Soc 63993, Homework #9 Answer Key:
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
------------------------------------------------------------------------------
```

Note that, in the sample, 8/14, or 57.14%, of those in Psi got As, compared to 3/18, or 16.67% of those not in Psi. We should be able to reproduce those numbers from the model's parameters.

Using log odds, note that

$$LogOdds = \alpha + \sum_{k=1}^{K} \beta_k X_k = \alpha + \beta * Psi = -1.609 + 1.897 * Psi$$

For those not in Psi, this simplifies to

$$LogOdds = \alpha + \beta * Psi = -1.609 + 1.897 * 0 = -1.609$$

Hence, the probability of someone who is not in Psi getting an A is

$$P = \frac{1}{1+\exp(-Z)} = \frac{1}{1+\exp(1.609)} = \frac{1}{1+5} = \frac{1}{6} = 16.67\%$$

which is what we found in the sample. Similarly, for those in Psi, we get

$$LogOdds = \alpha + \beta * Psi = -1.609 + 1.897 * 1 = .288$$

$$P == \frac{1}{1+\exp(-Z)} = \frac{1}{1+\exp(-.288)} = \frac{1}{1+.75} = \frac{1}{1.75} = 57.14\%$$

which again is what we found in the sample.

To confirm using the `margins` command,

```
. margins psi

Adjusted predictions                              Number of obs   =         32
Model VCE    : OIM

Expression   : Pr(grade), predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         psi |
          0  |   .1666667    .087841     1.90   0.058    -.0054986     .338832
          1  |   .5714286     .13226     4.32   0.000     .3122037    .8306534
------------------------------------------------------------------------------
```

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.

Here is the output:

```
. use https://www3.nd.edu/~rwilliam/statafiles/lrb.dta, clear
. logit happymar i.church i.female educ

Iteration 0:   log likelihood = -39.881468
Iteration 1:   log likelihood = -25.667639
Iteration 2:   log likelihood = -24.652305
Iteration 3:   log likelihood = -24.633826
Iteration 4:   log likelihood = -24.633783
Iteration 5:   log likelihood = -24.633783

Logistic regression                               Number of obs   =         61
                                                  LR chi2(3)      =      30.50
                                                  Prob > chi2     =     0.0000
Log likelihood = -24.633783                       Pseudo R2       =     0.3823

------------------------------------------------------------------------------
    happymar |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.church |   2.907538   .9207651     3.16   0.002     1.102871    4.712204
    1.female |     2.3945   .8773269     2.73   0.006      .674971    4.114029
        educ |   .5266878   .2651831     1.99   0.047     .0069384    1.046437
       _cons |   -8.15857   3.285418    -2.48   0.013    -14.59787   -1.719269
------------------------------------------------------------------------------
```

```
. logit, or

Logistic regression                               Number of obs   =         61
                                                  LR chi2(3)      =      30.50
                                                  Prob > chi2     =     0.0000
Log likelihood = -24.633783                       Pseudo R2       =     0.3823

------------------------------------------------------------------------------
    happymar | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.church |   18.31166   16.86073     3.16   0.002     3.012805    111.2972
    1.female |   10.96272   9.617886     2.73   0.006     1.963976    61.19278
        educ |   1.693314   .4490384     1.99   0.047     1.006963    2.847488
       _cons |   .0002863   .0009405    -2.48   0.013     4.57e-07    .1791971
------------------------------------------------------------------------------
```

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

- Errors would not be homoscedastic
- It is extremely unlikely (albeit not impossible) that the X's would have linear and additive effects on Y (or, more specifically, on the P[Y = 1]). For example, an OLS model would say that someone can go from a 50% chance of happiness to 60%, and someone else can go from 98% to 108%, which is not possible. A logistic regression model, which says that X's have a linear effect on the log odds of an event occurring, is generally more plausible.

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?

All three IVs have statistically significant effects. According to these results, those who attend church often (i.e. are coded 1 on CHURCH), and women (coded 1 on female) are more likely to say their marriages are very happy (in each case, a score of one on the variable increases the odds of happiness more than 10 fold). Better educated individuals also are more likely to say their marriages are happy. The answers to part 3 will give us a better feel for what these numbers mean in practice.

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.

Let's construct the following table:

Church	Female	Educ	Log odds	Odds	P(Happy)
0	0	8	-3.945	0.0194	1.90%
1	0	8	-1.0375	0.3543	26.16%
0	1	16	2.6631	14.3407	93.48%
1	1	16	5.5706	262.5916	99.62%

To get the last 3 columns:

Log odds = -8.1586 + (2.9075 * Church) + (2.3945 * Female) + (0.5267 * Educ)
Odds = Exp(Log odds)
P(Happy) = Odds/(1 + Odds)

Confirming the results with Stata's `margins` command,

```
. * Log odds
. margins church, at(educ = 8 female = 0) predict(xb)

Adjusted predictions                              Number of obs   =         61
Model VCE    : OIM

Expression   : Linear prediction (log odds), predict(xb)
at           : female          =           0
               educ            =           8

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      church |
          0  |  -3.945068   1.324316    -2.98   0.003    -6.540681   -1.349456
          1  |   -1.03753   1.180121    -0.88   0.379    -3.350525    1.275464
------------------------------------------------------------------------------

. * Odds
. margins church, at(educ = 8 female = 0) expression(exp(predict(xb)))

Adjusted predictions                              Number of obs   =         61
Model VCE    : OIM

Expression   : exp(predict(xb))
at           : female          =           0
               educ            =           8

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      church |
          0  |   .0193499   .0256254     0.76   0.450    -.0308749    .0695747
          1  |   .3543287   .4181507     0.85   0.397    -.4652316    1.173889
------------------------------------------------------------------------------
```

```
. * Predicted probabilities
. margins church, at(educ = 8 female = 0)

Adjusted predictions                              Number of obs   =         61
Model VCE    : OIM

Expression   : Pr(happymar), predict()
at           : female          =           0
               educ            =           8

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      church |
          0  |   .0189826   .0246617     0.77   0.441    -.0293536    .0673187
          1  |   .2616268   .2279737     1.15   0.251    -.1851934    .7084469
------------------------------------------------------------------------------
```

Confirming the results with Stata's older `adjust` command,

```
. quietly logit happymar church female educ
. adjust educ = 8 female = 0, by(church) xb

--------------------------------------------------------------------------------------
     Dependent variable: happymar     Command: logit
Covariates set to value: educ = 8, female = 0
--------------------------------------------------------------------------------------

-----------------------------
Church           |
attendance       |         xb
-----------------+-----------
rarely in church |   -3.94507
 often in church |   -1.03753
-----------------------------
     Key:  xb  =  Linear Prediction

. adjust educ = 8 female = 0, by(church) exp

--------------------------------------------------------------------------------------
     Dependent variable: happymar     Command: logit
Covariates set to value: educ = 8, female = 0
--------------------------------------------------------------------------------------

-----------------------------
Church           |
attendance       |    exp(xb)
-----------------+-----------
rarely in church |     .01935
 often in church |    .354329
-----------------------------
     Key:  exp(xb)  =  exp(xb)
```

```
. adjust educ = 8 female = 0, by(church) p

-------------------------------------------------------------------------------------
     Dependent variable: happymar     Command: logit
Covariates set to value: educ = 8, female = 0
-------------------------------------------------------------------------------------

-----------------------------
Church           |
attendance       |         pr
-----------------+-----------
rarely in church |    .018983
 often in church |    .261627
-----------------------------
     Key:  pr  =  Probability

. adjust educ = 16 female = 1, by(church) xb

-------------------------------------------------------------------------------------
     Dependent variable: happymar     Command: logit
Covariates set to value: educ = 16, female = 1
-------------------------------------------------------------------------------------

-----------------------------
Church           |
attendance       |         xb
-----------------+-----------
rarely in church |    2.66293
 often in church |    5.57047
-----------------------------
     Key:  xb  =  Linear Prediction

. adjust educ = 16 female = 1, by(church) exp

-------------------------------------------------------------------------------------
     Dependent variable: happymar     Command: logit
Covariates set to value: educ = 16, female = 1
-------------------------------------------------------------------------------------

-----------------------------
Church           |
attendance       |    exp(xb)
-----------------+-----------
rarely in church |    14.3383
 often in church |    262.558
-----------------------------
     Key:  exp(xb)  =  exp(xb)

. adjust educ = 16 female = 1, by(church) p

-------------------------------------------------------------------------------------
     Dependent variable: happymar     Command: logit
Covariates set to value: educ = 16, female = 1
-------------------------------------------------------------------------------------

-----------------------------
Church           |
attendance       |         pr
-----------------+-----------
rarely in church |    .934804
 often in church |    .996206
-----------------------------
     Key:  pr  =  Probability
```

These numbers match up with what we got before.

According to the model, a male with 8 years of education who did not attend church regularly would have only a 1.9% chance of having (or claiming to have) a very happy marriage. If that same male attended church regularly, the chance for happiness would jump to over 26%.

For a female with 16 years of education who doesn't attend church, the probability of a happy marriage is more than 93%. If such a female attends church, her probability of a happy marriage is extremely high, 99.62%.

For both the poorly educated man and the well-educated woman in our example, attending church increases the odds of happiness by a factor of 18. In terms of percentages, the increase is much greater for the man. This is because the well-educated woman already has a very high probability of happiness, and hence can't improve her chances much more, whereas the poorly educated man has room to improve his chances considerably.

If we really thought this model was correct, and we were determined to increase marital happiness, we might recommend that

- People start attending church more
- People should try to get more education
- Men should have sex-change operations (however, this might have its own direct effect on marital happiness)

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.)

From the Stata printout, we can tell that

G_M = LR chi2(3) = 30.50

DEV_M = -2LLM = -2 * -24.633783 = 49.268

$DEV_0 = G_M + DEV_M$ = 30.50 + 49.268 = 79.76; or equivalently, DEV_0 = -2LL0 = -2*-39.88 = 79.76.

McFadden's $R^2 = G_M/DEV_0$ = 30.50/79.763 = .382. (Or, if you prefer, just read it off the printout since Stata already reports it!)

DEV_0 is analogous to the total sum of squares in OLS regression; it is the variability that you are trying to explain. DEV_M is like the error sums of squares in OLS; it is the variation that is still unexplained after taking the model's variables into account. G_M is like the regression sums of squares; it tells you how much of the variability the model's variables account for. It is also like the global F test in OLS regression: G_M has a chi-square distribution and if it is significant, it tells you that one or more of the variables in the model has a non-zero effect.

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?

```
. estat class

Logistic model for happymar

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        36             9  |         45
     -     |         3            13  |         16
-----------+--------------------------+-----------
   Total   |        39            22  |         61

Classified + if predicted Pr(D) >= .5
True D defined as happymar != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   92.31%
Specificity                     Pr( -|~D)   59.09%
Positive predictive value       Pr( D| +)   80.00%
Negative predictive value       Pr(~D| -)   81.25%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   40.91%
False - rate for true D         Pr( -| D)    7.69%
False + rate for classified +   Pr(~D| +)   20.00%
False - rate for classified -   Pr( D| -)   18.75%
--------------------------------------------------
Correctly classified                        80.33%
--------------------------------------------------

. predict phappy
(option pr assumed; Pr(happymar))
. predict rstandard, rstandard
. extremes rstandard  happymar phappy church female educ

  +-----------------------------------------------------------------+
  | obs:   rstandard   happymar     phappy     church   female   educ |
  |-----------------------------------------------------------------|
  |  41.   -1.381519   not as h   .6324015   often in     male     11 |
  |   9.   -1.297267   not as h   .1372505   rarely i     male     12 |
  |  16.   -1.297267   not as h   .1372505   rarely i     male     12 |
  |  19.   -1.297267   not as h   .1372505   rarely i     male     12 |
  |  20.   -1.297267   not as h   .1372505   rarely i     male     12 |
  +-----------------------------------------------------------------+

  +----------------------------------------------------------------+
  |  43.   1.077104   very hap   .5039612   often in     male     10 |
  |   2.   1.355046   very hap   .3782007   rarely i   female     10 |
  |   6.   3.920076   not as h   .0858805   rarely i     male     11 |
  |  13.   3.920076   very hap   .0858805   rarely i     male     11 |
  |  36.   3.920076   very hap   .0858805   rarely i     male     11 |
  +----------------------------------------------------------------+

note: 6 values of -1.297267
```

Of the 22 cases that were "Not as happy", the model correctly classified 13, or 59%. For the 39 who were very happy, the model got 36 right. Overall, 80% of the cases were correctly classified.

Still, 12 cases were misclassified (the off-diagonal elements in the classification table). An examination of residuals can give us a better feel for those cases. The largest errors were for cases 13 and 36, so we might want to look more closely at them sometime. (Both of them happen to be non-churchgoing males with 11 years of education who said they had happy marriages.) With 61 cases, though, we would expect about 3 to have standardized residuals of 2 or above in magnitude, so actually we are doing pretty good. (Case 6 also has a large standard residual, even though the case is correctly classified. This is because it has the same values on the independent variables as the two cases with large standardized residuals. I don't fully understand the logic behind giving it the same standardized residual, but Stata says that this is the right way to do it!)

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.

Some descriptive statistics will also help:

```
. sum

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    happymar |        61    .6393443    .4841758          0          1
      church |        61    .4918033    .5040817          0          1
      female |        61    .5737705     .498632          0          1
        educ |        61    12.27869    1.817793          7         17
      cheduc |        61    6.278689     6.55523          0         16
-------------+--------------------------------------------------------
       educx |        61    9.38e-08    1.817793  -5.278688   4.721312
     cheducx |        61    .2399893    1.271463  -3.278688   3.721312

. logit happymar church female educ, nolog

Logistic regression                               Number of obs   =         61
                                                  LR chi2(3)      =      30.50
                                                  Prob > chi2     =     0.0000
Log likelihood = -24.633783                       Pseudo R2       =     0.3823

------------------------------------------------------------------------------
    happymar |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      church |   2.907538   .9207651     3.16   0.002     1.102871    4.712204
      female |     2.3945   .8773269     2.73   0.006      .674971    4.114029
        educ |   .5266878   .2651831     1.99   0.047     .0069384    1.046437
       _cons |   -8.15857   3.285418    -2.48   0.013    -14.59787   -1.719269
------------------------------------------------------------------------------

. di exp(-8.15857)
.00028627

. di exp(-8.15857)/ (1 + exp(-8.15857))
.00028619
```

```
. logit  happymar church female educx, nolog

Logistic regression                               Number of obs   =         61
                                                  LR chi2(3)      =      30.50
                                                  Prob > chi2     =     0.0000
Log likelihood = -24.633783                       Pseudo R2       =     0.3823

------------------------------------------------------------------------------
    happymar |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      church |   2.907538   .9207651     3.16   0.002     1.102871    4.712204
      female |     2.3945   .8773269     2.73   0.006      .674971    4.114029
       educx |   .5266878   .2651831     1.99   0.047     .0069384    1.046437
       _cons |  -1.691535   .7831351    -2.16   0.031    -3.226452   -.1566188
------------------------------------------------------------------------------
. di exp(-1.691535)
.1842365

. di exp(-1.691535)/ (1 + exp(-1.691535))
.15557408
```

In the original model, the intercept (-8.159) reflects the predicted log odds of a person who scores zero on every variable, i.e. a male who does not go to church very much and who has 0 years of education. This person has virtually no chance at being very happy; but luckily, he doesn't exist, at least in this sample, because descriptive statistics show us everyone has at least 7 years of education. In the model with centered education, the constant is the predicted log odds for a male who doesn't go to church very much who has an average level of education, 12.28 years. The odds of this person being very happy are .184 (exp(-1.691535)), which implies a 15.56% chance of being very happy.

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?

We will use the `nestreg` command to make this easy (remember to use the `lr` option):

```
. nestreg, lr: logit  happymar (church female educx)  cheducx, nolog

Block  1: church female educx

Logistic regression                               Number of obs   =         61
                                                  LR chi2(3)      =      30.50
                                                  Prob > chi2     =     0.0000
Log likelihood = -24.633783                       Pseudo R2       =     0.3823

------------------------------------------------------------------------------
    happymar |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      church |   2.907538   .9207651     3.16   0.002     1.102871    4.712204
      female |     2.3945   .8773269     2.73   0.006      .674971    4.114029
       educx |   .5266878   .2651831     1.99   0.047     .0069384    1.046437
       _cons |  -1.691535   .7831351    -2.16   0.031    -3.226452   -.1566188
------------------------------------------------------------------------------
```

```
Block  2: cheducx

Logistic regression                               Number of obs   =         61
                                                  LR chi2(4)      =      31.32
                                                  Prob > chi2     =     0.0000
Log likelihood = -24.219269                       Pseudo R2       =     0.3927

------------------------------------------------------------------------------
    happymar |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      church |    3.17948   1.019044     3.12   0.002      1.18219     5.17677
      female |   2.309507   .8724212     2.65   0.008     .5995924    4.019421
       educx |   .3712194   .2865789     1.30   0.195     -.190465    .9329037
     cheducx |   .5493444   .6495377     0.85   0.398    -.7237261    1.822415
       _cons |  -1.707305   .7794918    -2.19   0.029    -3.235081    -.179529
------------------------------------------------------------------------------


  +----------------------------------------------------------------+
  | Block |        LL       LR     df  Pr > LR       AIC       BIC |
  |-------+--------------------------------------------------------|
  |     1 | -24.63378    30.50      3   0.0000  57.26757  65.71106 |
  |     2 | -24.21927     0.83      1   0.3626  58.43854  68.99291 |
  +----------------------------------------------------------------+
```

The incremental chi-square is only .83 with 1 d.f., and is not significant. (To do it by hand, the differences between the two Model chi-squares are 31.32 – 30.50 = .82 with 1 d.f.) The z statistic (0.85) is similar. Ergo, we conclude that cheducx is not statistically significant, i.e. the effect of education does not vary by church attendance.