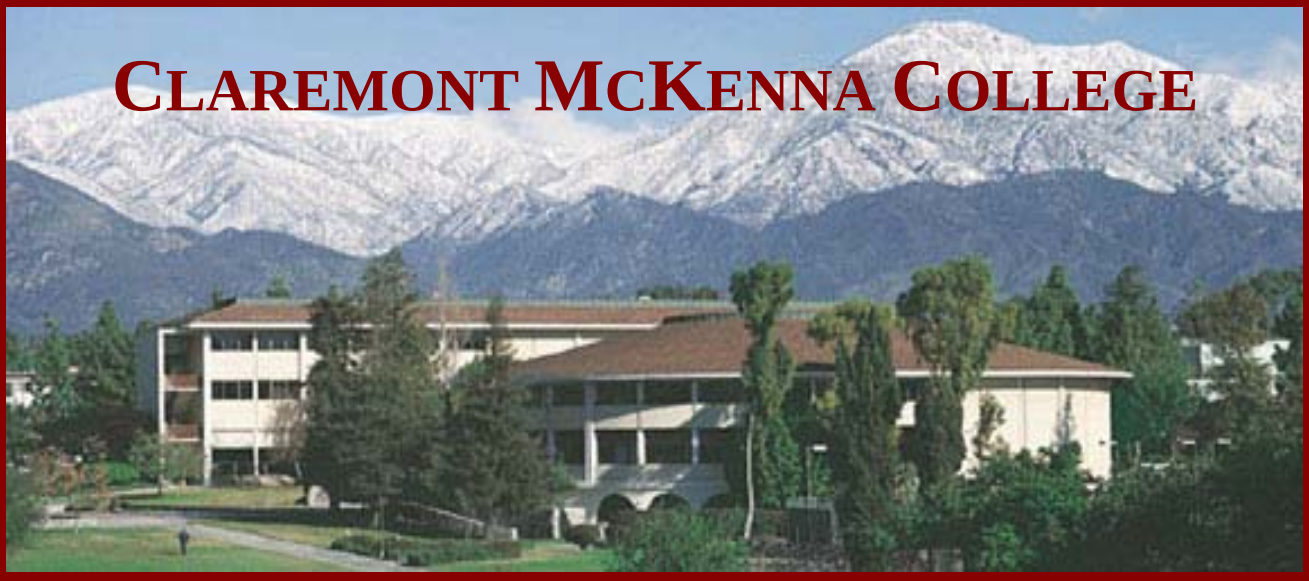
CLAREMONT MCKENNA COLLEGE



Fletcher Jones
Student Peer to Peer
Technology Training Program

Basic Statistics using Stata

ETS Educational Technology Services
CLAREMONT MCKENNA COLLEGE

An Introduction to Stata

Stata 8.0 for Windows is an integrated statistical package used for data analysis, data management, and graphics creation. Its range of capability extends from simple summary statistics to extremely complex statistical models. It is primarily designed for researchers in the fields of econometrics, social science and biostatistics.

A Comparison of Statistical Packages

SPSS

- SPSS is often used by those with little data analysis and management experience because it is relatively easy to use. It has a point-and-click interface, and it is rarely, if ever, necessarily to utilize the syntax language.

- SPSS has a friendly data editor that resembles Excel that allows you to enter your data and attributes of your data (missing values, value labels, etc.) However, SPSS has a difficult time working with multiple data files.

- SPSS performs most general statistical analyses (regression, logistic regression, survival analysis, analysis of variance, factor analysis, and multivariate analysis). The greatest strengths of SPSS are in the area of analysis of variance and multivariate analysis. One major weakness is the inability to create robust regressions or obtain robust standard errors.

- Creating graphs is extremely easy in SPSS, given its point-and-click interface. They can easily be pasted into other applications.

- **Conclusion**: SPSS is the easiest package to use, but if you have multiple data files or want to utilize some advanced statistical techniques, another package may be better.

Stata

- While Stata is initially more difficult to use than SPSS, it is fairly easy to learn how to use it. Stata uses one line commands which can be entered one command at a time (a mode favored by beginners) or can be entered many at a time in a Stata program (a mode favored by power users).

- Stata has more capabilities in the area of data management than SPSS. Intercooled Stata can handle 2,047 variables while Stata/SE can handle up to 32,766. However, it still can experience problems with multiple data files.

- Stata performs most general statistical analyses. The greatest strengths of Stata are probably in regression (it has very easy to use regression diagnostic tools) and logistic regression. Stata also has an array of robust methods that are very easy to use, including robust regression, regression with robust standard errors, and many other estimation commands include robust standard errors as well. Stata also excels in the area of survey data analysis. The greatest

weaknesses in data analysis would probably be in the area of analysis of variance and traditional multivariate methods (e.g., manova, discriminant analysis, etc.).

- Like SPSS, Stata graphics can be created using Stata commands or using a point and click interface. Unlike SPSS, the graphs cannot be edited using a graph editor. The syntax of the graph commands is the easiest of the three packages and is also the most powerful. Stata graphs are publication quality graphs.

- **Conclusion**: Stata offers a good combination of ease of use and power. While Stata is easy to learn, it also has very powerful tools for data management, many cutting edge statistical procedures, the ability to easily download programs developed by other users and the ability to create your own Stata programs that seamlessly become part of Stata.

SAS

- SAS is generally the most powerful package, but it is also the most difficult to learn and use. To use SAS, you write SAS programs that manipulate your data and perform your data analyses. If you make a mistake in a SAS program, it can be hard to see where the error occurred or how to correct it.

- SAS is very powerful in the area of data management, allowing you to manipulate your data in just about any way possible. SAS allows you to perform sql queries on your SAS data files. SAS can work with many data files at once easing tasks that involve working with multiple files at once and can handle up to 32,768 variables. However, many complex operations can be done much easier in Stata or SPSS.

- SAS performs most general statistical analyses. The greatest strengths of SAS are probably in its ANOVA, mixed model analysis and multivariate analysis, while it is probably weakest in ordinal and multinomial logistic regression (because these commands are especially difficult), and robust methods. While there is some support for the analysis of survey data, it is quite limited as compared to Stata.

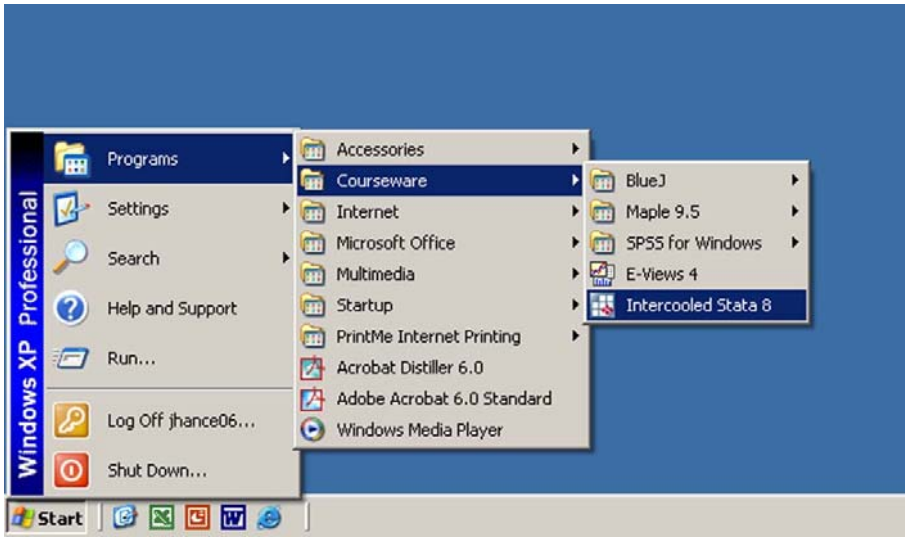
- SAS may have the most powerful graphic tools among all of the packages via SAS/Graph. However, SAS/Graph is also very technical and difficult to learn. The graphs are created largely using syntax language; however, SAS 8 does have a point-and-click interface for creating graphs but it is not as easy to use as SPSS.

- **Conclusion:** SAS is a package geared towards advanced users. It has a steep learning curve and can be extremely frustrating in the early going. However, advanced users enjoy its powerful data management and ability to work with numerous data files at once.

Opening Stata

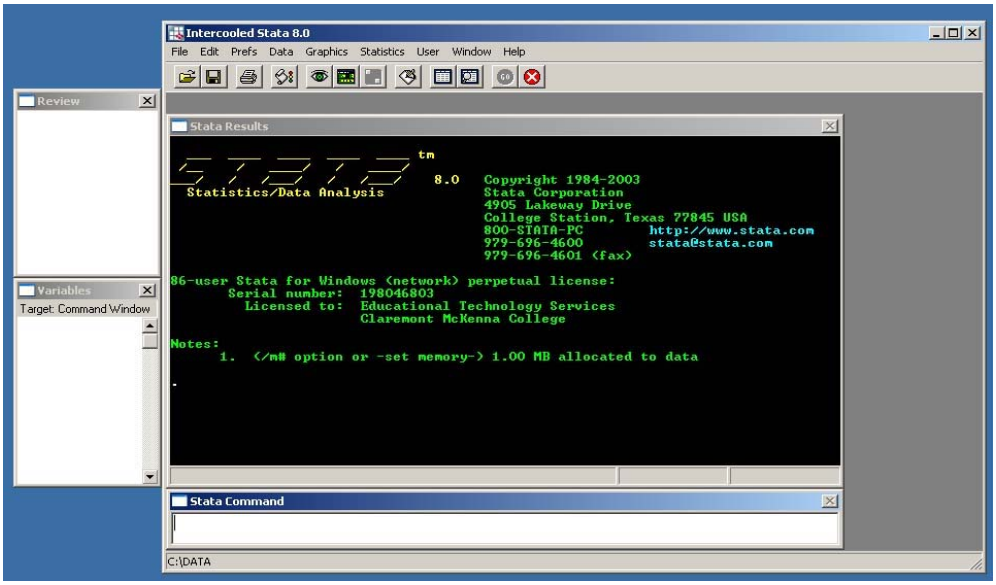
You can open Stata from the Start Menu.

1 **Start>Programs>Courseware>Intercooled Stata 8.0**



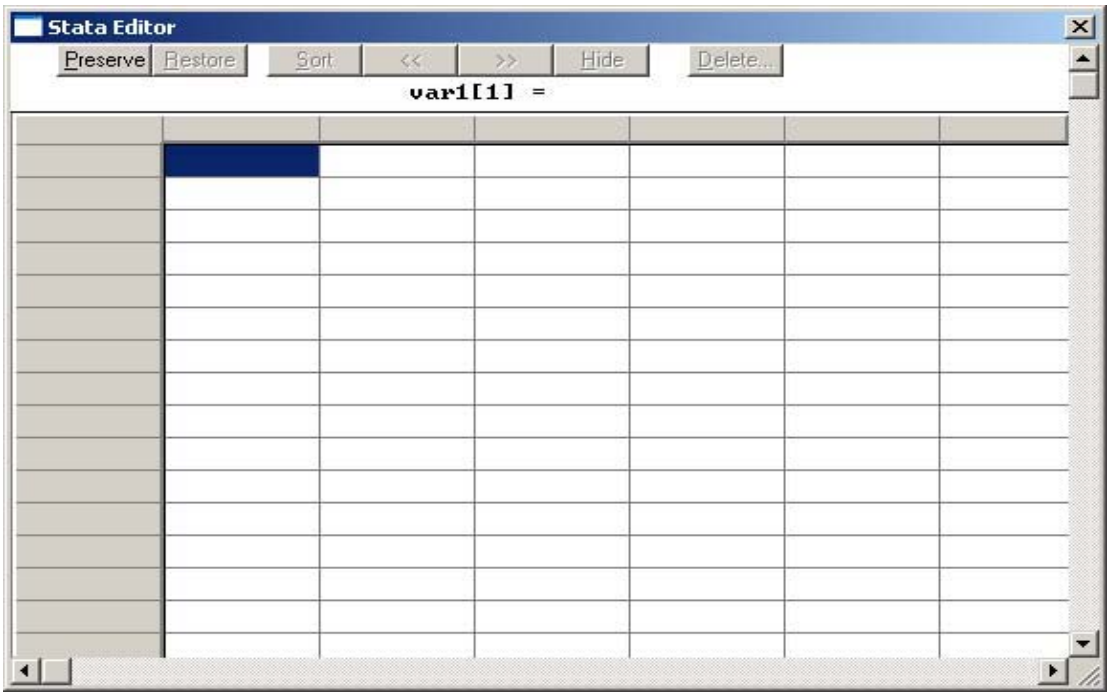
The session will open with four windows showing:

1) A results window, which displays the results of any commands you put into Stata
2) A command window, which is where all syntax commands are entered.
3) A variables window, which lists all the variables that are part of the data set with which you are currently working.
4) A review window, which details the last several commands you have entered into Stata.



How to Manually Enter Data

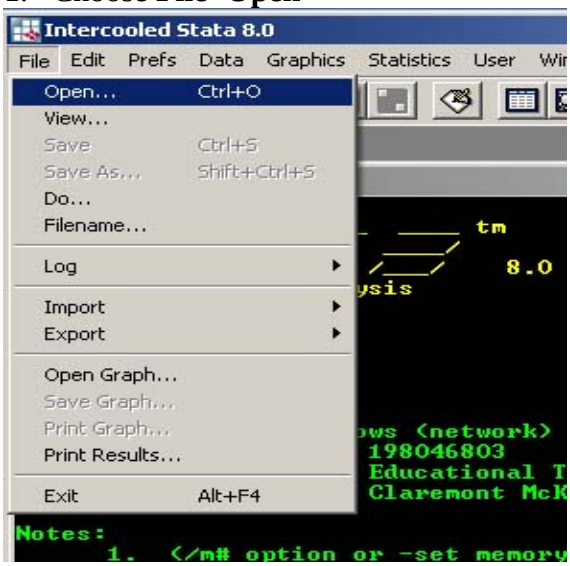
1. Opening Stata will automatically open a blank data sheet. Access this datasheet by choosing **Data>Data Editor**



***Note: When the Data Editor window is open, you lose the ability to enter commands or access most menus in Stata. You can solve this by closing the Data Editor.**

Opening an Existing Data File

1. Choose **File>Open**



Opening Data in an Excel Spreadsheet

Workbooks with more than one spreadsheet cannot be entered into Stata simultaneously. In order to enter data from multiple spreadsheets into Stata, *each spreadsheet* must be saved as a separate comma delimited file (.csv).

1. To save a spreadsheet as a comma delimited file in Excel, click **File>Save As…**

2. In the Save As Window, go to **Save as Type drop-down list**, and select CSV (Comma delimited)

3. In the Stata command window, type "**Insheet using (the path for the file you wish to use), names**". (There cannot be spaces anywhere in the path.) Below is an example.

Note: This command will only work if the top row in the spreadsheet is the names of the data in each column. If this is not the case, a more complex command must be entered.

Command Structure of Stata

Stata is primarily used through the entering of commands. The basic structure for the commands is as follows:

Command variable(s), options where

Command tells Stata what operation you want to execute

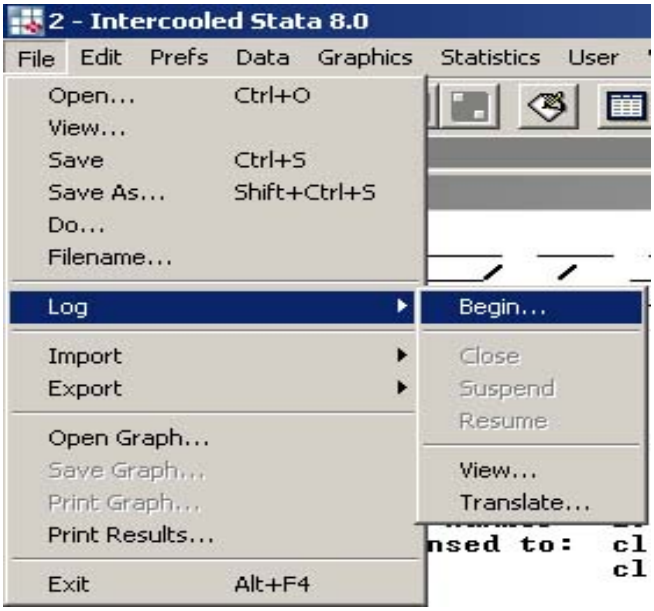
Variable(s) is (are) the variable(s) used to perform the command

Options tells Stata in what way you want to perform the command

Opening a Log File

If you are working on thesis or a long assignment, it would be wise to open a log file. This file will save both the commands you enter and the output generated by Stata. To open a log file:

1. Select **File>Log>Begin…**



2. You can close the log by selecting **File>Log>Close**.

Creating Variables as Functions of Existing Variables

The most common example of this is making a logarithmic transformation to a variable.

1. Open the Stata data file ceosal2.dta"

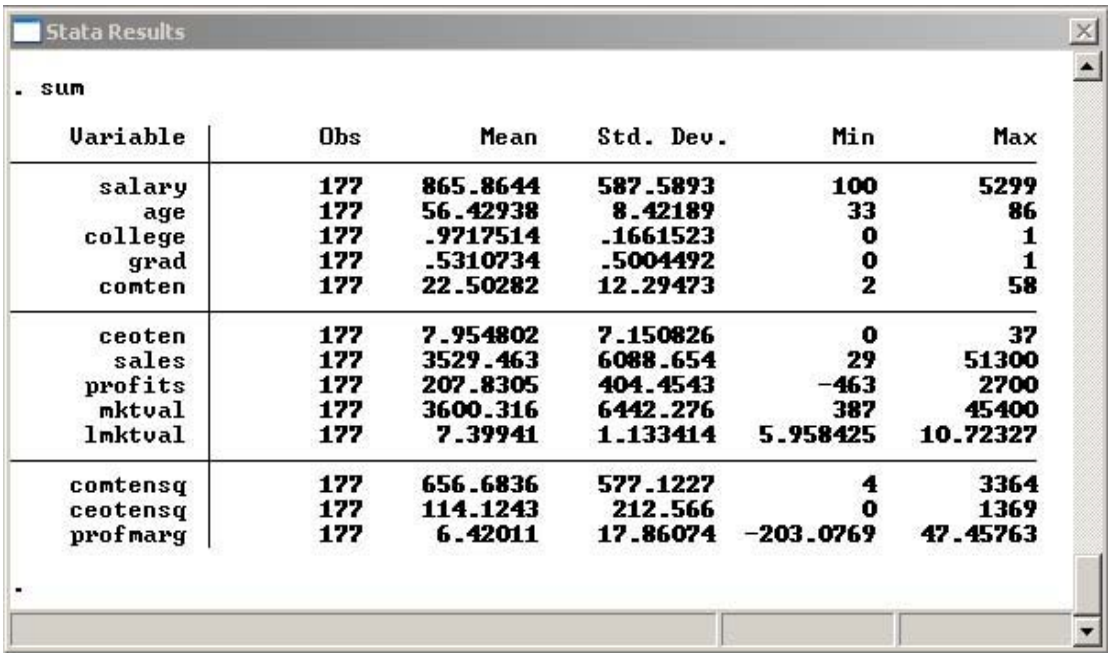
2. If you wanted to make this transformation to the variable “salary”, you would type the following into the command window:

The new variable should now appear at the bottom of the list of variables in the variable window.

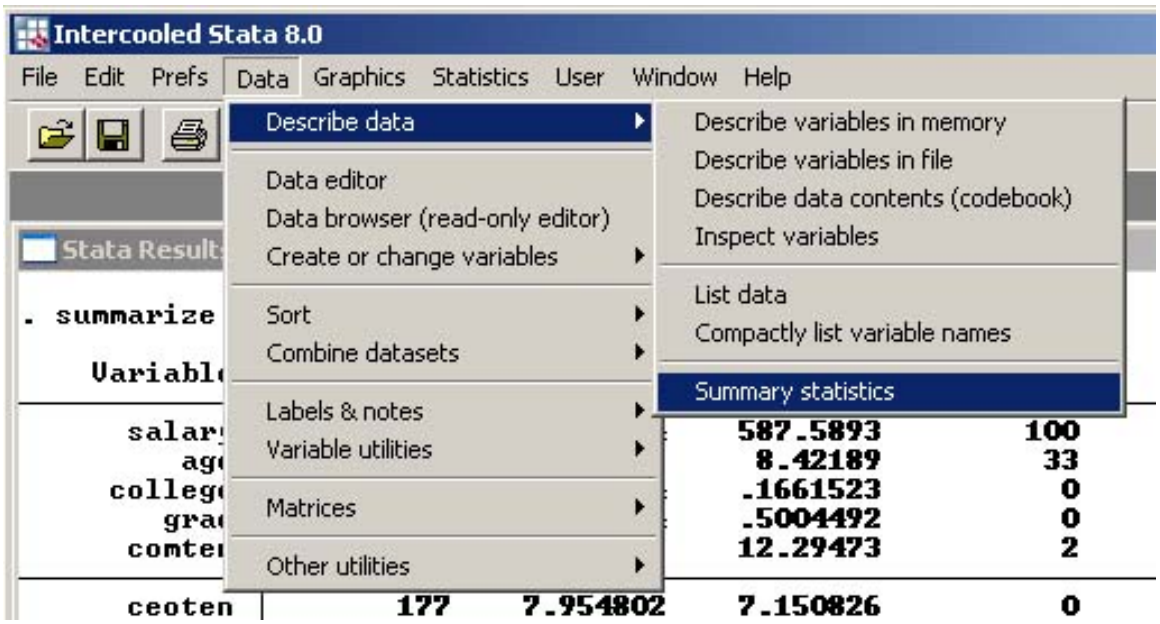
Analyzing Data

Summary Statistics

1. The **SUM** command will provide the number of observations, mean, standard deviation, and range for each variable in the data set.



2. Alternatively, you can select **Data > Describe Data > Summary Statistics**



*For more detailed statistics, including median, percentiles, variance and skewness, type, **SUM, DETAIL**.

Count

The **Count** command allows you to see how many observations meet certain conditions. Thus, if we wanted to see how many CEOs in our data had salaries of over $2 million:

1. Enter **Count if salary>2000**. Remember, we use 2000 because salary is measured in 1000s in this data set.

Note: If statements can be used with many commands in Stata, including regression.

Correlation

The **Corr** command will give you the correlation coefficient, which measures the similarity in magnitude and direction of the simultaneous changes of two variables. To find the correlation between *sales* and *profits* in the ceosal2 data set:

1. Enter **Corr sales profit**

```
. corr sales profits
(obs=177)

             |    sales  profits
-------------+------------------
       sales |   1.0000
     profits |   0.7983   1.0000
```

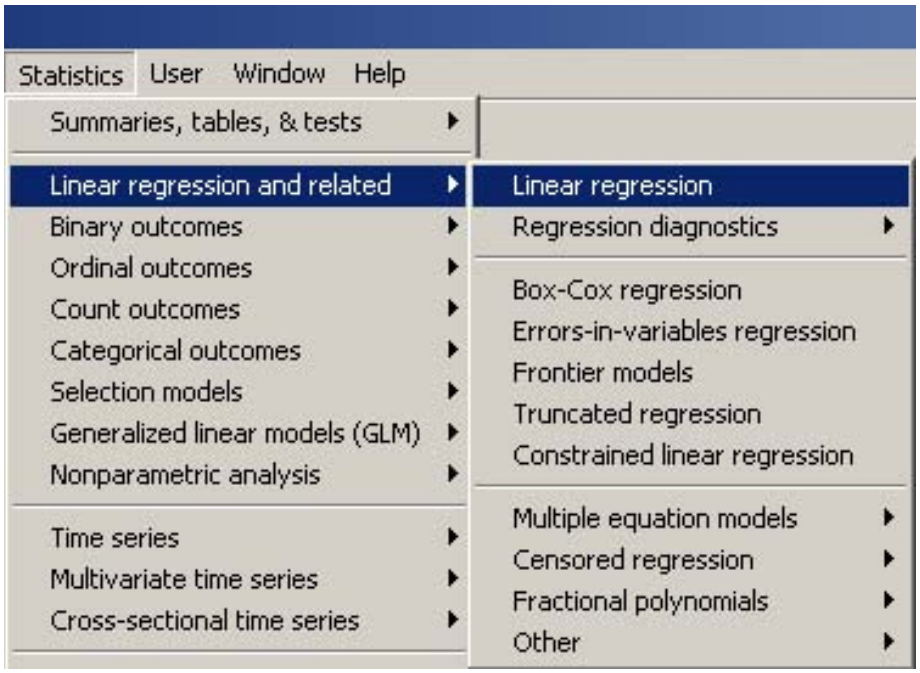
Regression

A regression attempts to explain the relationship between one or more independent variables and a dependent variable. A line (or curve) will be produced that best fits a single set of data.

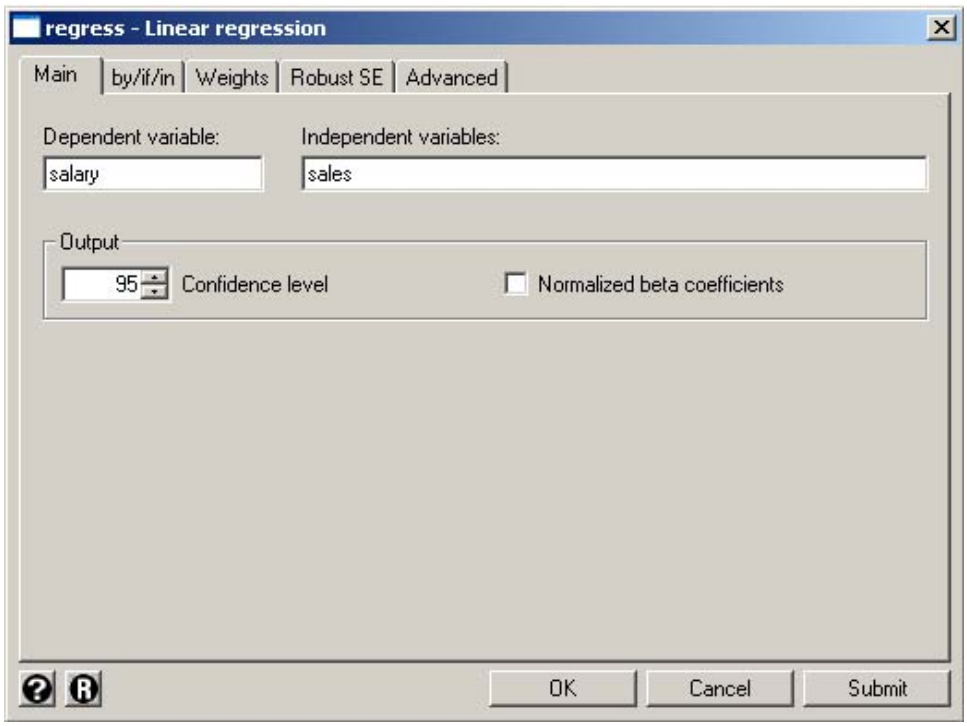
1. To examine the effect of sales on CEO salary, enter: **reg salary sales**. This will examine the impact that a change in sales has on the change in salary.

   Note: the dependent variable (the one you are attempting to measure the change in) must be listed first.

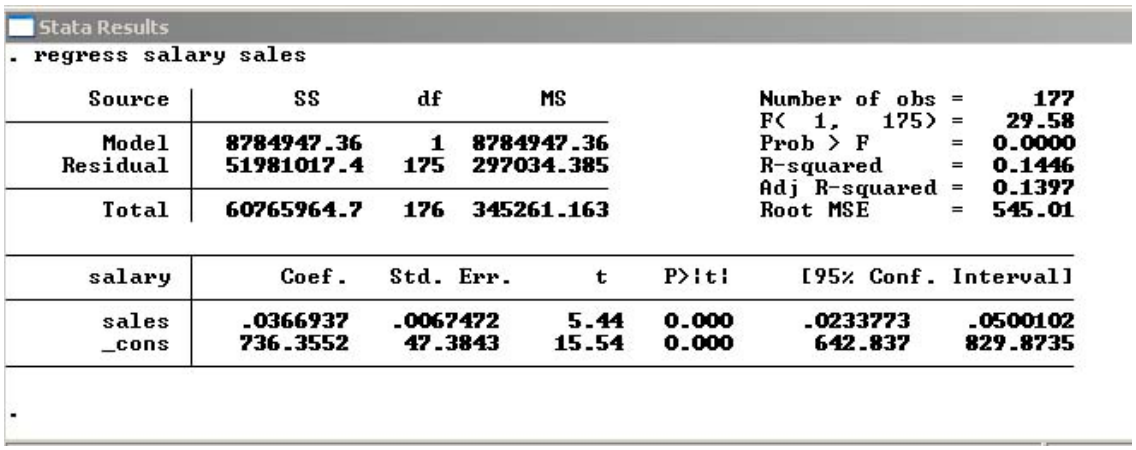
2. Alternatively, you can use the regression menu in Stata. Select **Statistics>Linear Regression and related>Linear Regression** to get to the regression input screen.



3. In the regression window, input your dependent and independent variables (in this case *salary* and *sales*.) Then click Submit.



4. After entering your regression using either of these methods, the results will appear in the Results Window.



5. You can also regress on multiple variables. In this example, let's say you want a more complete model. Therefore, you could include not only firm sales (*sales)*, but also firm profits (*profits)*, firm value *(mktval)*, tenure with the company (*comten)*, tenure as CEO of the company (*ceoten)*, age (*age)*, whether or not the CEO attended college *(college)*, and whether or not the CEO attended graduate school (*grad)*.

Saving Predicted Values from a Regression

When you run a regression, Stata will compute estimated values for the dependent variable. Sometimes, these values are needed for further analysis. To capture these values:

1. After running the regression above, enter: **predict salary_hat** where salary_hat is the new variable created using the estimated (or fitted) values from the regression. A new variable called salary_hat will now appear at the bottom of your Variables Window.

Inference

Once we have a regression, we need to know if the results we have actually mean anything—that is, could they be replicated? There are a number of significance tests available. The simplest is the T-statistic to measure the significance of a single coefficient in a regression. Another important test utilizes the F-statistic, which can test whether multiple variables are *jointly significant*.

T Tests

1. Whenever you run a regression, Stata will automatically provide the t-statistic, p-value, and confidence interval for every coefficient in the regression.

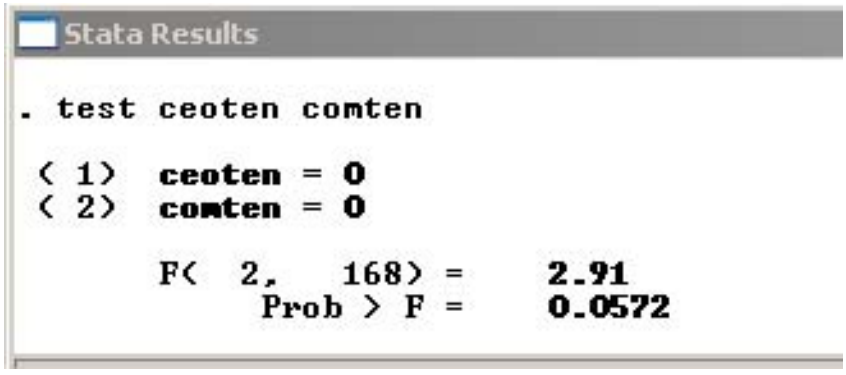
Stata Results

salary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sales	.0176717	.0111253	1.59	0.114	-.0042918	.0396351
profits	.0555287	.2756505	0.20	0.841	-.4886564	.5997138
mktval	.022517	.0159491	1.41	0.160	-.0089694	.0540035
comten	-5.321873	3.905433	-1.36	0.175	-13.03192	2.388176
ceoten	13.5699	6.164801	2.20	0.029	1.399442	25.74036
age	3.223325	5.663321	0.57	0.570	-7.957119	14.40377
college	-133.3714	250.3074	-0.53	0.595	-627.5246	360.7819
grad	-55.17274	84.71653	-0.65	0.516	-222.4189	112.0734
_cons	699.7096	400.5616	1.75	0.082	-91.07312	1490.492

F Test

1. Stata does not automatically generate F-statistics. If we wanted to test whether CEO tenure and Company tenure are jointly significant (they are not individually significant), we would enter: **test ceoten comten** immediately after running the regression listed above.

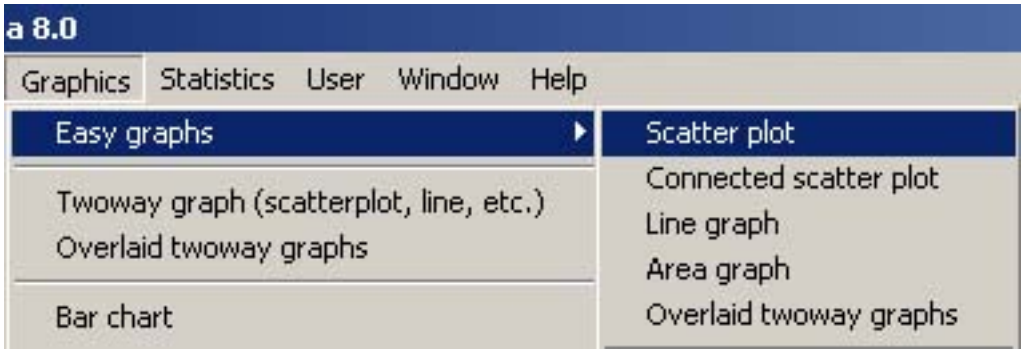
2. Stata will display an F-statistic and a p-value.



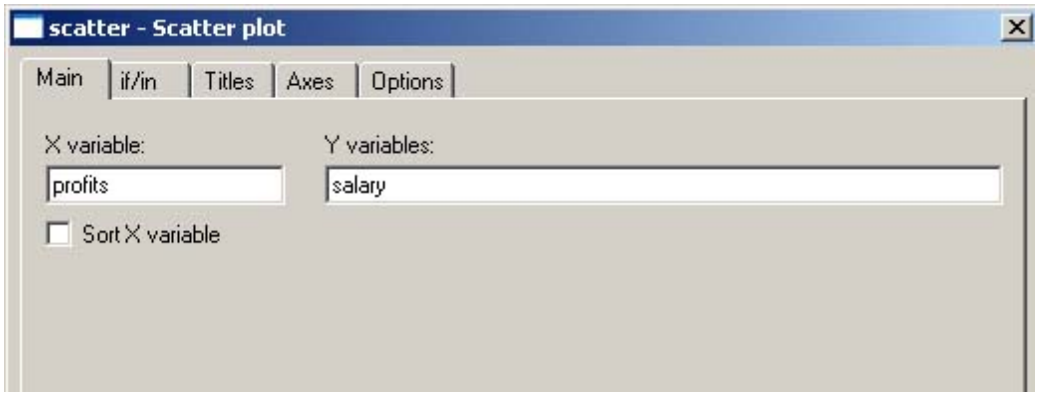
Graphing

Graphing is best completed using the point-and-click menus. Graphs can also be made via the command line, but the graphical menus in Stata (unlike many other menus) are pretty user-friendly. Stata has a vast array of graphical options. To make a scatter plot for all observations of *salary* and *profits*:

1. Select **Graphics>Easy Graphs>Scatter Plot**



2. In the Scatter Plot window, enter *profits* as the X variable and *salary* as the Y variable.



3. Then click Submit.

Using the Help Menu

The Help menu can be accessed through the point-and-click menus or through the command line. To search for help on the **reg** command:

1. Enter **help reg**

2. Stata will now produce every entry having to do with the regression command.

If you need to look for help, you should remember that many other people have probably had the problems you are having. Google searches can be extremely helpful. Many schools have posted detailed tutorials on how to certain types of analysis in Stata.

In addition, Stata provides a number of answers to questions you might have at the following address:

http://www.stata.com/support/faqs/