Lab #10 Discussion

Question 1. *Assess the fitted model based on the confidence intervals for the coefficients and the results of the likelihood ratio test.*

```
. logistic chd69 chol sbp dibpat age smoke bmi, coef
Logistic regression                               Number of obs   =       3140
                                                  LR chi2(6)      =     184.43
                                                  Prob > chi2     =     0.0000
Log likelihood = -794.79384                       Pseudo R2       =     0.1040

------------------------------------------------------------------------------
       chd69 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        chol |   .0106234   .0015272     6.96   0.000     .0076301    .0136167
         sbp |   .0182514   .0041456     4.40   0.000     .0101262    .0263766
      dibpat |   .6972925   .1443713     4.83   0.000       .41433     .980255
         age |   .0603351   .0119705     5.04   0.000     .0368734    .0837968
       smoke |   .6038674   .1410698     4.28   0.000     .3273756    .8803592
         bmi |   .0535404   .0267973     2.00   0.046     .0010187    .1060621
       _cons |  -12.25074   .9832063   -12.46   0.000    -14.17779    -10.3237
------------------------------------------------------------------------------
```

The default likelihood ratio test statistic `LR chi2(6)`=184.43, and the associated *P*-value is very, very close to zero. This statistic compares the likelihood of this model with the corresponding (nested) model that excludes all of the predictors. The null hypothesis evaluated is that the model including predictors does not improve on the corresponding model including only an intercept term.

The output of the model with no predictors is included below:

```
. logistic chd69, coef
Logistic regression                               Number of obs   =       3140
                                                  LR chi2(0)      =       0.00
                                                  Prob > chi2     =          .
Log likelihood = -887.00953                       Pseudo R2       =     0.0000

------------------------------------------------------------------------------
       chd69 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -2.421756    .065215   -37.13   0.000    -2.549575   -2.293937
------------------------------------------------------------------------------
```

With this simplistic model, we are trying to predict CHD outcomes using a constant (`-2.421756`). This is just the log-odds of CHD in the sample. To verify this, check out the following commands:

```
. summarize chd69
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       chd69 |      3140    .0815287    .2736889          0          1
. dis log(r(mean)/(1 - r(mean)))
-2.4217561
```

The first returns the average of the binary indicator chd69 (i.e. the proportion of individuals with CHD in the sample); the second computes the log odds of this quantity (returned via the `r(mean)` function). Thus, we are comparing a model including six

fairly established predictors of CHD with a simpler model that specifies that our best guess for each individual's CHD risk is the overall prevalence of CHD in the entire sample! The Wald test for the coefficient in this model (`z = -37.13, P=0.000`) is testing the hypothesis that the true intercept coefficient is zero. This compares the model that bases the estimated log odds for the outcome on the observed disease prevalence with the null model that specifies that the best guess for the log odds is zero (i.e. that individuals are as likely to have disease as not). This hypothesis is typically not of much interest.

We can verify the value of the LR statistic for the model including predictors by subtracting the value of the of the log-likelihood from the model with no predictors from the corresponding value from the larger model, and multiplying the difference by 2:

```
. dis 2*(-794.79384- (-887.00953))
184.43138
```

The value required for significance with 6 degrees of freedom is 12.592, hence the very small *P*-value.

Question 2. *Examine the differences between the model excluding the most influential observation and the model including all observations. What issues should you consider in deciding whether to retain this observation in further analyses?*

```
. list id chd69 chol sbp dibpat age smoke bmi if id==10078

      +-----------------------------------------------------------------+
      |    id   chd69   chol   sbp   dibpat   age   smoke        bmi |
      |-----------------------------------------------------------------|
 501. | 10078     yes    188   166    B3,B4    43       0   38.94737 |
      +-----------------------------------------------------------------+

. logistic chd69 chol sbp dibpat age smoke bmi if id != 10078, coef

Logistic regression                               Number of obs   =       3139
                                                  LR chi2(6)      =     184.79
                                                  Prob > chi2     =     0.0000
Log likelihood = -792.10558                       Pseudo R2       =     0.1045

------------------------------------------------------------------------------
       chd69 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        chol |   .0107656   .0015299     7.04   0.000      .007767    .0137642
         sbp |   .0180712   .0041567     4.35   0.000     .0099241    .0262183
      dibpat |   .7113087   .1449475     4.91   0.000     .4272168    .9954007
         age |   .0610197   .0119898     5.09   0.000     .0375202    .0845192
       smoke |   .6065078   .1414284     4.29   0.000     .3293132    .8837024
         bmi |   .0453055   .0270511     1.67   0.094    -.0077137    .0983247
       _cons |  -12.10452   .9856004   -12.28   0.000    -14.03626   -10.17278
------------------------------------------------------------------------------
```

Comparing the above coefficients to the output from the model including the observation in question, we see that they are quite comparable. We can also compare the values of the predictors in question in the entire sample to the values for the individual being considered to look for possible explanations for the relatively high influence.

```
. list bmi weight height chol if id==10078

      +-----------------------------------+
      |      bmi   weight   height   chol |
      |-----------------------------------|
3140. | 38.94737      320       76    188 |
```

We see that this person has CHD, less than average cholesterol and a higher than average BMI. These differences might be enough to explain the results but do not seem particularly alarming. Given that the observed change in the coefficients is relatively small, and that we have no reason to suspect that any of the predictors reflect errors in measurement (although we may want to check this in practice), keeping the observation seems warranted.

Question 3. *Interpret the results of the goodness of fit test.*

```
. lfit, group(10) table
Logistic model for chd69, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.0191 |     2 |   4.5 |   312 | 309.5 |   314 |
  |     2 | 0.0273 |     9 |   7.3 |   305 | 306.7 |   314 |
  |     3 | 0.0367 |     8 |  10.0 |   306 | 304.0 |   314 |
  |     4 | 0.0464 |    10 |  13.0 |   304 | 301.0 |   314 |
  |     5 | 0.0587 |    20 |  16.4 |   294 | 297.6 |   314 |
  |-------+--------+-------+-------+-------+-------+-------|
  |     6 | 0.0729 |    15 |  20.5 |   299 | 293.5 |   314 |
  |     7 | 0.0941 |    24 |  26.0 |   290 | 288.0 |   314 |
  |     8 | 0.1220 |    47 |  33.5 |   267 | 280.5 |   314 |
  |     9 | 0.1719 |    48 |  45.2 |   266 | 268.8 |   314 |
  |    10 | 0.5528 |    73 |  79.5 |   241 | 234.5 |   314 |
  +--------------------------------------------------------+
       number of observations =      3140
             number of groups =        10
     Hosmer-Lemeshow chi2(8) =        12.47
                 Prob > chi2 =         0.1314
```

Recall that the hypothesis being evaluated in the goodness of fit (GOF) test is that the model predictions agree with the observed counts over the (ten) ordered outcome categories specified. Rejection indicates significant lack of agreement. In this case, we do not reject at the 5% level. Although this indicates that the model fit may not be unacceptably poor, we are certainly not justified to conclude that we have a "good" fitting model. As indicated in lecture, before concluding anything final about fit, further evaluation of model assumptions and application of diagnostic procedures are necessary. Particular next steps that might be taken include exploration of nonlinearity of risk with included predictors, evaluation of potential interactions and graphical examination of influence and residuals. Assuming that our model escapes unscathed after these evaluations, presenting the result of the GOF test (along with a description of the other efforts) as indicating that the model fit is *acceptable* is justified. Trying the goodness of fit test again after including the suggested interaction appears to improve the fit considerably.

Question 4. *Comment on the differences between the sensitivity estimates obtained in the last two commands and their implications for expected classification performance*.

The output below gives the results from the `lstat` command applied to both the development and validation groups:

```
. lstat if group==1, cutoff(0.085)

Logistic model for chd69
             -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |       157           758  |        915
     -     |        65          1846  |       1911
-----------+--------------------------+-----------
   Total   |       222          2604  |       2826

Classified + if predicted Pr(D) >= .085
True D defined as chd69 != 0
-------------------------------------------------
Sensitivity                     Pr( +| D)   70.72%
Specificity                     Pr( -|~D)   70.89%
Positive predictive value       Pr( D| +)   17.16%
Negative predictive value       Pr(~D| -)   96.60%
-------------------------------------------------
False + rate for true ~D        Pr( +|~D)   29.11%
False - rate for true D         Pr( -| D)   29.28%
False + rate for classified +   Pr(~D| +)   82.84%
False - rate for classified -   Pr( D| -)    3.40%
-------------------------------------------------
Correctly classified                        70.88%
-------------------------------------------------

. lstat if group==2, cutoff(0.085)

Logistic model for chd69
             -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        24            90  |        114
     -     |        10           190  |        200
-----------+--------------------------+-----------
   Total   |        34           280  |        314

Classified + if predicted Pr(D) >= .085
True D defined as chd69 != 0
-------------------------------------------------
Sensitivity                     Pr( +| D)   70.59%
Specificity                     Pr( -|~D)   67.86%
Positive predictive value       Pr( D| +)   21.05%
Negative predictive value       Pr(~D| -)   95.00%
-------------------------------------------------
False + rate for true ~D        Pr( +|~D)   32.14%
False - rate for true D         Pr( -| D)   29.41%
False + rate for classified +   Pr(~D| +)   78.95%
False - rate for classified -   Pr( D| -)    5.00%
-------------------------------------------------
Correctly classified                        68.15%
-------------------------------------------------
```

The validation sample results slightly inferior overall classification performance characterized by reduced specificity. As discussed in lecture, cross validation and

bootstrapping methods would yield more reliable estimates of optimism-corrected values for these quantities.

Question 4. *Compare the estimated relative risks with the odds ratios estimated from the logistic model. Are there any notable differences?*

```
. poisson chd69 dibpat age smoke c.bmic##c.cholc c.sbpc##c.bmic, irr robust
note: bmic omitted because of collinearity

Iteration 0:   log pseudolikelihood = -809.03452
Iteration 1:   log pseudolikelihood = -809.00598
Iteration 2:   log pseudolikelihood = -809.00597

Poisson regression                                Number of obs   =       3140
                                                  Wald chi2(8)    =     259.34
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -809.00597                 Pseudo R2       =     0.0988

------------------------------------------------------------------------------
             |               Robust
       chd69 |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      dibpat |   1.885479   .2429655     4.92   0.000     1.464653    2.427218
         age |   1.052683   .0108212     4.99   0.000     1.031686    1.074107
       smoke |   1.696515   .2062327     4.35   0.000     1.336852     2.15294
        bmic |   1.101012   .0273448     3.87   0.000       1.0487    1.155932
       cholc |   1.009803   .0011662     8.45   0.000      1.00752    1.012091
             |
c.bmic#c.cholc |   .9985268   .0003806    -3.87   0.000     .9977812    .9992729
             |
        sbpc |   1.017992   .0033253     5.46   0.000     1.011496    1.024531
        bmic |          1  (omitted)
             |
 c.sbpc#c.bmic |   .9969651   .0011988    -2.53   0.011     .9946184    .9993174
             |
       _cons |    .003102    .001585   -11.30   0.000     .0011395    .0084447
------------------------------------------------------------------------------
```

No big differences are apparent. This is a case where deciding which results to present might be based on investigator preference and target audience. As always, we recommend reporting what was actually estimated (i.e. always report odds ratios as such).