

Scatterplots

Scatterplots display the joint distribution of two quantitative variables. This handout 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>).

This handout focuses mainly on the Stata menu approach to obtaining scatterplots, but display equivalent command-line language.

Good, basic scatterplots can be produced relatively easily, as follows:

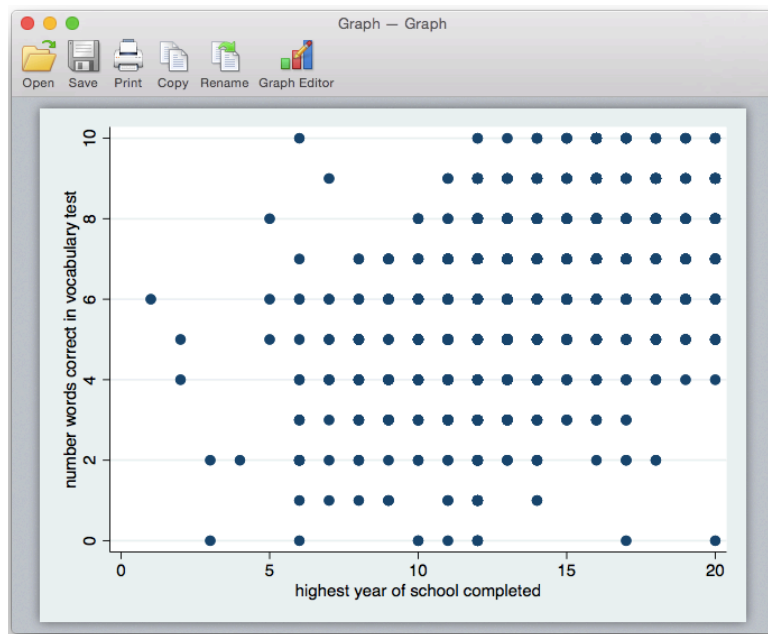
Via the command line, one issues the following generic command to obtain a scatterplot:

```
scatter <yvariable> <xvariable>
```

To create a scatterplot of number of words correct on a vocabulary test and years of schooling, the command would be:

```
scatter wordsum educ
```

A separate “Stata Graph” window will appear, displaying this scatterplot:

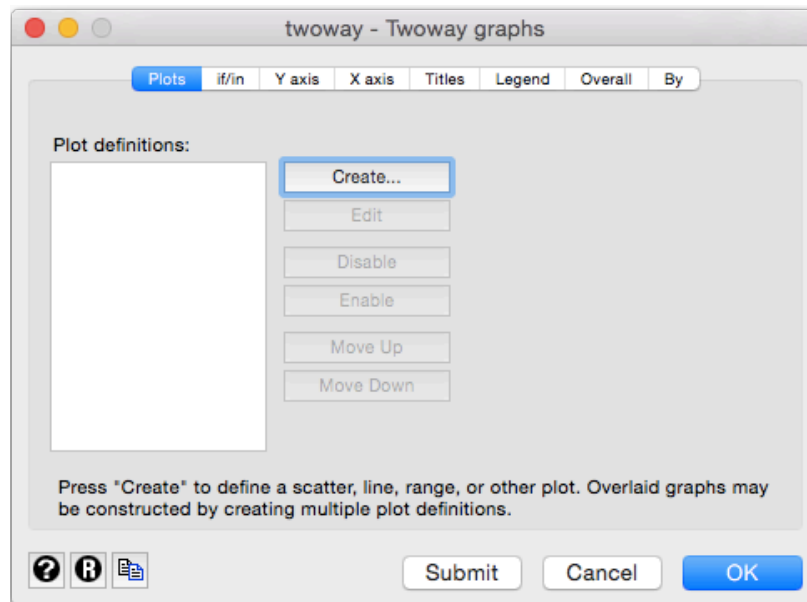


Via the menu options:

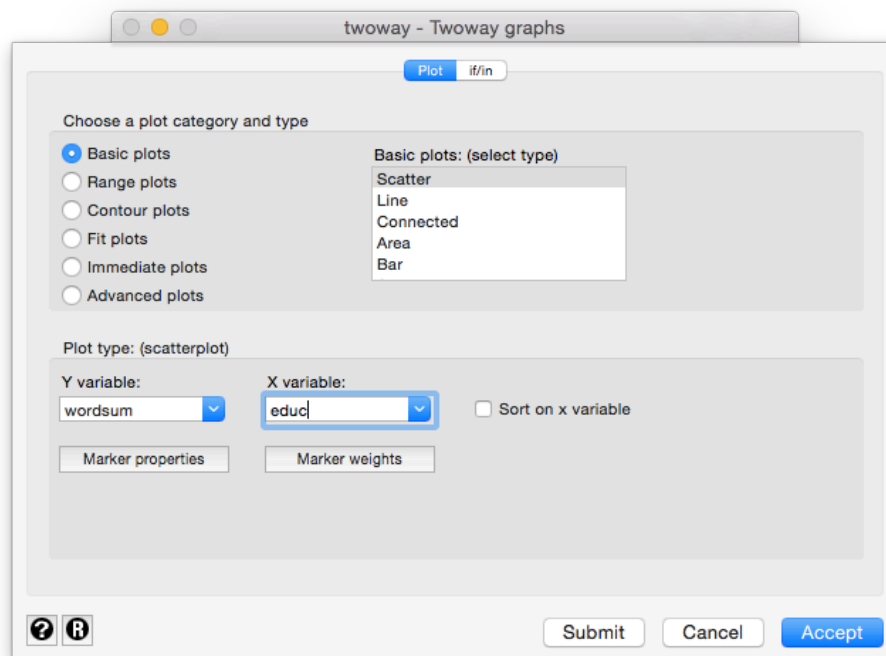
Click on "Graphics"

Click on "Twoway graph", which will open a "Twoway Graphs" window

Click on "Create"

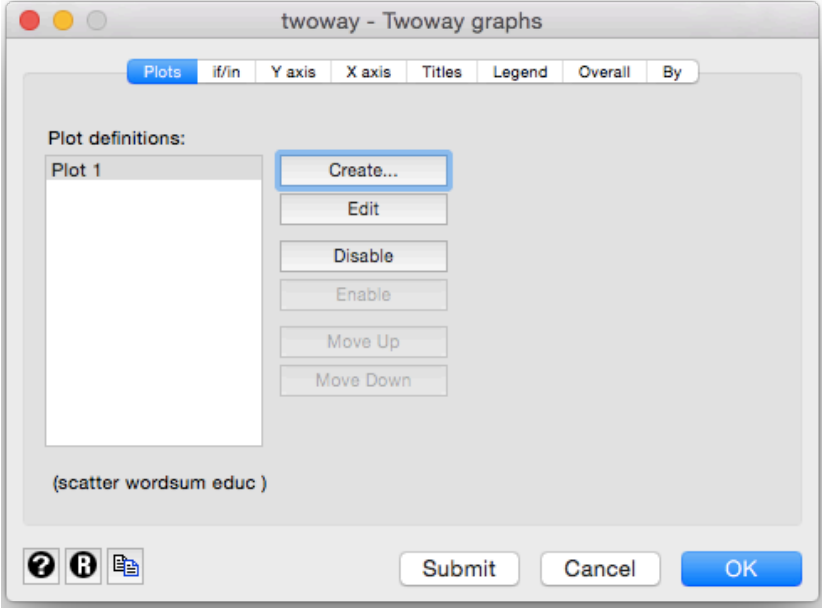


When you do this, a window like this opens:

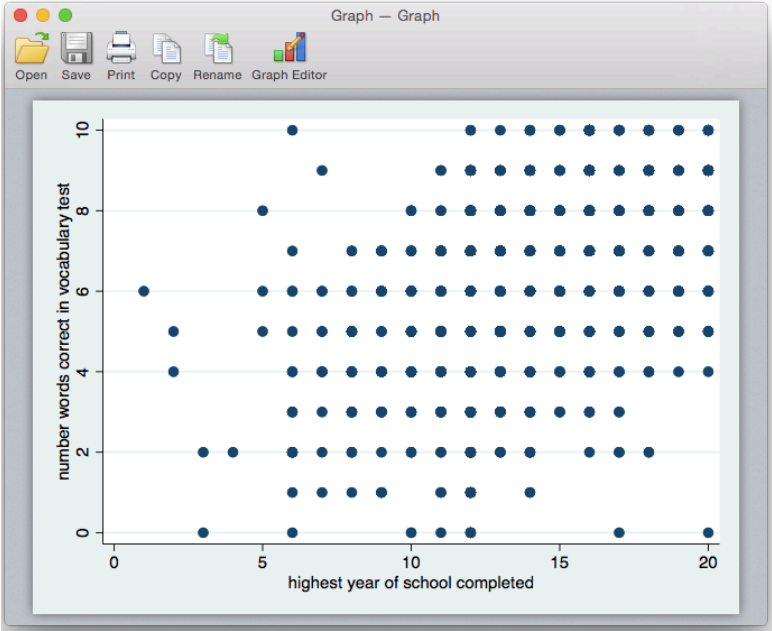


Select the button for "Basic plots", choose "Scatter" as the plot type in the box to the right, and then fill in your Y variable ("wordsum" in this example) and X variable ("educ" in this example). Then click "Accept."

This will return you to the "Tway Graph" screen, which will look like this:



Click "OK" to display the scatterplot, which for this example looks like this:

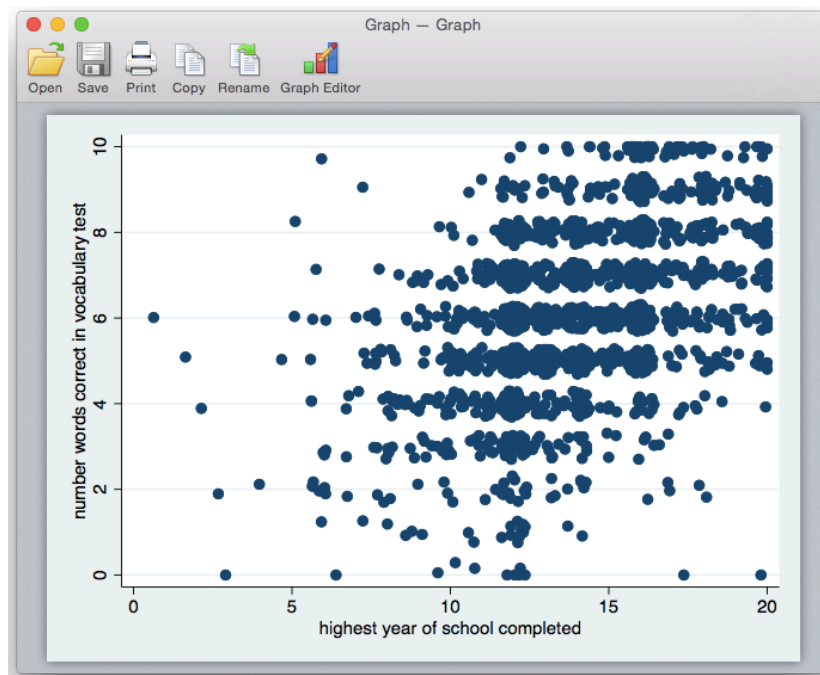


Stata plots observations with identical X and Y values directly on top of one another. So in the example above, it *appears* that there about a hundred observations or so, however *in fact* there are over 1,800 observations. Most of the observations are identical to others and therefore are plotted on top of one another. The “jitter” option is a useful tool to help better visualize the data. “Jitter” adds random noise to the observations before generating the scatterplot, yielding a better visual sense of how many observations have each pair of X and Y values. “Jitter” does not modify the data permanently.

To include the “jitter” option for the example above type the following command:

```
scatter wordsum educ, jitter(5)
```

It is recommended to specify either jitter(5) or jitter(6). The command above produces the following graph.



Via the menu options:

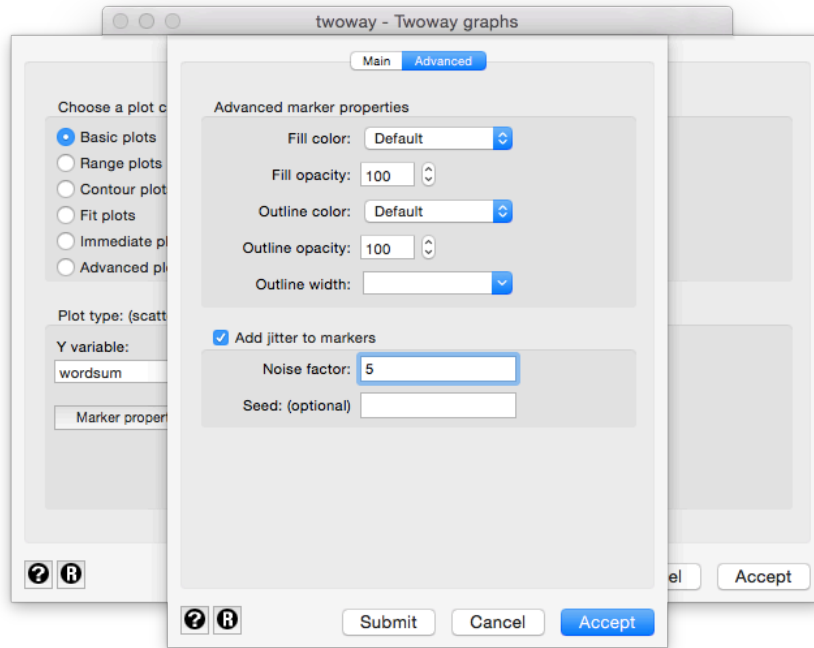
- Click on “Graphics”

- Click on “Twoway graph”, which will open a “Twoway Graphs” window

- Click on “Plot 1” (*the scatterplot just generated*)

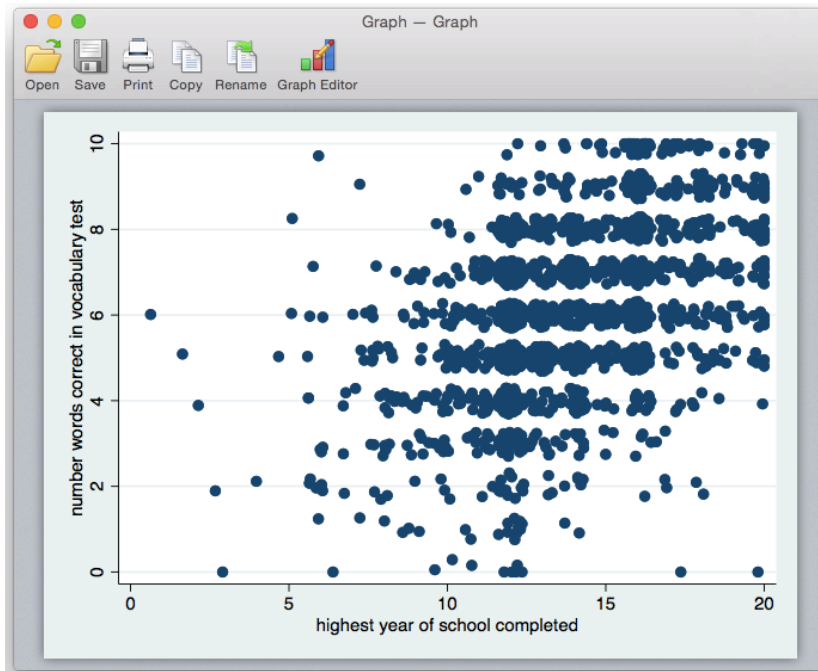
- Click on “Edit”

This will bring up the window in which the variables to plot are selected. In this window, click on “Marker Properties” and toggle over to the “Advanced” tab:



Check "Add jitter to markers." Enter "5" next to "Noise Factor." Click "Accept."

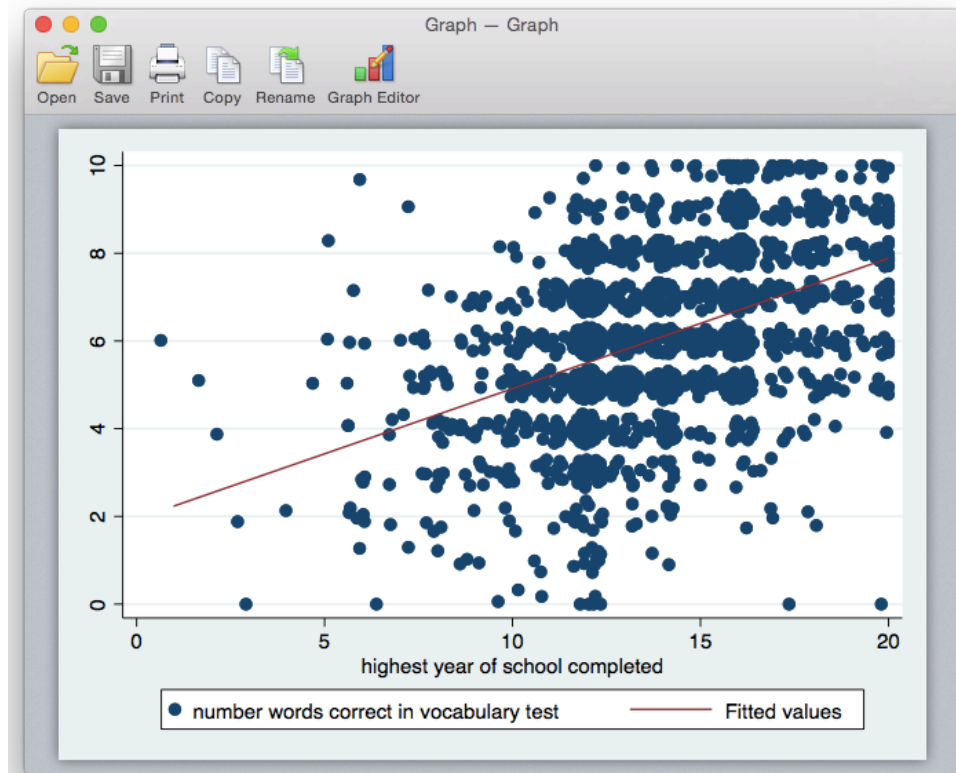
Next, click "Accept" again. Then click "OK" to display the jittered scatterplot, which for this example looks like this:



To superimpose an estimated regression line on a scatterplot, proceed as follows:

```
twoway (scatter wordsum educ, jitter(5)) (lfit wordsum educ)
```

A separate “Stata Graph” window will appear, displaying the following scatterplot:



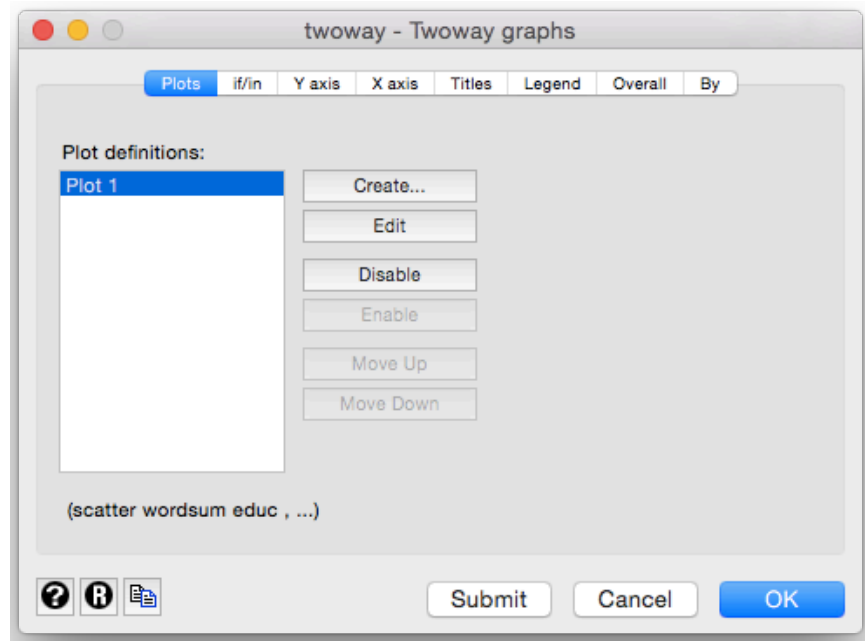
Via the menu options:

Click on "Graphics"

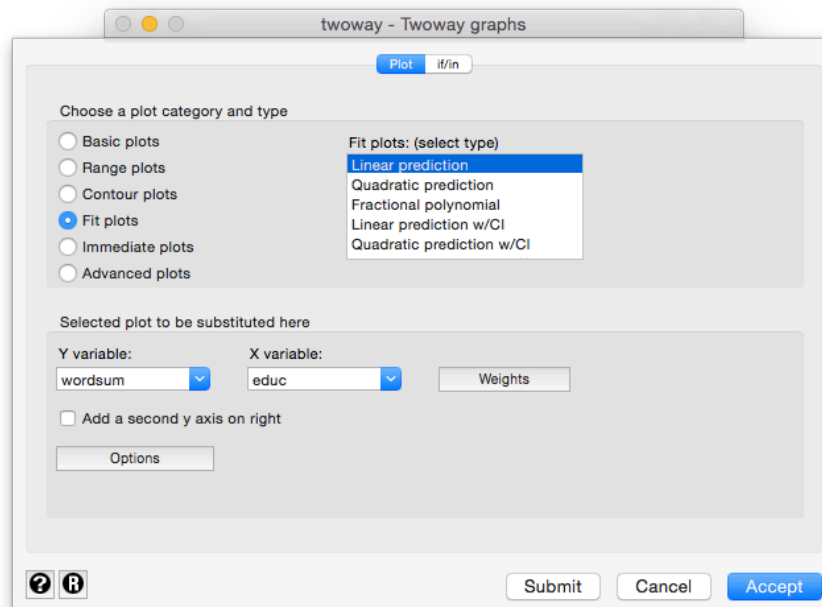
Click on "Twoway graph" to open the "Twoway Graphs" menu

(You will see that Plot1, the scatterplot above, is already created.)

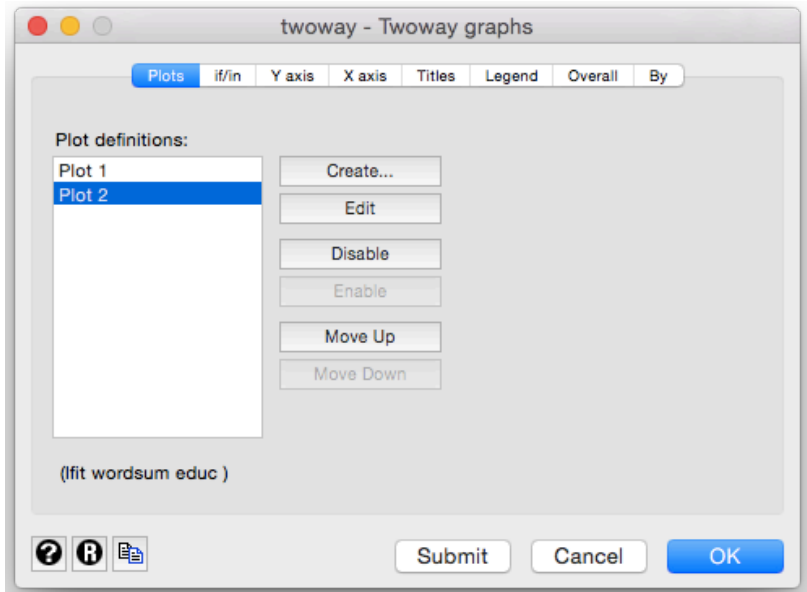
Click on "Create" to create the regression line plot.



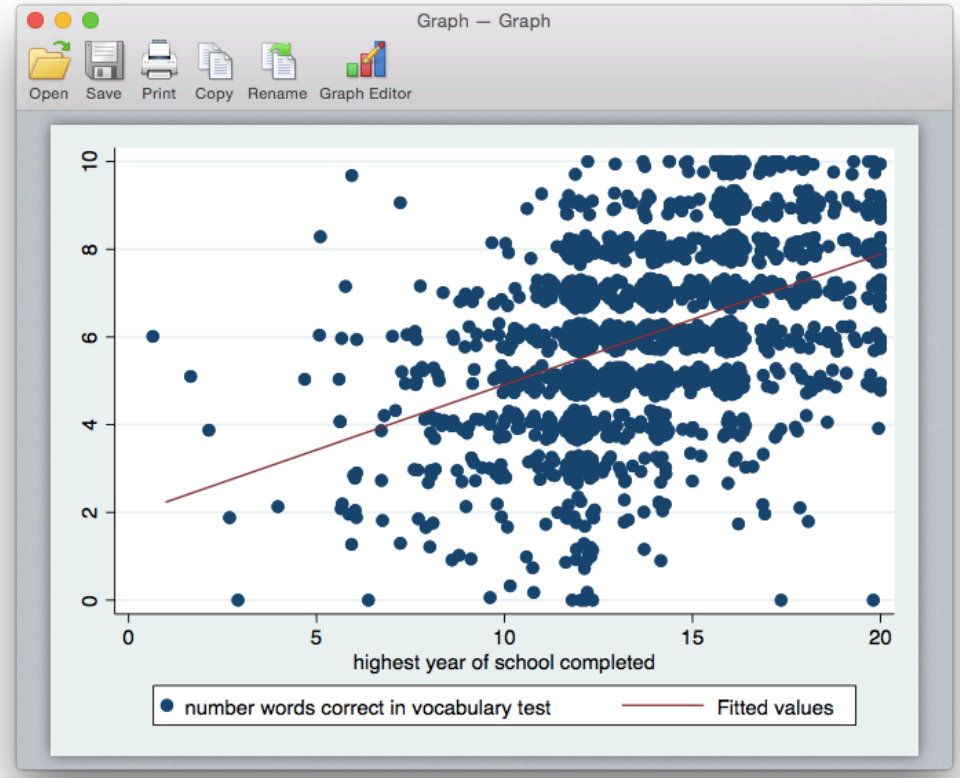
Select the button for "fit plots" and then choose "linear prediction" as the type. Choose the same Y and X variables you used for the scatterplot. This creates "Plot 2" and maps the regression line for Y on X on top of the scatterplot created on the Plot 1 screen.



Click "Accept" to return to the "Twoway Graphs" menu:

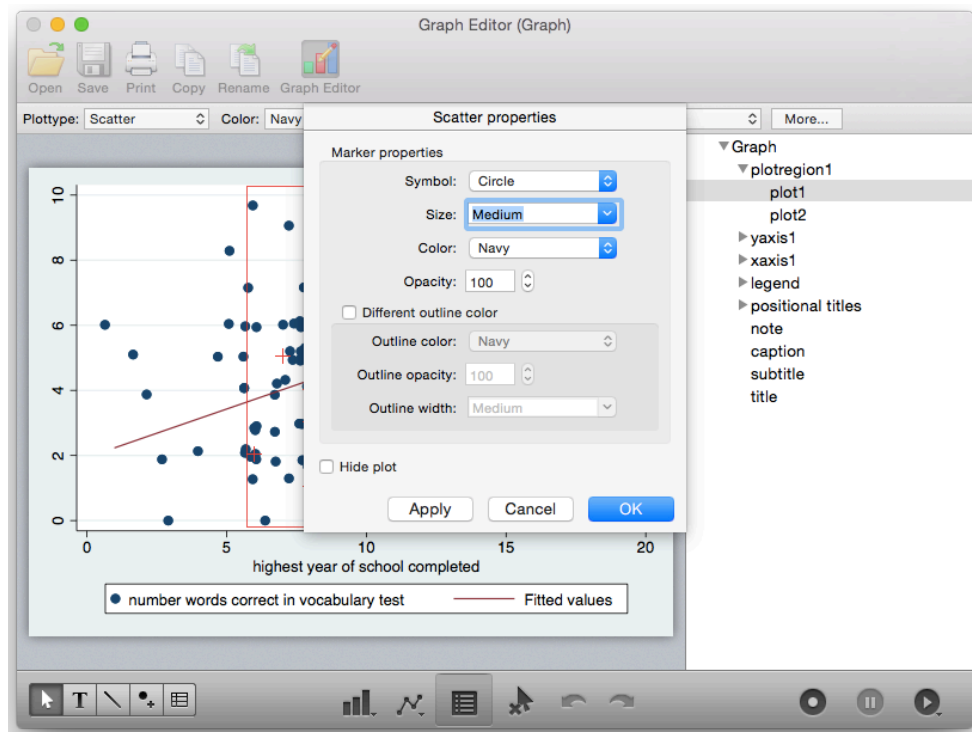


Then click OK to display the scatterplot with the regression line, which looks like this:



Enhancements and extensions.

You can change plot symbols and plot colors in the plot. For the example above, click on the "Graph Editor" button in the Graph window. Then click on the arrow next to the word "plotregion1." Double click on "plot1" to modify the plot symbols and colors in the scatterplot. Make your selections and then click "OK."



Each of the command menus offers tabs and options that you can use to create titles, subtitles, etc., label axes, and do various other tricks. If, for example, you want to label the dots in the scatterplot—proceed as follows (this will only be useful for data sets that identify the cases/units in some way; for surveys in which respondents are kept anonymous, there is no reason to do it, though this handout explains how to do so anyway):

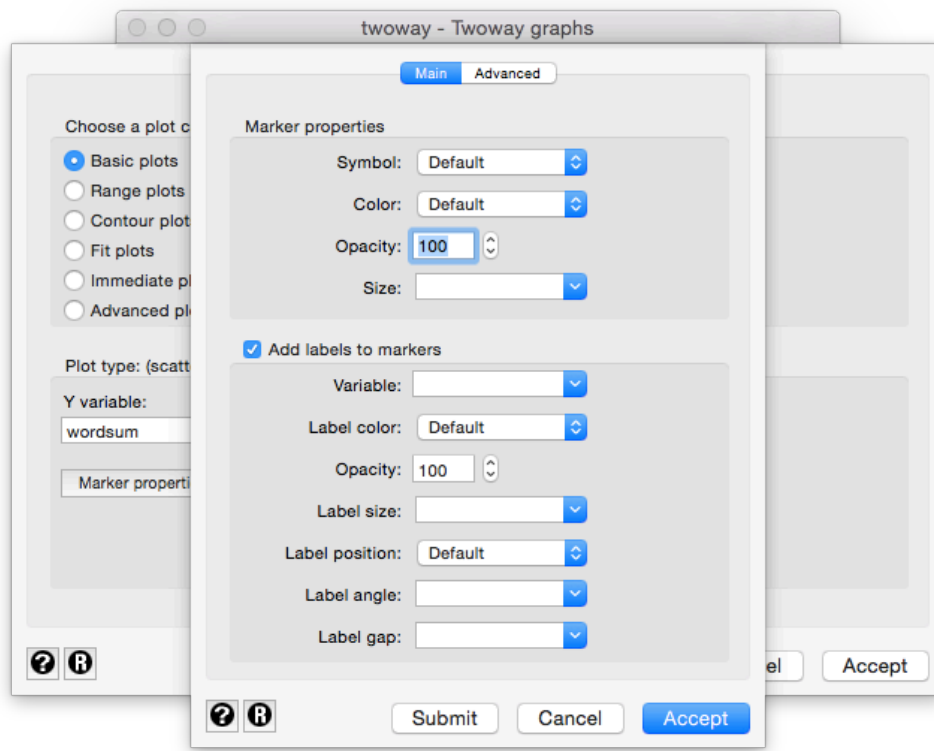
Click on "Graphics"

Click on "Twoway graph" to open the "Twoway Graphs" menu

(You will see that Plot1, the scatterplot above, is already created.)

Select "Plot1" and click "Edit."

Click the "Marker Properties" button. The following screen will appear:



Check the "Add labels to markers" box, and then select the variable to be used to label points (there isn't one for this example, but with other data sets you might use a variable that names a city, state, country, etc. that identifies the observations you are studying). Click "Accept" to return to the "Plot 1" screen, and then click "Accept" again to return to the "Twoway Graph" screen. Then click "OK."

Saving Graphs

To save your scatterplot(s), see the "Saving Graphs" handout.