Models In Epidemiology And Biostatistics
Gordon Hilton Fick
Session 3 : A Detailed Annotated Example

```
. cc dis exp,exact
                                                      Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |       210          90  |        300       0.7000
        Controls |       120         180  |        300       0.4000
-----------------+------------------------+------------------------
           Total |       330         270  |        600       0.5500
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |              3.5       |    2.461396    4.980692 (exact)
 Attr. frac. ex. |         .7142857       |    .5937264    .7992247 (exact)
 Attr. frac. pop |               .5       |
                 +-------------------------------------------------
                                 1-sided Fisher's exact P = 0.0000
                                 2-sided Fisher's exact P = 0.0000
. disp r(p_exact)
1.830e-13
```

*** r(p_exact) provides the p-value with significant digits

```
. cc dis exp if smo==1,exact
                                                      Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |       197          48  |        245       0.8041
        Controls |        77          19  |         96       0.8021
-----------------+------------------------+------------------------
           Total |       274          67  |        341       0.8035
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         1.012716       |    .5270515    1.888145 (exact)
 Attr. frac. ex. |         .0125568       |   -.8973477    .4703796 (exact)
 Attr. frac. pop |         .0100967       |
                 +-------------------------------------------------
                                 1-sided Fisher's exact P = 0.5383
                                 2-sided Fisher's exact P = 1.0000

. cc dis exp if smo==2,exact
                                                      Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |        13          42  |         55       0.2364
        Controls |        43         161  |        204       0.2108
-----------------+------------------------+------------------------
           Total |        56         203  |        259       0.2162
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         1.158915       |    .5225342    2.444502 (exact)
 Attr. frac. ex. |         .1371237       |   -.9137503    .5909188 (exact)
 Attr. frac. pop |         .0324111       |
                 +-------------------------------------------------
                                 1-sided Fisher's exact P = 0.4039
                                 2-sided Fisher's exact P = 0.7131
```

```
. cc dis expo,by(smo)

              smo |       OR       [95% Conf. Interval]   M-H Weight
------------------+--------------------------------------------------
                1 |   1.012716      .5270515   1.888145     10.83871 (exact)
                2 |   1.158915      .5225342   2.444502     6.972973 (exact)
------------------+--------------------------------------------------
            Crude |        3.5      2.461396   4.980692              (exact)
    M-H combined |   1.069951       .679206   1.685489
-------------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     0.08  Pr>chi2 = 0.7746

                   Test that combined OR = 1:
                               Mantel-Haenszel chi2(1) =      0.08
                               Pr>chi2 =    0.7708
```

*** There is no evidence that smoking is a modifier from the test of homogeneity of odds ratios (p = 0.7746). Further, the smoking group specific estimated odds ratios are nearly the same as are the confidence intervals and group specific p-values. So here it makes sense to explore whether smoking is a confounder.
*** There is evidence that smoking is a confounder since the M-H "adjusted" OR estimate (1.069951) is different from the crude OR estimate (3.5). The confidence interval for the "assumed common" OR includes one while the confidence interval for the "crude" OR does not include one. These 2 intervals do not overlap. A test of significance for a difference between the "crude" OR and the "assumed common" OR is not warranted in most epidemiologic settings. We can test whether the assumed common OR is equal to 1 [p-value =0.7708] and hence there is no evidence against the hypothesis that there is no disease-exposure association.

```
. gen smc=2-smo
```

*** gen here generates an indicator variable smc for smoking group

```
. gen ds=dis*smc
```

*** gen here creates the ds variable needed for model 2

*** model 2 fit

```
. logit expo dis smc ds

Iteration 0:   log likelihood = -412.88329
Iteration 1:   log likelihood =  -304.1997
Iteration 2:   log likelihood =  -304.0952
Iteration 3:   log likelihood = -304.09518

Logistic regression                               Number of obs     =        600
                                                  LR chi2(3)        =     217.58
                                                  Prob > chi2       =     0.0000
Log likelihood = -304.09518                       Pseudo R2         =     0.2635

------------------------------------------------------------------------------
        expo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         dis |   .1474841   .3608318     0.41   0.683    -.5597332    .8547015
         smc |    2.71957   .3083595     8.82   0.000     2.115197    3.323944
          ds |  -.1348479   .4708796    -0.29   0.775    -1.057755    .7880592
       _cons |  -1.320204   .1716595    -7.69   0.000    -1.656651   -.9837577
------------------------------------------------------------------------------
```

*** same model fit but now showing the exponents of the coefficients

```
. logit expo dis smc ds,or

Iteration 0:   log likelihood = -412.88329
Iteration 1:   log likelihood =  -304.1997
Iteration 2:   log likelihood =  -304.0952
Iteration 3:   log likelihood = -304.09518

Logistic regression                               Number of obs     =        600
                                                  LR chi2(3)        =     217.58
                                                  Prob > chi2       =     0.0000
Log likelihood = -304.09518                       Pseudo R2         =     0.2635

------------------------------------------------------------------------------
        expo | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         dis |   1.158915   .4181733     0.41   0.683     .5713615    2.350673
         smc |    15.1738   4.678986     8.82   0.000     8.291219    27.76966
          ds |   .8738488   .4114776    -0.29   0.775     .3472345    2.199124
       _cons |   .2670807    .045847    -7.69   0.000     .1907769    .3739034
------------------------------------------------------------------------------
```

*** To acclimatize to the models and the fits, it is best to write them down:
Here p = Pr(exposure) and the model is:

$$\log(p/(1-p))=\beta_0+\beta_1 D+\beta_2 S+\beta_3 DS$$

The fit is:

$$\log(\hat{p}/(1-\hat{p}))=b_0+b_1 D+b_2 S+b_3 DS=-1.320204+0.1474841\,D+2.71957\,S-0.1348479\,DS$$

Notice that:

$$\exp(b_0)=\exp(-1.320204)=0.2670807 \quad \exp(b_1)=\exp(0.1474841)=1.158915$$ and so on...

*** So 1.158915 is an estimate (for the non-smokers) of the ratio of odds of exposure among those with disease to the odds of exposure among those without disease

*** notice that 0.873849 is actually an estimate of a ratio of two odds ratios. The label on the output is incorrect.

```
. est stor M2
```

*** used for a LR test

```
. logit expo dis smc

Logistic regression                               Number of obs     =        600
                                                  LR chi2(2)        =     217.49
                                                  Prob > chi2       =     0.0000
Log likelihood = -304.13608                       Pseudo R2         =     0.2634

------------------------------------------------------------------------------
        expo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         dis |   .0676887   .2319921     0.29   0.770    -.3870075    .5223848
         smc |    2.66253   .2331505    11.42   0.000     2.205563    3.119496
       _cons |  -1.302446   .1592554    -8.18   0.000     -1.61458   -.9903109
------------------------------------------------------------------------------
```

```
. logit expo dis smc,or

Iteration 0:   log likelihood = -412.88329
Iteration 1:   log likelihood = -304.24066
Iteration 2:   log likelihood =  -304.1361
Iteration 3:   log likelihood = -304.13608

Logistic regression                               Number of obs     =        600
                                                  LR chi2(2)        =     217.49
                                                  Prob > chi2       =     0.0000
Log likelihood = -304.13608                       Pseudo R2         =     0.2634

------------------------------------------------------------------------------
        expo | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         dis |   1.070032    .248239     0.29   0.770      .679086    1.686044
         smc |    14.3325   3.341629    11.42   0.000     9.075362    22.63497
       _cons |   .2718661   .0432961    -8.18   0.000     .1989741    .3714612
------------------------------------------------------------------------------
```

*** 1.070032 is the maximum likelihood estimate (MLE) of the assumed common OR. It is slightly different from the M-H estimate (1.069951). Confidence limits are almost the same (but not quite the same) *** In other words, under the assumption that the OR among smokers is the same as the OR among non-smokers, an estimate of this assumed common OR_{ac} is 1.070032

If $H_0: \mathrm{OR}_{ac} = 1$ then the p-value=0.770

```
. est stor M3

. logit expo dis

Iteration 0:   log likelihood = -412.88329
Iteration 1:   log likelihood = -385.21491
Iteration 2:   log likelihood = -385.16279
Iteration 3:   log likelihood = -385.16279

Logistic regression                               Number of obs     =        600
                                                  LR chi2(1)        =      55.44
                                                  Prob > chi2       =     0.0000
Log likelihood = -385.16279                       Pseudo R2         =     0.0671

------------------------------------------------------------------------------
        expo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         dis |   1.252763   .1725164     7.26   0.000      .914637    1.590889
       _cons |  -.4054651   .1178511    -3.44   0.001    -.6364491   -.1744811
------------------------------------------------------------------------------

. logit expo dis,or

Iteration 0:   log likelihood = -412.88329
Iteration 1:   log likelihood = -385.21491
Iteration 2:   log likelihood = -385.16279
Iteration 3:   log likelihood = -385.16279
Logistic regression                               Number of obs     =        600
                                                  LR chi2(1)        =      55.44
                                                  Prob > chi2       =     0.0000
Log likelihood = -385.16279                       Pseudo R2         =     0.0671

------------------------------------------------------------------------------
        expo | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         dis |        3.5   .6038074     7.26   0.000     2.495869     4.90811
       _cons |   .6666667   .0785674    -3.44   0.001     .5291681    .8398927
```

```
------------------------------------------------------------------------------
```

*** We see that 3.5 is an estimate of the crude OR. This number is the same as that number computed in the first table [cc dis exp, exact]

```
. est stor M1

. lrtest M2 M3

Likelihood-ratio test                                 LR chi2(1)  =      0.08
(Assumption: M3 nested in M2)                         Prob > chi2 =    0.7749
```

*** For this example, the Likelihood Ratio (LR) test gives almost the same p-value as the Wald test. In principle, the LR test has better sampling properties than the Wald test.

```
. lrtest M1 M2

Likelihood-ratio test                                 LR chi2(2)  =    162.14
(Assumption: M1 nested in M2)                         Prob > chi2 =    0.0000
```

*** This test is not needed here

```
. disp 1/(1+exp(1.320204)) .21078436
```

*** This is the proportion exposed among the non-smoking controls (or the estimated probability of exposure among the non-smoking controls

```
. disp log(.21078/(1-.21078)) -1.3202302
```

*** This is the estimated log of odds of exposure among the non-smoking controls

*** Lets now consider modeling the log odds of disease

*** Interpret all the regression coefficients and their exponents.

*** When are the interpretations the same as modeling the log odds of exposure? When are the interpretations different? Why?

```
. gen es = expo * smc

. logit dis expo smc es

Iteration 0:   log likelihood = -415.88831
Iteration 1:   log likelihood = -336.60557
Iteration 2:   log likelihood = -336.52171
Iteration 3:   log likelihood =  -336.5217

Logistic regression                               Number of obs     =        600
                                                  LR chi2(3)        =     158.73
                                                  Prob > chi2       =     0.0000
Log likelihood =  -336.5217                       Pseudo R2         =     0.1908

------------------------------------------------------------------------------
         dis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        expo |    .147484   .3608318     0.41   0.683    -.5597334    .8547013
         smc |   2.270497   .3216918     7.06   0.000     1.639992    2.901001
          es |  -.1348477   .4708797    -0.29   0.775    -1.057755    .7880595
       _cons |  -1.343735   .1732648    -7.76   0.000    -1.683328   -1.004142
------------------------------------------------------------------------------
```

```
. logit dis expo smc es, or

Iteration 0:   log likelihood = -415.88831
Iteration 1:   log likelihood = -336.60557
Iteration 2:   log likelihood = -336.52171
Iteration 3:   log likelihood =  -336.5217

Logistic regression                             Number of obs     =        600
                                                LR chi2(3)        =     158.73
                                                Prob > chi2       =     0.0000
Log likelihood =  -336.5217                     Pseudo R2         =     0.1908

------------------------------------------------------------------------------
         dis | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        expo |   1.158915   .4181733     0.41   0.683     .5713614    2.350672
         smc |    9.68421   3.115331     7.06   0.000      5.15513    18.19235
          es |    .873849   .4114777    -0.29   0.775     .3472345    2.199125
       _cons |   .2608696   .0451995    -7.76   0.000     .1857548    .3663589
------------------------------------------------------------------------------

. logit dis expo smc

Iteration 0:   log likelihood = -415.88831
Iteration 1:   log likelihood = -336.64237
Iteration 2:   log likelihood = -336.56261
Iteration 3:   log likelihood =  -336.5626

Logistic regression                             Number of obs     =        600
                                                LR chi2(2)        =     158.65
                                                Prob > chi2       =     0.0000
Log likelihood =  -336.5626                     Pseudo R2         =     0.1907

------------------------------------------------------------------------------
         dis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        expo |   .0676888   .2319921     0.29   0.770    -.3870074    .5223849
         smc |   2.208323   .2354759     9.38   0.000     1.746799    2.669847
       _cons |  -1.325646   .1604954    -8.26   0.000    -1.640211    -1.01108
------------------------------------------------------------------------------

. logit dis expo smc, or

Iteration 0:   log likelihood = -415.88831
Iteration 1:   log likelihood = -336.64237
Iteration 2:   log likelihood = -336.56261
Iteration 3:   log likelihood =  -336.5626

Logistic regression                             Number of obs     =        600
                                                LR chi2(2)        =     158.65
                                                Prob > chi2       =     0.0000
Log likelihood =  -336.5626                     Pseudo R2         =     0.1907

------------------------------------------------------------------------------
         dis | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        expo |   1.070032    .248239     0.29   0.770     .6790861    1.686044
         smc |   9.100444   2.142935     9.38   0.000     5.736212    14.43777
       _cons |   .2656314   .0426326    -8.26   0.000     .1939392    .3638258
------------------------------------------------------------------------------
```

```
. logit dis expo, or

Iteration 0:   log likelihood = -415.88831
Iteration 1:   log likelihood = -388.19658
Iteration 2:   log likelihood = -388.16781
Iteration 3:   log likelihood = -388.16781

Logistic regression                             Number of obs     =        600
                                                LR chi2(1)        =      55.44
                                                Prob > chi2       =     0.0000
Log likelihood = -388.16781                     Pseudo R2         =     0.0667

------------------------------------------------------------------------------
         dis | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        expo |        3.5   .6038074     7.26   0.000     2.495869     4.90811
       _cons |         .5   .0645497    -5.37   0.000     .3882222    .6439611
------------------------------------------------------------------------------
```