Revised July 2018

Multiple Linear Regression Analysis

This set of notes shows how to use Stata in multiple regression analysis. 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).

To estimate the linear regression of a response variable on K explanatory variables, issue the following command:

```
regress <depvar> <indepvar1> <indepvar2> … <indepvarK>
```

Where you fill in the name of your response variable in place of "<depvar>" and the name of your explanatory variables in place of "<indepvar1>", "<indepvar2>", etc. The response variable must always be listed first in the list. The order in which the explanatory variables are listed does not matter. The explanatory variables may include indicator variables (see the "Multiple Linear Regression Analysis with Indicator Variable" handout) or other terms that you construct yourself, as well as quantitative variables.

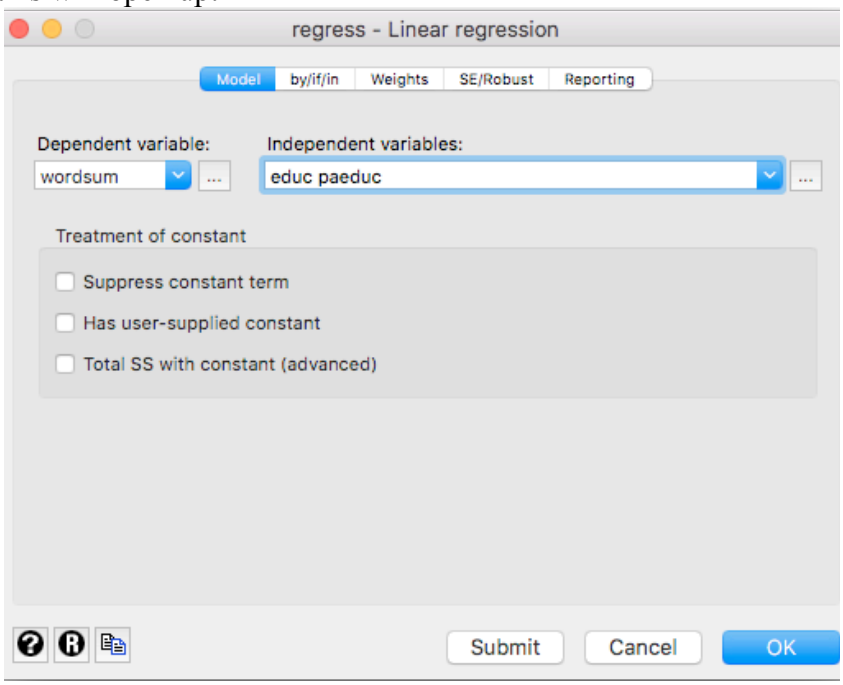
For a regression of vocabulary test score ("wordsum") on education ("educ") and father's education ("paeduc"), the command would be:

```
regress wordsum educ paeduc
```

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

click on "Statistics"
click on "Linear models and related"
click on "Linear regression"

A window like this will open up:



Fill in the name of your response variable in the "Dependent variable:" box and the name of your explanatory variables in the "Independent variables:" box, and click "OK".

Whether via the menus or via the command line approach, the following output will appear in the Stata "Results" window:

```
. reg wordsum educ paeduc

      Source |       SS           df       MS      Number of obs   =     1,355
-------------+----------------------------------   F(2, 1352)      =    167.68
       Model |  963.189642         2  481.594821   Prob > F        =    0.0000
    Residual |  3883.11589     1,352  2.87212714   R-squared       =    0.1987
-------------+----------------------------------   Adj R-squared   =    0.1976
       Total |  4846.30554     1,354  3.57925076   Root MSE        =    1.6947

------------------------------------------------------------------------------
     wordsum |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .2456815    .017593    13.96   0.000     .2111689    .2801941
      paeduc |   .0720562   .0126831     5.68   0.000     .0471755     .096937
       _cons |   1.912368   .2409929     7.94   0.000     1.439607    2.385129
------------------------------------------------------------------------------
```

The estimates that determine the regression equation (a plane, when there are two explanatory variables) are in the "Coef." column in the bottom panel of this chart; the number beside "cons" is the "constant" or "intercept" **a**, here 1.91; the number beside each explanatory variable is its "partial regression coefficient" or "slope" **b**. Here, the coefficient of 0.25 for respondent's education ("educ") indicates that a 1-year rise in respondent's education is associated with a 0.25 word rise in vocabulary test score, among persons having the same level of father's education; the coefficient of 0.07 for father's education ("paeduc") indicates that a 1-year rise in father's

education is associated with a 0.07 word rise in vocabulary test score, among persons having the same level of education.

The coefficient of determination is reported in the upper right panel as "R-squared": about 20% of the variation in vocabulary test score is "explained" by respondent's education and father's education. The standard error of estimate (or residual standard deviation) is reported as "Root MSE." The standard deviation of residuals or "errors" (that is, a typical distance of an actual observation from its predicted value under the regression) around the regression plane is about 1.7.

At the upper left is an analysis of variance table that leads to the F statistic reported at the upper right (167.7, with 2 and 1352 df). At the bottom are standard errors and t-statistics for the slopes and intercept, as well as 95% confidence intervals for those statistics.

If you want Stata to calculate and print the standardized (beta) coefficients, select the "Reporting" tab of the above screen and check the box for "Standardized beta coefficients" there, before clicking OK. Alternatively, add the "beta" option to the command-line version of this command, as follows for this example:

regress wordsum educ paeduc, beta

In this case, you will see the following output:

```
. regress wordsum educ paeduc, beta

      Source |       SS           df       MS      Number of obs   =     1,355
-------------+----------------------------------   F(2, 1352)      =    167.68
       Model |  963.189642         2  481.594821   Prob > F        =    0.0000
    Residual |  3883.11589     1,352  2.87212714   R-squared       =    0.1987
-------------+----------------------------------   Adj R-squared   =    0.1976
       Total |  4846.30554     1,354  3.57925076   Root MSE        =    1.6947

------------------------------------------------------------------------------
     wordsum |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        educ |   .2456815    .017593    13.96   0.000                .3672722
      paeduc |   .0720562   .0126831     5.68   0.000                .1494177
       _cons |   1.912368   .2409929     7.94   0.000                      .
------------------------------------------------------------------------------
```

As you can see, this is identical to the above, except that the "Beta" coefficient appears in place of the confidence interval.

Predicted values of the response variable based on a regression are often useful. If you want to compute and save them, follow the exact same procedures used in doing this for bivariate regression (see the Stata handout for bivariate regression).

Testing hypotheses about subsets of explanatory variables. The output shown earlier for regression analysis allows you to test hypotheses about single partial coefficients (via the t statistic) or about all of the partial regression coefficients taken together (via the F statistic).

Sometimes one will wish to test a hypothesis about a subset of two or more–but not all–the regression coefficients. You can do this by using one of Stata's "post-estimation" commands, immediately after you have estimated the regression. For example, here is the same regression from page 1 but with the variable age added to the set of explanatory variables:

```
.  regress wordsum educ paeduc age

      Source |       SS           df       MS      Number of obs   =     1,350
-------------+----------------------------------   F(3, 1346)      =    124.90
       Model |  1051.18418         3  350.394727   Prob > F        =    0.0000
    Residual |  3776.00174     1,346  2.80535048   R-squared       =    0.2178
-------------+----------------------------------   Adj R-squared   =    0.2160
       Total |  4827.18593     1,349   3.5783439   Root MSE        =    1.6749

------------------------------------------------------------------------------
     wordsum |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .2357611   .0174946    13.48   0.000     .2014413    .2700808
      paeduc |   .0965124   .0131911     7.32   0.000     .0706351    .1223897
         age |   .0154274   .0026967     5.72   0.000     .0101372    .0207175
       _cons |   .9948572   .2872937     3.46   0.001     .4312652    1.558449
------------------------------------------------------------------------------
```

Perhaps you would like to test the hypothesis that the partial coefficients for father's education and age are both zero; if retained (i.e., not rejected), this hypothesis would suggest that only education shapes vocabulary test score.

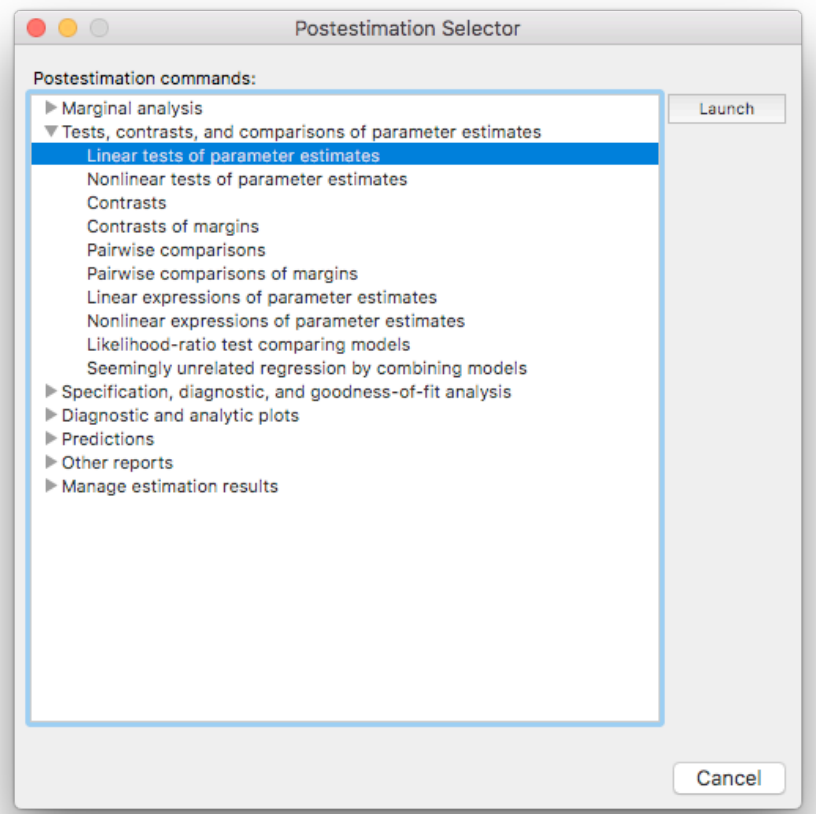
Using the command line, you can accomplish this by typing

```
test paeduc age
```

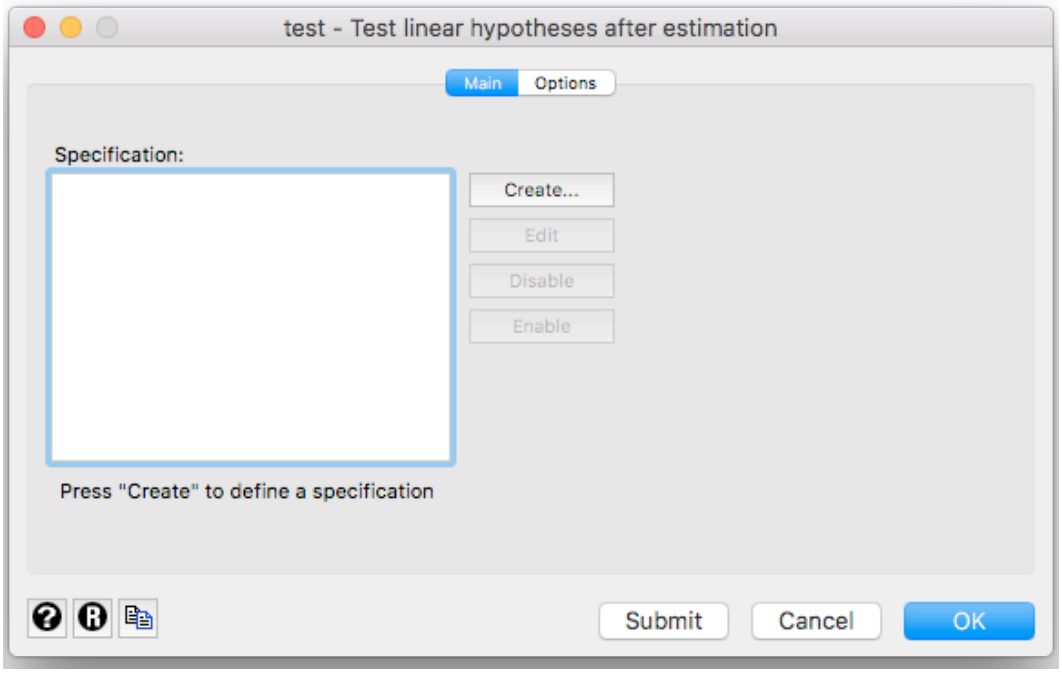
right after you estimate the regression.

Alternately, you can use the menu options to do this, again immediately after the regression.

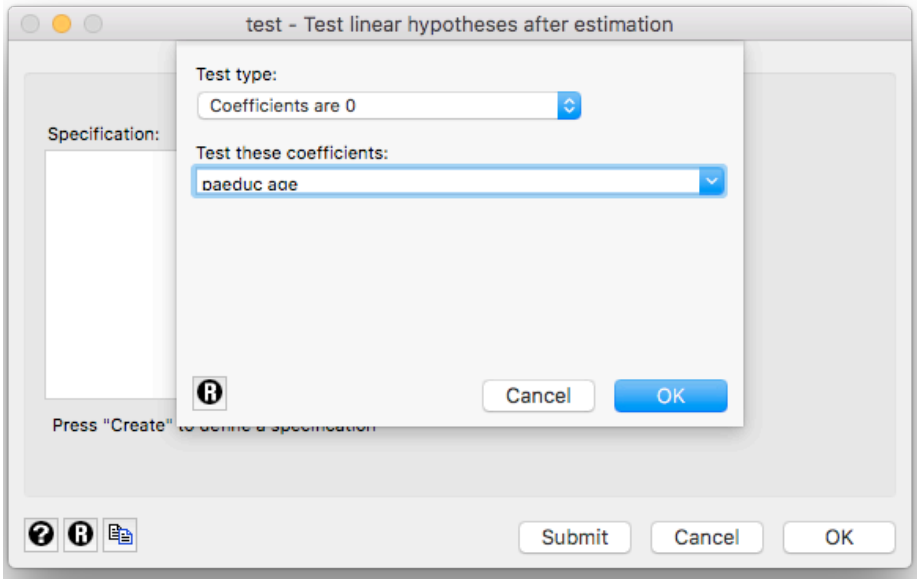
Click on "Post-estimation"
Click on the arrow next to "Tests, contrasts, and comparisons…"
Select "Linear tests of parameter estimates"
Click "Launch"



The screen below will appear:



Click “create.” The following window will pop up:



Fill in the names of the variables corresponding to the coefficients in your joint hypothesis. Then click "OK."

Either way, the following output will appear:

```
. test paeduc age

 ( 1)  paeduc = 0
 ( 2)  age = 0

       F(  2,  1346) =   33.40
            Prob > F =    0.0000
```

This F statistic for removing both paeduc and age from the regression (while leaving education in the regression) offers strong evidence against the hypothesis that the regression coefficients for father's education and age, controlling for education, are simultaneously 0.