533    9.789192
           6 |    7.000000      25.20       0.0000      2.887903   16.967329
           7 |   17.733333      62.46       0.0000      6.619385   47.507599
           8 |   56.000000      96.34       0.0000     12.906495  242.978441
           9 |    3.783784       5.85       0.0156      1.186680   12.064768
          10 |    8.936170      31.25       0.0000      3.522942   22.667175
          11 |   16.800000      34.01       0.0000      4.598697   61.373908
          12 |   39.200000      62.41       0.0000      8.627518  178.109165
          13 |    6.086957      11.35       0.0008      1.834218   20.199907
          14 |   12.600000      34.00       0.0000      4.200670   37.793967
          15 |   39.200000      62.41       0.0000      8.627518  178.109165
          16 |   93.333333     103.21       0.0000     14.766136  589.938431
-------------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(15)  =   173.42
                                  Pr>chi2   =   0.0000

Score test for trend of odds:     chi2(1)   =    57.58
                                  Pr>chi2   =   0.0000
```

Then some intermediate steps with adjustment:

```
. tabodds case alc,or adjust(tob age)

Mantel-Haenszel odds ratios adjusted for tob and age

-------------------------------------------------------------------------------
         alc |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+-----------------------------------------------------------------
        0-39 |    1.000000          .          .              .          .
       40-79 |    4.049572      33.40       0.0000      2.422136   6.770484
      80-119 |    7.490421      51.90       0.0000      3.929177  14.279432
        120+ |   22.797747     112.60       0.0000      9.810159  52.979492
-------------------------------------------------------------------------------
Score test for trend of odds: chi2(1)  =  118.37
                              Pr>chi2  =  0.0000
```

```
. tabodds case tob,or adjust(alc age)

Mantel-Haenszel odds ratios adjusted for alc and age

------------------------------------------------------------------------------
         tob |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         0-9 |    1.000000          .            .               .           .
       10-19 |    1.505990       3.49       0.0619        0.976845    2.321766
       20-29 |    1.556719       2.63       0.1050        0.907592    2.670114
         30+ |    8.099868      29.72       0.0000        3.303340   19.861064
------------------------------------------------------------------------------
Score test for trend of odds: chi2(1)  =   20.07
                              Pr>chi2  =  0.0000
```

Then they offer a series of models using this grouped data [with discussion]

```
. logit case i.alc##i.tob

Logistic regression                               Number of obs     =        975
                                                  LR chi2(15)       =     166.89
                                                  Prob > chi2       =     0.0000
Log likelihood =  -411.2976                       Pseudo R2         =     0.1687

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         alc |
       40-79 |   1.881831   .3890857     4.84   0.000     1.119237    2.644425
      80-119 |   2.538974   .4376305     5.80   0.000     1.681234    3.396714
        120+ |   4.025352   .5500721     7.32   0.000      2.94723    5.103473
             |
         tob |
       10-19 |   1.330725   .4781139     2.78   0.005     .3936385     2.26781
       20-29 |   1.330725   .5848986     2.28   0.023     .1843443    2.477105
         30+ |   1.806148   .5987968     3.02   0.003      .632528    2.979768
             |
      alc#tob |
 40-79#10-19 |  -1.266646   .5817479    -2.18   0.029    -2.406851   -.1264407
 40-79#20-29 |  -1.022449   .6829041    -1.50   0.134    -2.360916    .3160187
   40-79#30+ |  -1.154283   .7456387    -1.55   0.122    -2.615708    .3071423
80-119#10-19 |  -.9942523    .625299    -1.59   0.112    -2.219816    .2313113
80-119#20-29 |  -1.048319   .8277766    -1.27   0.205    -2.670732    .5740929
  80-119#30+ |  -.6764453   .8819614    -0.77   0.443    -2.405058    1.052167
  120+#10-19 |  -1.330725   .8161451    -1.63   0.103     -2.93034    .2688906
  120+#20-29 |  -1.687399   .9340576    -1.81   0.071    -3.518119    .1433197
    120+#30+ |  -1.295323   .9896418    -1.31   0.191    -3.234985    .6443398
             |
        _cons |  -3.332205   .3392335    -9.82   0.000     -3.99709   -2.667319
------------------------------------------------------------------------------
```

They proceed to provide a number of analyses using the actual [ungrouped] values of alc and tob. They note that there were 163 values for alc ranging from 0 to 268 g/day. They call alc a "true continuous" variable. There were in fact 9 levels for tob ranging from 0 to 50+ g/day. They call tob "discrete".

They consider transformations of alc and tob using log(alc+1) and log(tob+1). They then consider models based on linearity on these transformed scales.