Soc 63993, Homework #9:
Logistic Regression

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/
Last revised March 28, 2015

I. As we saw in the class handout on the PSI teaching example, 8 of the 14 students who were in PSI got A's compared to only 3 of the 18 students who were in a conventional classroom. Verify that those numbers are consistent with the following results that we get when GRADE is (logistically) regressed on PSI only. Recall that GRADE = 1 if grade is an A, 0 otherwise, PSI = 1 if in psi, 0 otherwise. [HINT: Compute the log odds for those in psi and those not in psi, and then take it from there.]

```
. use https://www3.nd.edu/~rwilliam/statafiles/logist.dta, clear
. logit grade i.psi, nolog

Logistic regression                                     Number of obs   =         32
                                                        LR chi2(1)      =       5.84
                                                        Prob > chi2     =     0.0156
Log likelihood = -17.670815                             Pseudo R2       =     0.1418

------------------------------------------------------------------------------
       grade |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.psi |    1.89712    .831665     2.28   0.023     .2670865    3.527153
       _cons |  -1.609438   .6324555    -2.54   0.011    -2.849028   -.3698478
------------------------------------------------------------------------------
```

II. Download *lrb.dta* from the course web page. We use a sample of Southern Baptists from the GSS in this homework. General Social Surveys from 1973 to 1991 are used to make one big sample. All married Southern Baptists between the ages of 20 to 25 (all 61 of them!) are in the data file. The dependent variable is `happymar`, respondent's marital happiness (1 = Very Happy, 0 = Otherwise). `church`, Church attendance (1 = Often attends, 0 = other), `female` (1 = female, 0 = male), and `educ`, Years of education, are the DVs.

Use Stata to run the logistic regression of happymar on church, female and educ. Then answer the following questions.

1. What assumptions of OLS would be violated if OLS was used to approach this problem?

2. Interpret the logistic regression coefficients. What do the parameters tell you about the determinants of marital happiness? What can you say about the size and magnitude of effects?

3. Determine the log odds, odds and probability of marital happiness for:
   (a) a male with 8 years of education who is not a regular churchgoer
   (b) a male with 8 years of education who is a regular churchgoer
   (c) a female with 16 years of education who is not a regular churchgoer
   (d) a female with 16 years of education who is a regular churchgoer.

   That is, complete the following table using the values above.

Church	Female	Educ	Log odds	Odds	P(Happy)

Do this first by hand. Then confirm your answers by using the `adjust` and/or `margins` commands.

4. What are the values of DEV_0, DEV_M, and G_M? Explain what each of these parameters means and, in the case of G_M, what hypothesis it is testing and whether or not you should reject that hypothesis given the results. Also, what does McFadden's Pseudo R^2 equal? (Note that some of these values are explicitly reported in the printout while others require minor computations.)

5. Run the following post-estimation commands and `extremes` command (you need to have the `extremes` command installed):

```
estat class
predict phappy
predict rstandard, rstandard
extremes rstandard  happymar phappy church female educ
```

What is the proportion of cases that have been correctly classified? Of the cases that have been improperly classified, which ones appear to be the most problematic?

6. The data set also includes a variable, `educx`, which is equal to education centered about its mean. Rerun the logistic regression using `educx`. Note that the value of the intercept (but no other coefficient) changes when you do this. Explain how to interpret the intercept once education is centered, and how that differs from the interpretation when education is not centered. Review your earlier notes on centering if necessary.

7. The data set also includes the interaction `cheducx` = `church` * `educx`. Add it to the model (or, if you prefer, add it via factor variable notation) and use a likelihood ratio chi-square test (i.e. don't just rely on the Wald statistic) to test whether the effect of `cheducx` is significant. What is the value of the test statistic and what does it tell you?