Models In Epidemiology And Biostatistics
Gordon Hilton Fick

Session 4 : The First Example With Annotation

```
. use Session_2_Examples.dta

. table nd ne ga [fw=ct1]

------------------------------------------------------------
          |            Gender/Age Groups and Exposure
          | --- YF ---     --- OF ---     --- YM ---     --- OM ---
  Disease |    E  notE        E  notE        E  notE        E  notE
----------+-------------------------------------------------------
        D |   13   101       39    84       56    80        4   124
     notD |   33   153       21   156       17   147       41   131
------------------------------------------------------------

. cs dis exp [fw=ct1], by(gender age) or

      gender age |       OR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
             0 0 |    .5967597     .3025872   1.178593        11.11 (Cornfield)
             0 1 |     3.44898     1.914292   6.211071         5.88 (Cornfield)
             1 0 |    6.052941     3.314122   11.04433     4.533333 (Cornfield)
             1 1 |    .1030685     .0374139   .2848443     16.94667 (Cornfield)
-----------------+-------------------------------------------------
           Crude |    1.508997     1.127973   2.018743
    M-H combined |    1.458193     1.096429   1.939319
-----------------------------------------------------------------
Test of homogeneity (M-H)      chi2(3) =   59.855  Pr>chi2 = 0.0000

                   Test that combined OR = 1:
                               Mantel-Haenszel chi2(1) =      7.06
                                                Pr>chi2 =    0.0079

. disp 13*153/(33*101)
.59675968
```

The odds ratios here are the odds of disease among the exposed divided by the odds of disease among the unexposed.

```
. disp chi2tail(3,59.855)
6.313e-13
```

We can see that, for the young females, there is no evidence of a disease-exposure relationship as the CI for this OR covers the null OR of 1. However, for the old females and the young males, there is a disease-exposure relationship. In both cases, the exposure is a risk. Curiously, for the old males, the exposure is protective.

The omnibus test for modification has chi2(3) =59.855 $p < 0.0001$ which indicates that there is strata modification. This test does not, per se, tell us how age and/or gender modify.

Since modification has been detected, we should not address confounding here. A comparison of crude and adjusted estimates of the OR would not be warranted. Further, the MH chi2(1) has no meaningful interpretation here.

Let us now consider the 'one-at-a-time' assessments:

```
. cs dis exp [fw=ct1], by(gender) or

          gender |       OR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
               0 |    1.608408     1.056361   2.449062        16.65 (Cornfield)
               1 |    1.409736     .9429081   2.107711        19.72 (Cornfield)
-----------------+-------------------------------------------------
           Crude |    1.508997     1.127973   2.018743
    M-H combined |    1.500687     1.120924   2.009112
-----------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =    0.196  Pr>chi2 = 0.6582

                   Test that combined OR = 1:
                               Mantel-Haenszel chi2(1) =      7.49
                                                Pr>chi2 =    0.0062

. cs dis exp [fw=ct1], by(age) or

             age |       OR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
               0 |    2.287293     1.523055   3.434835     15.08333 (Cornfield)
               1 |    .9569634     .6250081   1.465417     21.49333 (Cornfield)
-----------------+-------------------------------------------------
           Crude |    1.508997     1.127973   2.018743
    M-H combined |    1.505559     1.125901    2.01324
-----------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =    8.341  Pr>chi2 = 0.0039

                   Test that combined OR = 1:
                               Mantel-Haenszel chi2(1) =      7.71
                                                Pr>chi2 =    0.0055
```

Notice that neither 'one-at-a-time' [aka 'univariate'] analyses display the results correctly. The gender only analysis clearly involves combining OR estimates that are different. These two meaningless numbers are 'close' and we might then look a meaningless MH combined number that is quite close to the crude.

Similarly the age only analysis combines OR estimates that are different. These two meaningless numbers are now 'different' and so we get incorrectly determined 'evidence' of age modification that still misses the real issues in play.

So we are seeing that the simultaneous stratification on both age and gender is required here.

Now, let us consider modeling.
We need a few 'new' variables:

```
. gen genage=gender*age
. gen gae=genage*expo
. gen ge=gender*expo
. gen ae=age*expo
```

```
. logit dis age gender genage expo ae ge gae [fw=ct1]

Logistic regression                               Number of obs     =      1,200
                                                  LR chi2(7)        =      91.93
                                                  Prob > chi2       =     0.0000
Log likelihood = -769.40291                       Pseudo R2         =     0.0564

------------------------------------------------------------------------------
         dis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.2037218   .1864188    -1.09   0.274    -.5690959    .1616523
      gender |  -.1930885   .1890494    -1.02   0.307    -.5636185    .1774414
      genage |    .757212   .2641075     2.87   0.004     .2395708    1.274853
        expo |  -.5162408   .3516576    -1.47   0.142    -1.205477    .1729954
          ae |   1.754319   .4639376     3.78   0.000     .8450181     2.66362
          ge |   2.316785   .4686645     4.94   0.000      1.39822    3.235351
         gae |  -5.827225   .7754417    -7.51   0.000    -7.347063   -4.307388
       _cons |  -.4153174   .1282066    -3.24   0.001    -.6665978    -.164037
------------------------------------------------------------------------------

. disp log(101/153)
-.4153174
```

By referring directly to the data, we can check that $b_0 = -0.4153174$ is an estimate of the log of odds of disease for the unexposed young females.

```
. disp log(84/156/(101/153))
-.2037218
```

So, by direct calculation, we have that $b_1 = -0.2037218$ is an estimate of the log of the ratio of the odds of disease for the old relative to the odds of disease for young but specific to the unexposed females

```
. disp exp(-0.5162408)
.59675968
```

We have verified that $b_4 = -0.5162408$ is an estimate of the log odds ratio for the young females.

```
. disp exp(-5.827225)
.00294624

. disp (0.1030685/6.052941)/(3.44898/0.5967597)
.00294624
```

So $b_7 = -5.827225$ is an estimate of the log of the ratio of 2 ratios of odds ratios. Telling us about whether age modification is modified by gender. [and vice versa]

```
. disp exp(2.316785)
10.143012

. disp 6.052941/0.5967597
10.143012
```

So $b_6 = 2.316785$ is an estimate of the log of ratio of odds ratios [males relative to female] specific to the young. Here telling us about gender modification but specific to the young.

```
. disp exp(1.754319)
5.7795106

. disp 3.44898/0.596797
5.779151
```

So $b_5 = 1.754319$ is an estimate of the log of the ratio of odds ratios [old relative to young] specific to the females. Here telling us about age modification but specific to the females.

Let us look now at some other models and explore the challenges implicit:

```
. logit dis age gender genage expo ae ge [fw=ct1]


Logistic regression                               Number of obs     =      1,200
                                                  LR chi2(6)        =      18.93
                                                  Prob > chi2       =     0.0043
Log likelihood =  -805.9016                       Pseudo R2         =     0.0116

------------------------------------------------------------------------------
         dis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .2637525   .1774928     1.49   0.137    -.0841271     .611632
      gender |    .287099   .1795689     1.60   0.110    -.0648495    .6390475
      genage |  -.1757638   .2371684    -0.74   0.459    -.6406054    .2890778
        expo |   .9497358   .2762329     3.44   0.001     .4083292    1.491142
          ae |  -.8722452   .3064885    -2.85   0.004    -1.472952   -.2715388
          ge |  -.2615727    .306344    -0.85   0.393    -.8619959    .3388505
       _cons |   -.643299   .1289406    -4.99   0.000     -.896018   -.3905801
------------------------------------------------------------------------------
```

Now, for this fit, $b_5 = -0.87224452$ is the estimate of the log of the ratio of 2 odds ratios [young to old] assumed common to gender. It would be telling us about age modification adjusted for gender except that such a statement was discredited by the previous model. The 'assumed common' part is not correct.

And $b_6 = -0.2615727$ has the same issue in play. It is an estimate of the log of the ratio of 2 odds ratios [male to female] assumed common to age group. Again, the 'assumed common' part is incorrect.

The two 'one-at-a-time' models can be compared to the two classic 'one-at-a-time' analyses. Similar problems here.

```
. logit dis gender expo ge [fw=ct1]

Logistic regression                               Number of obs     =      1,200
                                                  LR chi2(3)        =      10.12
                                                  Prob > chi2       =     0.0176
Log likelihood = -810.30733                       Pseudo R2         =     0.0062

------------------------------------------------------------------------------
         dis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |   .2034843   .1309226     1.55   0.120    -.0531193     .460088
        expo |   .4752451   .2153856     2.21   0.027     .0530972    .8973931
          ge |  -.1318425     .29799    -0.44   0.658    -.7158922    .4522073
       _cons |  -.5129855   .0929605    -5.52   0.000    -.6951847   -.3307863
------------------------------------------------------------------------------
```

```
. logit dis age expo ae [fw=ct1]

Logistic regression                               Number of obs     =      1,200
                                                  LR chi2(3)        =      16.09
                                                  Prob > chi2       =     0.0011
Log likelihood = -807.31932                       Pseudo R2         =     0.0099

------------------------------------------------------------------------------
         dis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .1833413   .1309588     1.40   0.162    -.0733332    .4400158
        expo |   .8273689   .2082089     3.97   0.000      .419287    1.235451
          ae |  -.8713591   .3017082    -2.89   0.004    -1.462696    -.280022
       _cons |  -.5052854    .094118    -5.37   0.000    -.6897533   -.3208176
------------------------------------------------------------------------------
```

It is crucial to understand why the first model [with all eight regression coefficients] is needed here. All the other models detail oversimplifications and components that can be discredited.