Revised July 2018

Logistic Regression Analysis

This set of notes shows how to use Stata to estimate a logistic regression equation. It assumes that you have set Stata up on your computer (see the "Getting Started with Stata" handout), and that you have read in the set of data that you want to analyze (see the "Reading in Stata Format (.dta) Data Files" handout).

In Stata, most tasks can be performed either by issuing commands within the "Stata command" window, **or** by using the menus. These notes illustrate both approaches, using the data file "GSS2016.DTA" (this data file is posted here: https://canvas.harvard.edu/courses/53958).

Stata requires that the dependent variable for a logistic regression is a dichotomous variable that is coded 1 if someone has the outcome of interest, and 0 otherwise. If yours is not, you need to recode it before beginning (or you will get an error message). (See the "Recoding Variables…" handout for more information).

The example used here is the association between support for gay marriage and years of schooling. To recode a categorical variable that measures whether or not LGBTQ individuals have the right to marry ("marhomo," 1=strongly agree…5=strongly disagree) as a proportion, the recoding commands might be

```
recode marhomo (1/2=1 "Favor")(3/5=0 "Neutral or
oppose"),gen(marhomo_r)

label variable marhomo_r "Favorable view toward gay marriage"
```

(Note: enter the commands in Stata on one line. "oppose…" is displayed on the second line above due to Microsoft Word formatting.).

To estimate the logistic regression of a dependent variable on an independent variable, issue the following command:

```
logit <depvar> <indepvar>, nolog
```

where you fill in the name of your 0/1 response variable in place of "<depvar>" and the name of your explanatory variable in place of "<indepvar>". There may be more than one explanatory variable, but the dependent variable must always come first in the list. The "nolog" option suppresses some technical output that you needn't look at for the time being.

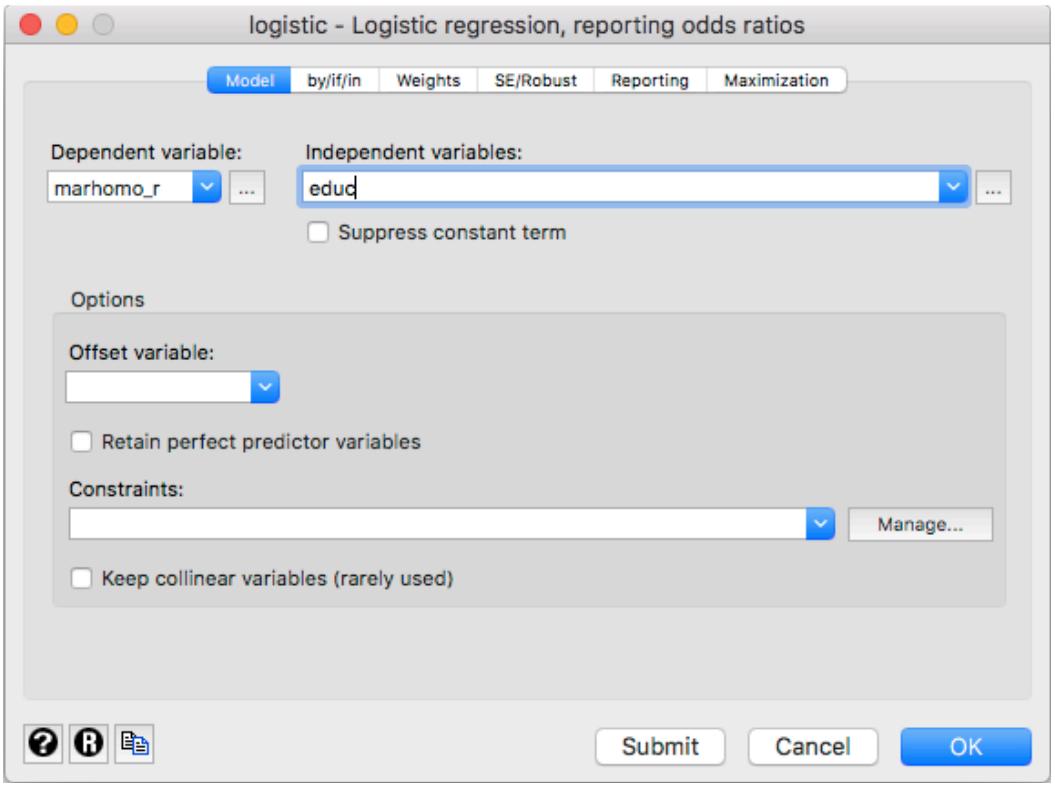
For the gay marriage-education example the command would be:

```
logit marhomo_r educ, nolog
```

Using the Stata menus, you can estimate a logistic regression as follows:

click on "Statistics"
click on "Binary outcomes"
click on "Logistic regression"

A window like the one below will open up:



Fill in the name of your 0/1 response variable in the "Dependent variable:" box and the name of your explanatory variable(s) in the "Independent variables:" box. To report the coefficients rather than the odds-ratios (we discuss odds-ratios later in this handout), go to the "Reporting" tab and click a button that says "Report estimated coefficients". If you want to suppress the log with technical output, go to the "Maximization" tab and click a button that says "Suppress" under "Iteration Log". Then click "OK".

Whether via the menus or via the command line approach, the following output appears:

```
. logit marhomo_r educ, nolog

Logistic regression                               Number of obs     =      1,858
                                                  LR chi2(1)        =      87.07
                                                  Prob > chi2       =     0.0000
Log likelihood = -1211.1706                       Pseudo R2         =     0.0347

------------------------------------------------------------------------------
   marhomo_r |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1529801   .0170933     8.95   0.000     .1194779    .1864823
       _cons |  -1.711666   .2370876    -7.22   0.000    -2.176349   -1.246983
------------------------------------------------------------------------------
```

The coefficient estimates are in the "Coef." column in the bottom panel of this chart; the number beside the independent variable (educ) is the regression coefficient or slope b. Here this tells us that the logit (or "log odds") for being in favor of gay marriage is estimated to rise by 0.15 with each year of education. Since bigger logits correspond to bigger probabilities, this means that the more educated are more apt to support gay marriage.

The z statistic and confidence interval for the regression coefficient permit us to draw inferences about the relationship between the variables. We are quite confident that the positive education-support for gay marriage association can be generalized beyond this sample.

The log-odds is not a terribly intuitive quantity. Analysts often prefer to interpret the results of logistic regression using the odds and odds ratios rather than the logits (or log-odds) themselves. Applying an exponential (exp) transformation to the regression coefficient gives the odds ratio; you can do this using most hand calculators.

You can, however, obtain odds ratios directly by requesting the "or" option as part of the "logit" command or, using the Stata menu, go to the "Reporting" tab and click a button that says "Report odds ratios" (Stata by default reports odds ratios if you run a logistic regression using the menus but not if you use the command-line approach).

This reports odds ratios–which give multiplicative effects on the odds–rather than additive effects on the log-odds or logits. It does not alter the type of analysis done at all, only the terms in which it is expressed.

For this example, this yields

```
. logit marhomo_r educ, nolog or

Logistic regression                               Number of obs     =      1,858
                                                  LR chi2(1)        =      87.07
                                                  Prob > chi2       =     0.0000
Log likelihood = -1211.1706                       Pseudo R2         =     0.0347

------------------------------------------------------------------------------
   marhomo_r | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.165302   .0199188     8.95   0.000     1.126908    1.205003
       _cons |   .1805648   .0428097    -7.22   0.000      .113455    .2873706
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

The odds of being in support of gay marriage are predicted to grow about 1.17 times larger for each additional year of education. So, if two people differ by 2 years of education, the person with more education has predicted odds of being in support of gay marriage that are 1.17*1.17 or 1.37 times larger than the person with less education. If two people differ by 10 years of education, the odds that the person with more education is in support of gay marriage are 1.17^{10} or 4.8 times larger than those of the person with less education.

Odds ratios greater than 1 correspond to "positive effects" because they increase the odds. Those between 0 and 1 correspond to "negative effects" because they decrease the odds. Odds ratios of exactly 1 correspond to "no association." An odds ratio cannot be less than 0.

As in "regular" regression, you can add control variables to a logit regression by extending the list of independent variables. For example, adding age (“age”) and sex differences (“female” indicator variable) to the above regression gives the following estimates:

```
. logit marhomo_r educ age female, nolog or

Logistic regression                               Number of obs     =      1,850
                                                  LR chi2(3)        =     156.71
                                                  Prob > chi2       =     0.0000
Log likelihood = -1170.6592                       Pseudo R2         =     0.0627

------------------------------------------------------------------------------
   marhomo_r | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.164621   .0201925     8.79   0.000     1.125709    1.204878
         age |   .9768814   .0028191    -8.11   0.000     .9713716    .9824224
      female |    1.14239   .1138453     1.34   0.182     .9396962    1.388804
       _cons |   .5499887   .1554764    -2.11   0.034     .3160293    .9571502
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Here we see that the odds of supporting gay marriage are predicted to

> grow 1.16 times larger (controlling for age and sex) with each additional year of education
>
> shrink by a factor of about 0.02 (controlling for education and sex) with each additional year of age: the odds that a person who is 1 year older are predicted to be about 0.98 times as large as those for a person who is 1 year younger)
>
> be about 1.14 times larger among women (controlling for age and education) than they are among men

So: being in support of gay marriage is more common among the educated, younger people, and women.

The "LR chi2" reported at the upper right here is analogous to the overall F-statistic in multiple regression. It asks if using the logistic regression improves our ability to predict the response variable.

Predicted values of the response variable can be obtained for logistic regression just as they are for "regular" regression. Stata produces them using the same kind of post-estimation command used in linear regression, but this handout will not go into the details on how to do that.