Revised July 2018

Bivariate (Simple) Regression Analysis

This set of notes shows how to use Stata to estimate a simple (two-variable) 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).

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

```
regress <depvar> <indepvar>
```

Where you fill in the name of your response variable in place of "<depvar>" and the name of your explanatory variable in place of "<indepvar>". The response variable must always be listed first in the list (in multiple regressions later on, you will add additional explanatory variables to the command).

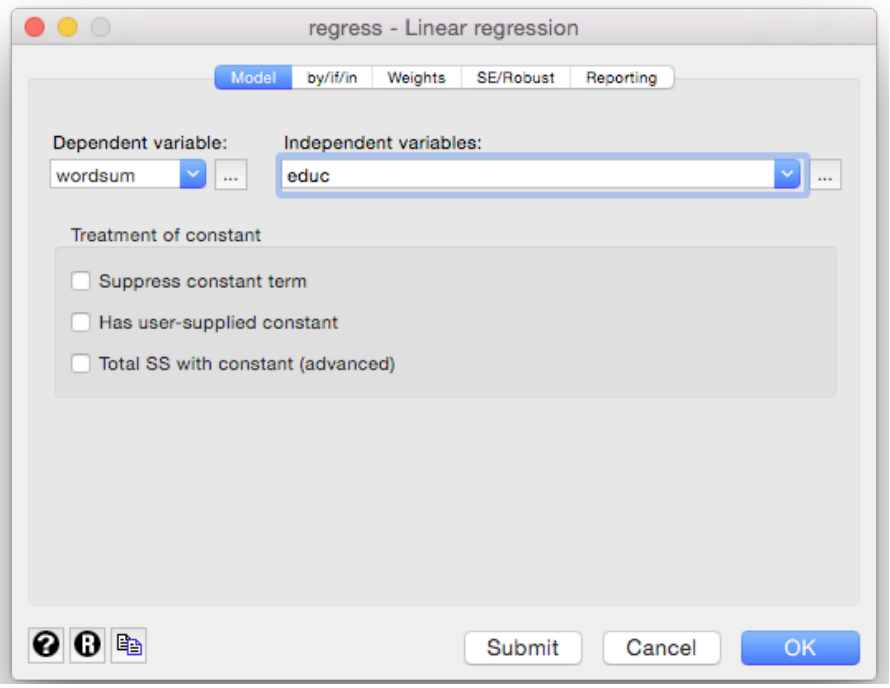
For a regression of vocabulary test score on educational attainment, the command would be:

```
regress wordsum educ
```

Using the Stata menus, you can estimate a simple 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 variable 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:

```
. regress wordsum educ

      Source |       SS           df       MS      Number of obs   =     1,861
-------------+----------------------------------   F(1, 1859)      =    454.69
       Model |  1348.95828         1  1348.95828   Prob > F        =    0.0000
    Residual |  5515.22442     1,859  2.96676946   R-squared       =    0.1965
-------------+----------------------------------   Adj R-squared   =    0.1961
       Total |   6864.1827     1,860  3.69042081   Root MSE        =    1.7224

------------------------------------------------------------------------------
     wordsum |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .2972705    .013941    21.32   0.000     .2699288    .3246122
       _cons |   1.941593   .1954309     9.93   0.000     1.558306     2.32488
------------------------------------------------------------------------------
```

The estimates that describe the regression line are in the "Coef." column in the bottom panel of this chart; the number beside "_cons" is the "constant" or "intercept" **a**, here 1.94; the number beside the independent variable (educ) is the "regression coefficient" or "slope" **b**, here 0.297.

The coefficient of determination is reported in the upper right panel as "R-squared" and the standard error of estimate (or residual standard deviation) is reported as "Root MSE".

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

If you want Stata to 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, beta
```

Either way, you will see the following output:

```
. regress wordsum educ, beta

      Source |       SS           df       MS      Number of obs   =     1,861
-------------+----------------------------------   F(1, 1859)      =    454.69
       Model |  1348.95828         1  1348.95828   Prob > F        =    0.0000
    Residual |  5515.22442     1,859  2.96676946   R-squared       =    0.1965
-------------+----------------------------------   Adj R-squared   =    0.1961
       Total |   6864.1827     1,860  3.69042081   Root MSE        =    1.7224

------------------------------------------------------------------------------
     wordsum |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        educ |   .2972705    .013941    21.32   0.000                .4433073
       _cons |   1.941593   .1954309     9.93   0.000                       .
------------------------------------------------------------------------------
```

As you can see, this is identical to the output on the previous page, except that the "Beta" coefficient (always equal to the Pearson correlation, for bivariate regression) appears in place of the confidence interval.

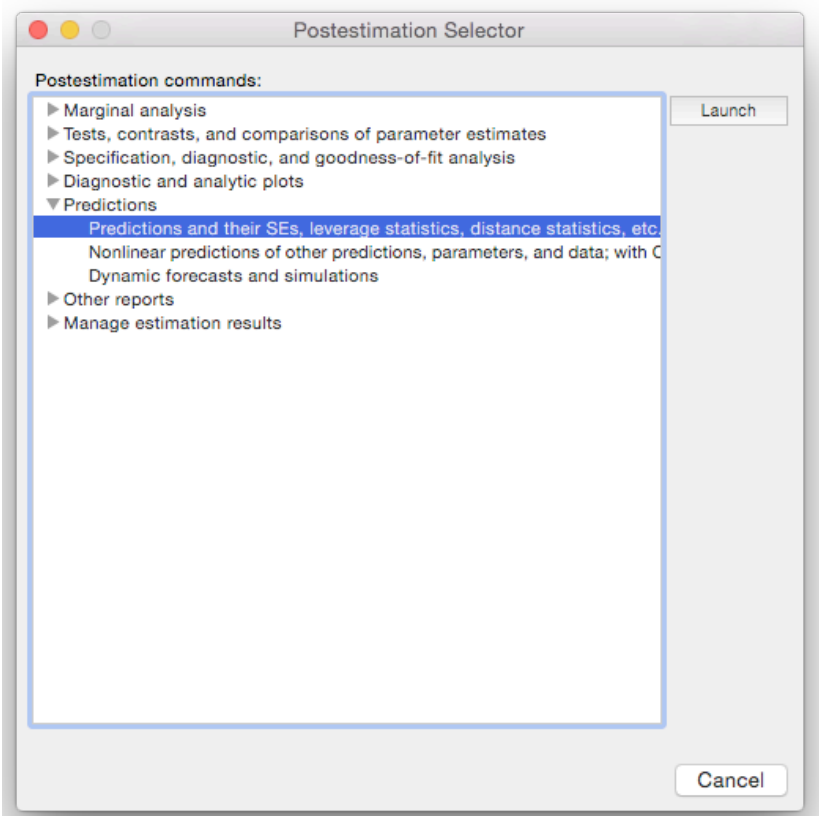
Predicted values of the dependent variable based on a regression are often useful. To obtain these, you issue the following command immediately after you estimate the regression:

```
predict pred, xb
```

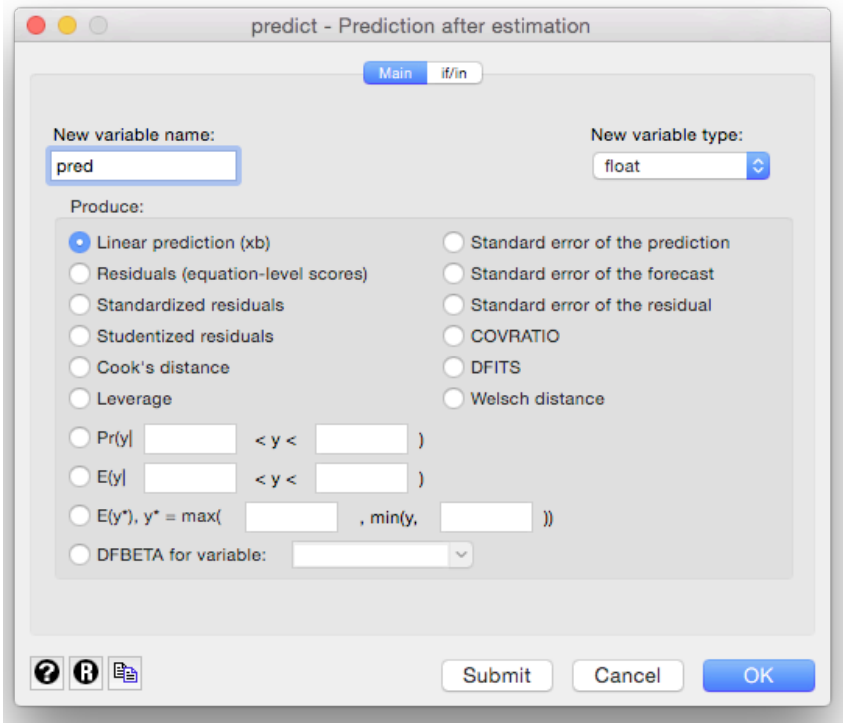
This will calculate the predicted value of the dependent variable for each observation and store it as a new variable named "pred" (you can select a different variable name than "pred" if you want).

Alternately, you can use the "Postestimation" menu to do this, again immediately after you estimate the regression.

Click on "Statistics"
Click on "Post-estimation"
Click on the arrow next to "Predictions"
Select "Predictions and their SEs…"
Click "Launch"



The screen shown below will appear:



Fill in the name of the variable that is to store the predicted values in the "New variable name:" box. Leave the "Linear prediction (xb)" button in the "Produce:" box checked. Then click "OK".

Either way (command line or menus), you will see little if any output in the Stata Results window. If there are any missing values on the new variable, you will be told so. For example:

```
(9 missing values generated)
```

You will see a new variable name in the "Variables" window at the right of the Stata screen, however. Henceforth you can refer to this ("pred", in this case) just like any other variable in the data set, when executing other Stata commands.