# Introduction to Stata

Written by John Rigg

Center for Social Science Computation & Research
145 Savery Hall
University of Washington
Seattle WA 98195 U.S.A.
(206)543-8110

April 2007
http://julius.csscr.washington.edu/pdf/stata.pdf

# WHAT IS STATA?

Stata is a command-driven statistical packages commonly used in the Social Science's. Along with SPSS and R, Stata is one of CSSCR's most popular statistical packages because of its interoperability, usability, potential for customization, and technical sophistication.

Though it isn't limited to these analyses, Stata is often used for the following:

- Panel and survey data analysis
- Discrete and limited dependent variable analysis
- Maximum likelihood estimation
- Regression analysis and regression diagnostics
- Data management and transformation.

Stata codes can also be reused and turned into routines that will help automate routine manipulations and analyses.

**What is *Introduction to Stata?***

In this hand-out, we hope to help the beginning Stata user learn to perform the following functions:

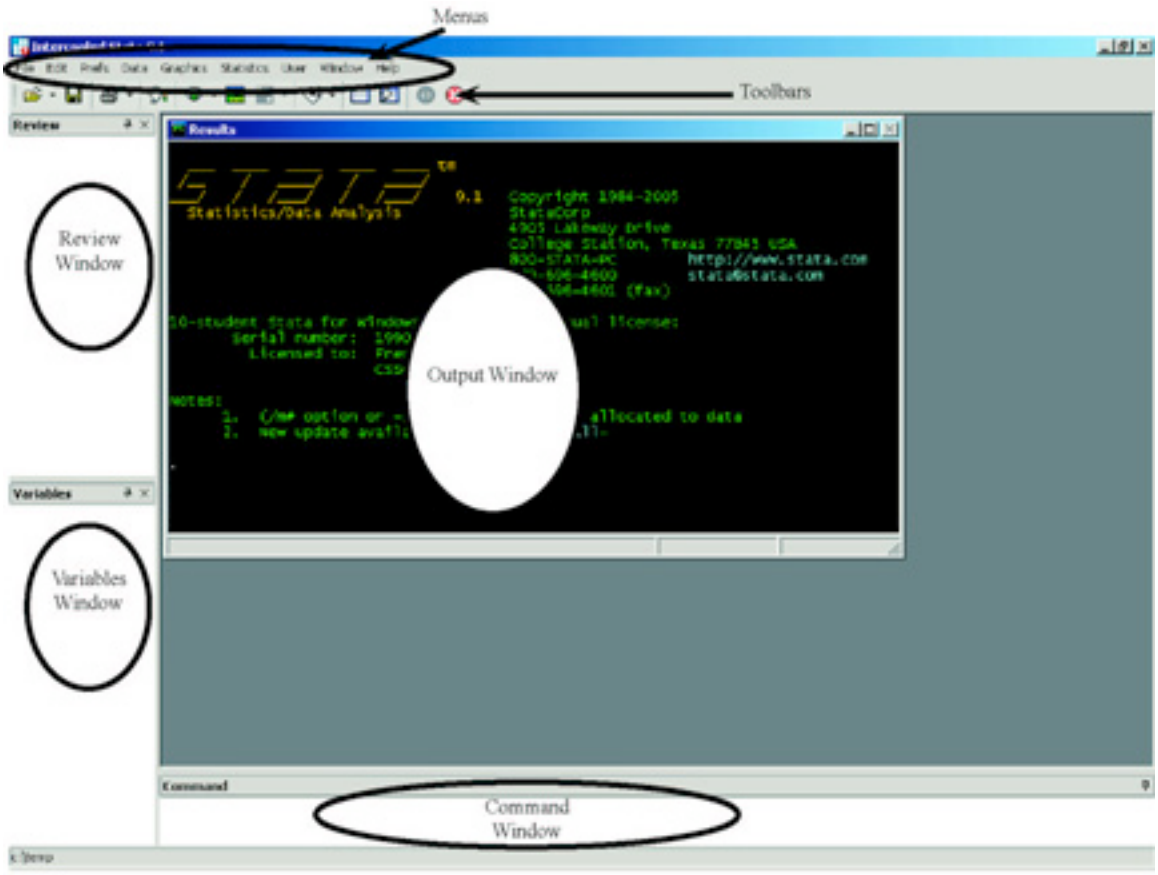
- Open a Stata data set
- Exploring data to find out more about them
- Perform simple descriptive statistics
- Plot simple graphs
- Examine simple correlation analysis
- Run linear regression models
- Perform hypothesis testing

This tutorial will guide you through the various tasks and exercises. Please note: this tutorial only works in the CSSCR labs! If you want to see FAQs, forums and links to on-line tutorials, go to http://www.stata.com/links/resources1.html

# Getting Started

This tutorial only works at CSSCR.

1. Download the tutorial data files by clicking on the start menu (in the lower left corner of the screen) and scrolling to "RUN".
2. Type the following text string: o:\classdata stata *(note the space between classdata and stata)*
3. Start Stata 9 from the desktop. This should bring up a blank Stata session that looks like this:

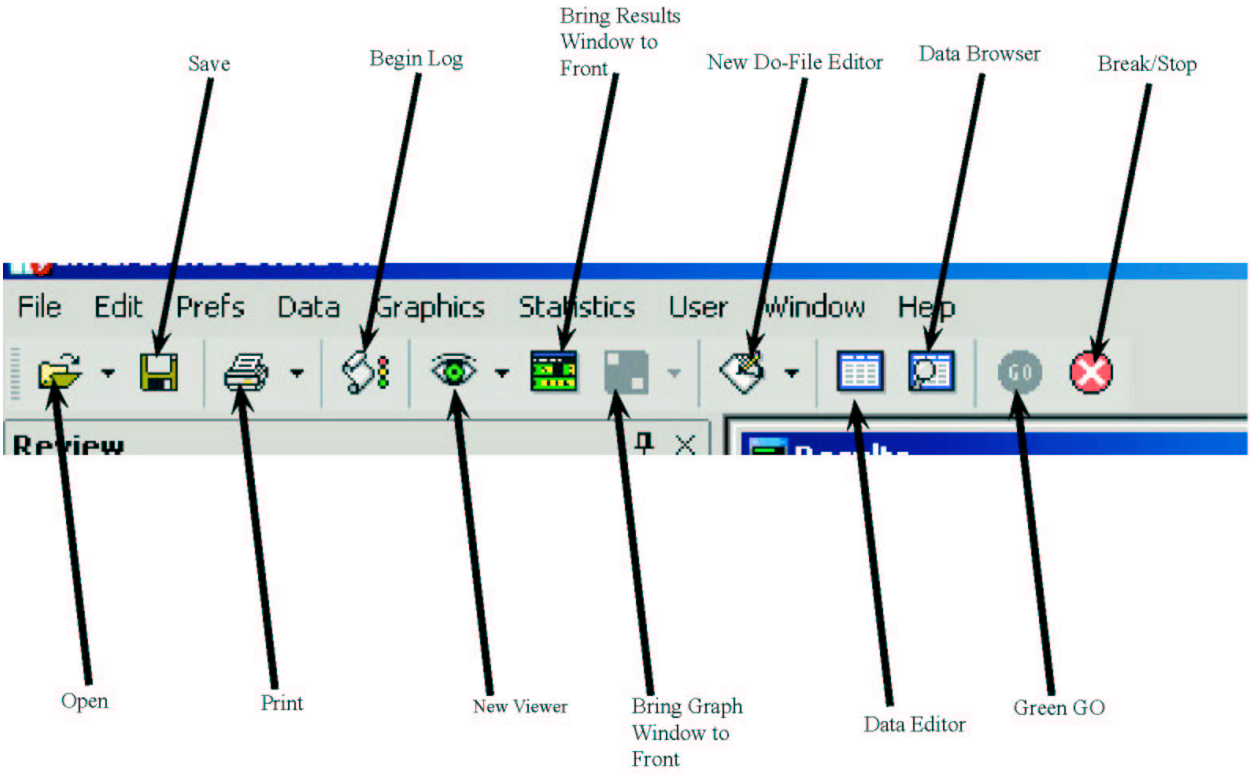



*Command Window* The command line user interface. This is where the user inputs commands for Stata. The command window is the only window that the user can type into.

*Review Window* All the commands you have entered for the current session are listed in this window. You can recall the previous command by clicking on the line in the window You can also scroll the commands you issued with Page Up or Page Down on the keyboard. No keyboard commands work in this window, but it can be manipulated with the mouse pointer.

*Variable Window* Lists all the variables contained in the dataset (in this example, we have not imported any datasets yet, so the variable list is empty). No keyboard commands work in this window, but it can be manipulated with the mouse pointer.

*Results Window* Any results or messages from Stata will be shown in this window. No commands by keyboard or mouse work in this window, though text can be highlighted and copied for use in other portions of the program.




The Stata toolbar is an important aspect of the program. The labeled picture above may not make much sense to you right now, but you can refer back to this picture to get help as we progress through the tutorial. Also, if you forget which buttons on the toolbar perform which function, you can hover your mouse pointer over it, and a box will pop up to remind you.

# Opening a Stata Data Set

When we ran the command Start➔Run➔o:\classdata stata (See *Getting started*), we downloaded a stata data set entitled *hsng.dta* into [c:\temp\stata]. This is data from the 1980 census housing data, and we will use it for the remainder of this tutorial.

To open a log file in Stata

Go to the File menu, select Log

Log➔Begin

Enter c:/temp/joemomma.smcl in the dialog box.

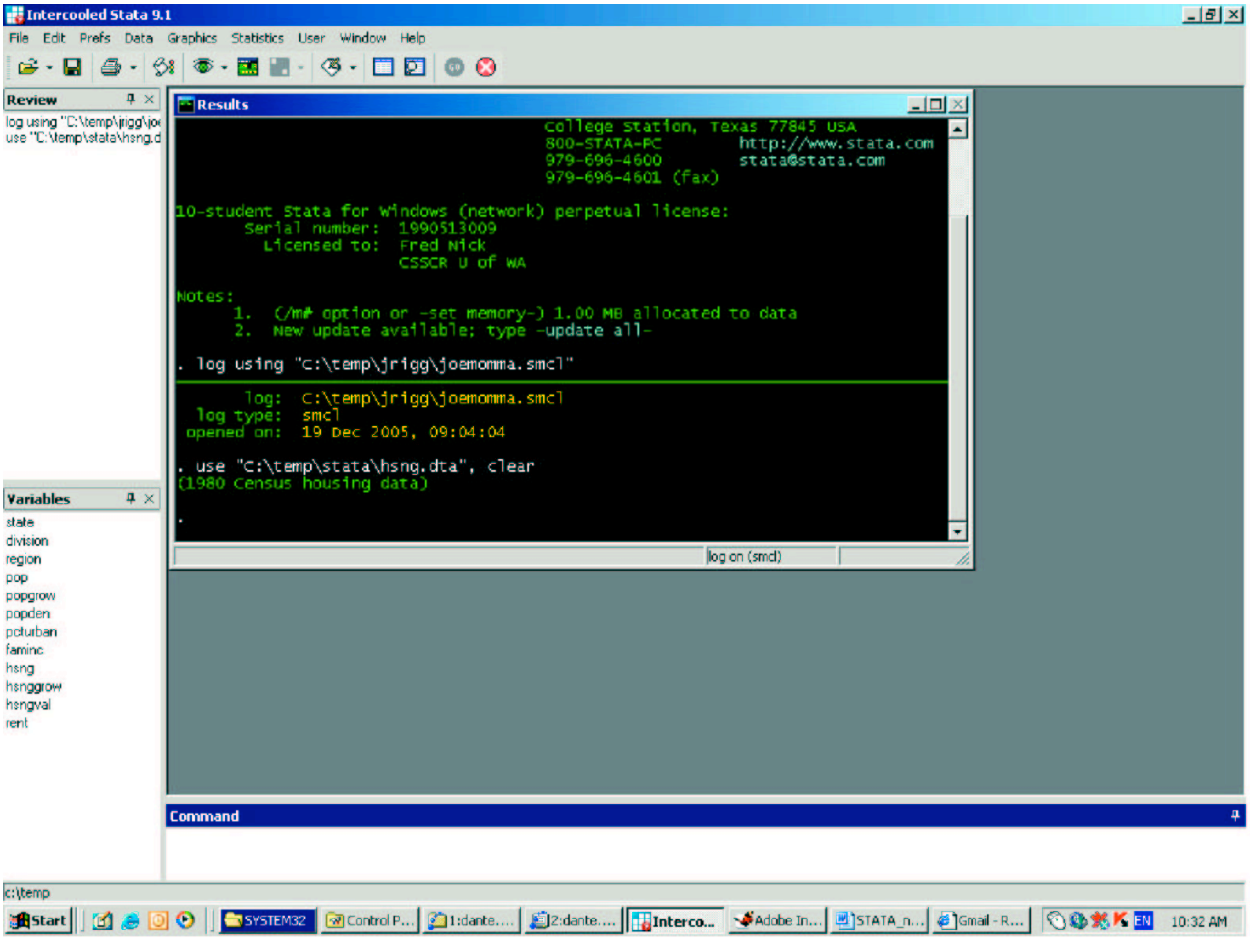
To open hsng.dta to be the active data set in Stata:

Go to the File menu, and select File➔Open

Navigate to c:\temp\stata

Double click on hsng.dta

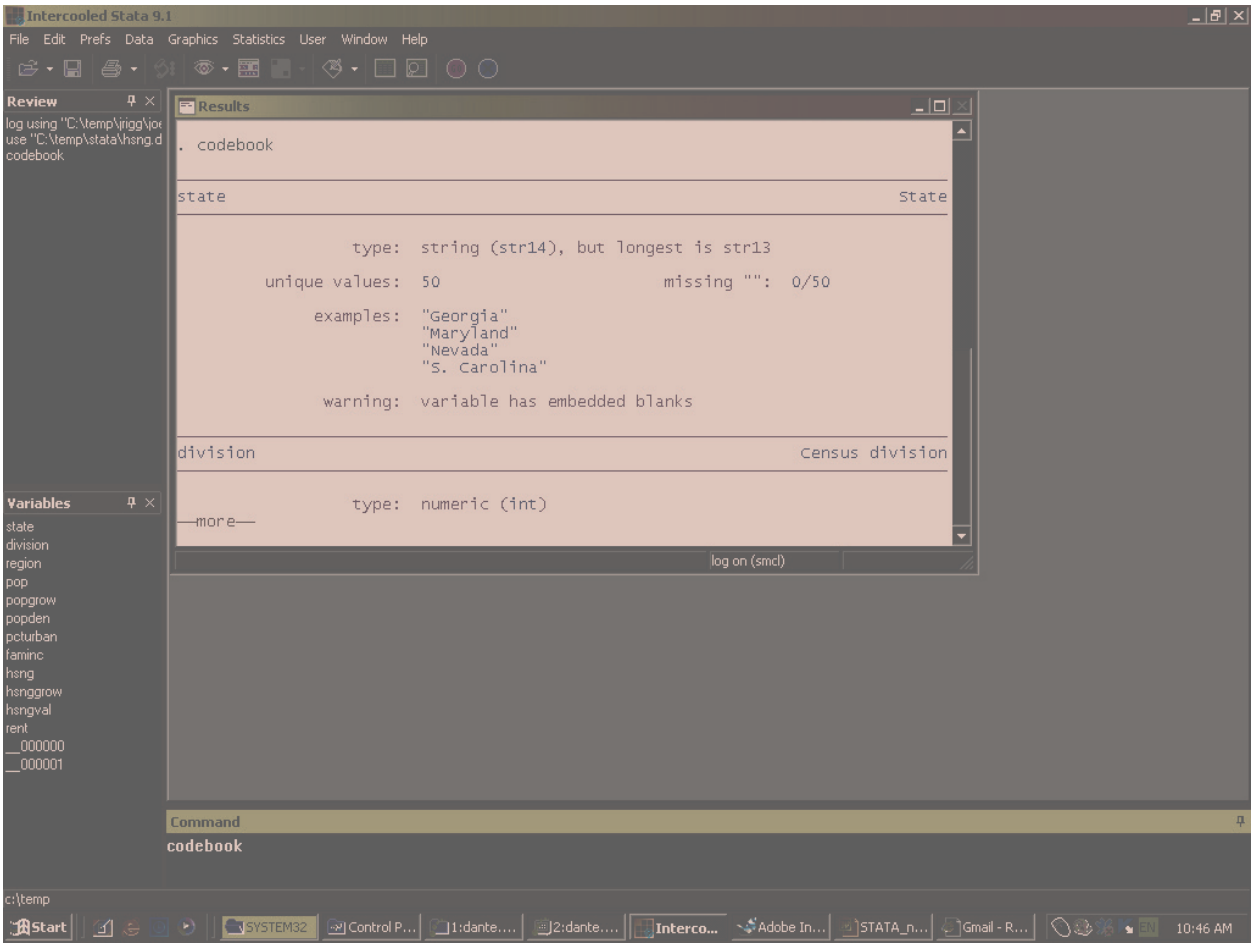
Stata should now look similar to this:




**hsing.dta** is now the active Stata data set, and **joemomma.smcl** is now the log. Any further manipulations you perform will be performed on the active set (hsing.dta), and recorded in the active log (joemomma.smcl).

# Exploring the Stata Data Set

Unless you're quite familiar with the data in the set already, it is a good idea to explore data sets with which you are unfamiliar by using the *codebook* command. Type *codebook* in the command window, and Stata will tell you more about hsing.dta:

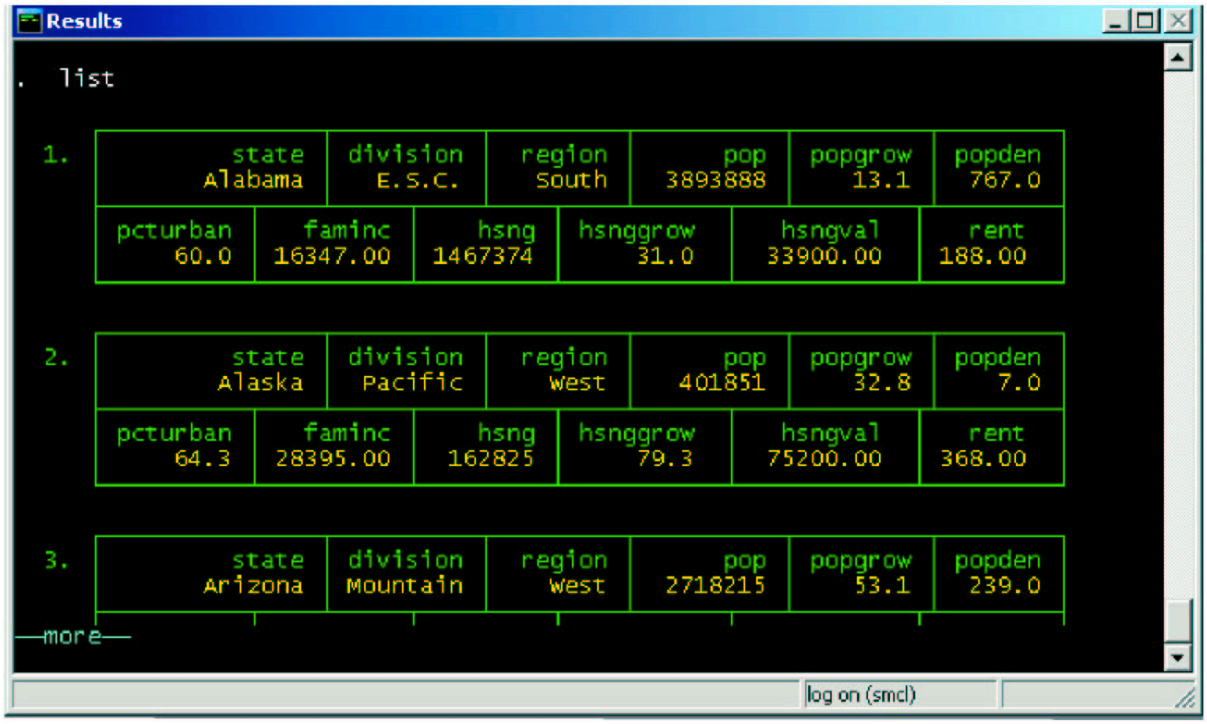



Another quick and easy exploration you can perform on data that is unfamiliar to you is the *describe* command. The *describe* command gives you information about the variables, but doesn't help you much with the individual cases. In the case of hsing.dta, the variables include *state* and *population* among others.

To use the *describe* command, simply type *describe* into the command window, or use the drop-down menus: Data➔Describe Data➔Describe variables in memory.

You can also explore the data themselves. In the case of hsing.dta, the data are by state, Alaska for example; each case is described by data in the variables. For example, Alaska is an example of a case, which is described by a name, a population, housing growth, and a host of other factors.

To use the *list* command, type list into the command window. To proceed through all of the output screens, press the spacebar each time –more– appears at the bottom of the screen.




## Wildcard Syntax for Beginners

As you explore hsng.dta, you'll quickly notice that there are considerably more variables than can be explored all at once with any precision. Using wildcards allows you to target your commands to certain variables or cases.

*Axtrix (*) wildcard:*

selects all variables of a similar type to whatever you have before or after the astrix.

**Example 1:** if you enter *list pop** Stata will return information on all variables with the text string *pop* in the beginning. In this case, it includes variables pop, popgrow, and popden.

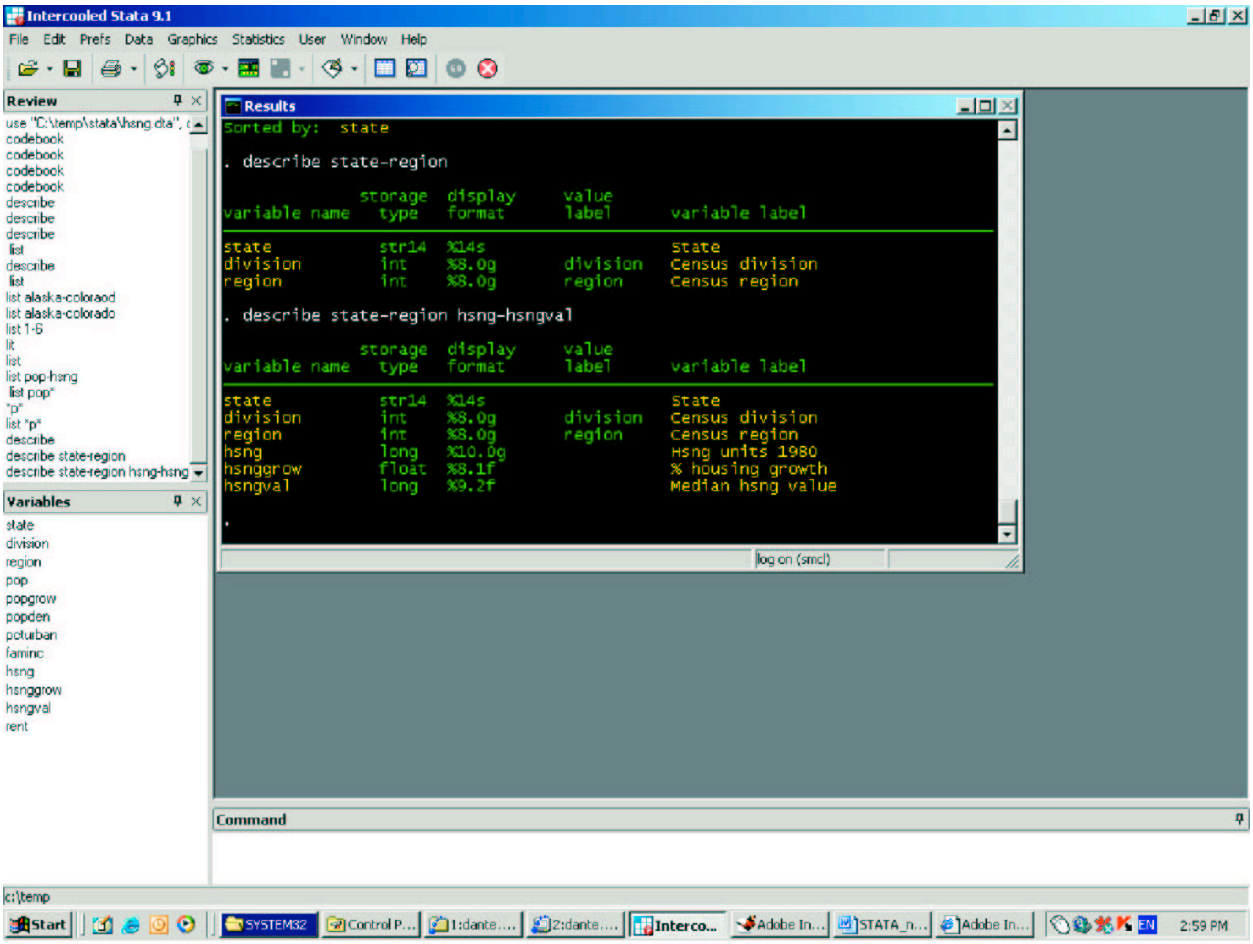
**Example 2:** if you enter *list *p** Stata will return all variables with p in the name of the variable. In this case, pop, popgrow, popden, and pcturban.

*Hyphen (-) wildcard:*

- selects all variables in a sequence. In hsng.dta for example, the first variable is *state* followed by *division* and ending in *rent*.

**Example 1:** entering *describe state-region* returns a description of three variables: *state*, *division*, and *region.*

**Example 2:** entering *describe state-region hsng-hsngval* describes the variables between state and region, and those between hsng and hsngval, while not including descriptions of those in between.



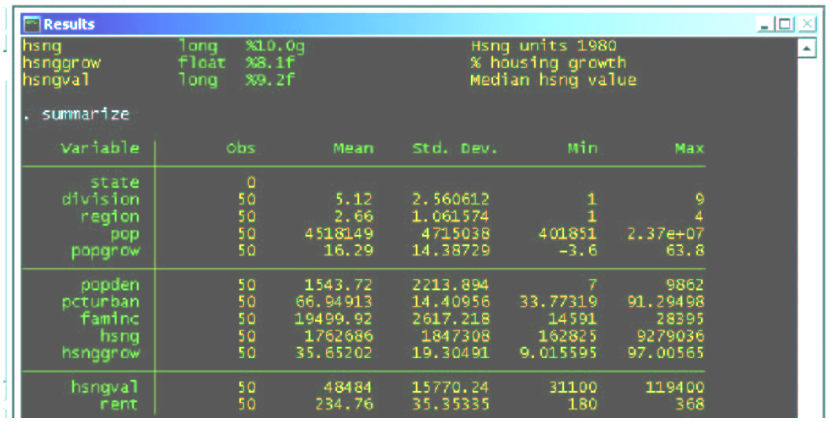

## Results Window Manipulation

You'll notice that output display one screen at a time. To proceed to the next screen, press spacebar, or use the green *GO* button on the toolbar.

On the other hand, if you wanted to stop the codebook from being displayed in the results window, note that the *break* button is no longer grayed-out, and will terminate the codebook command if you press it.

## Performing Descriptive Statistics

Now that you know a little bit more about the data, you can run some descriptives. The command for descriptive statistics is *summarize*. You can either summarize one or some of the variables using the procedures described in the wildcard syntax section above, or describe all variables by simply entering *summarize*.

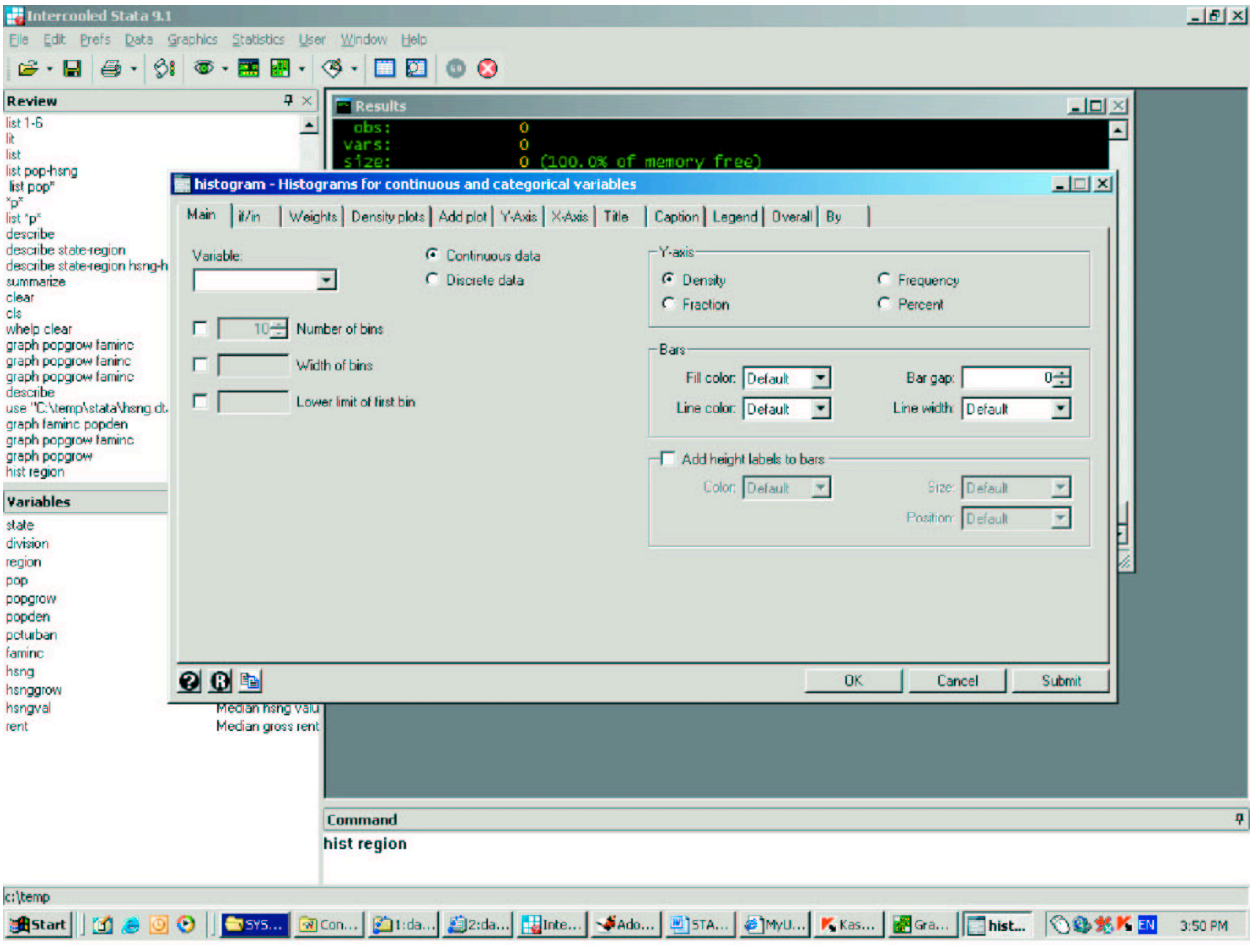


## Graphing in Stata

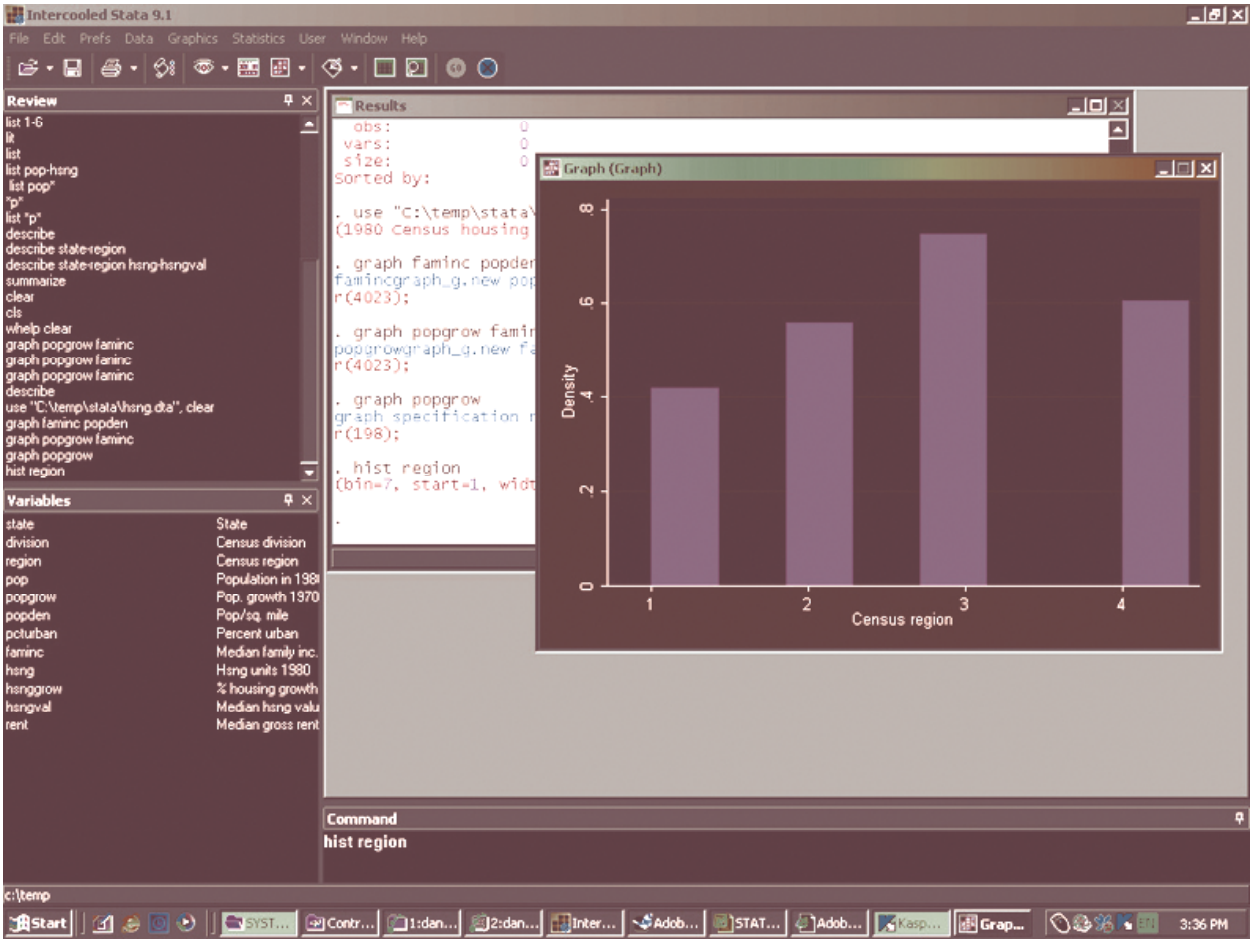
Stata 9 has a graphing feature that creates a separate window for graph output, rather than graphing it in the output window. For our first graph, we will create a histogram of the variable *region*. There are two ways to do this; either by entering *hist region* in the command window, or by using the menus to select Graphics➔Histogram.

The first method, entering the *hist region* command in the command window is the simplest, but it doesn't allow for customization.

The second method, by using the menus to select Graphics➔Histogram is more complex in that it allows you to customize your graph to an extent that you cannot by simply entering the command line. In this menu, the variable you want to run the histogram on is selected by the drop-down menu, and other aspects of the graph can be changed by selecting the appropriate tab and making changes. It is outside the scope of this pamphlet to describe these features in detail, but experimentation is an easy (and fun) way to learn.




This is the histogram tab user interface. This interface and others like it in other graphing commands allow you to change the way the graph appears.



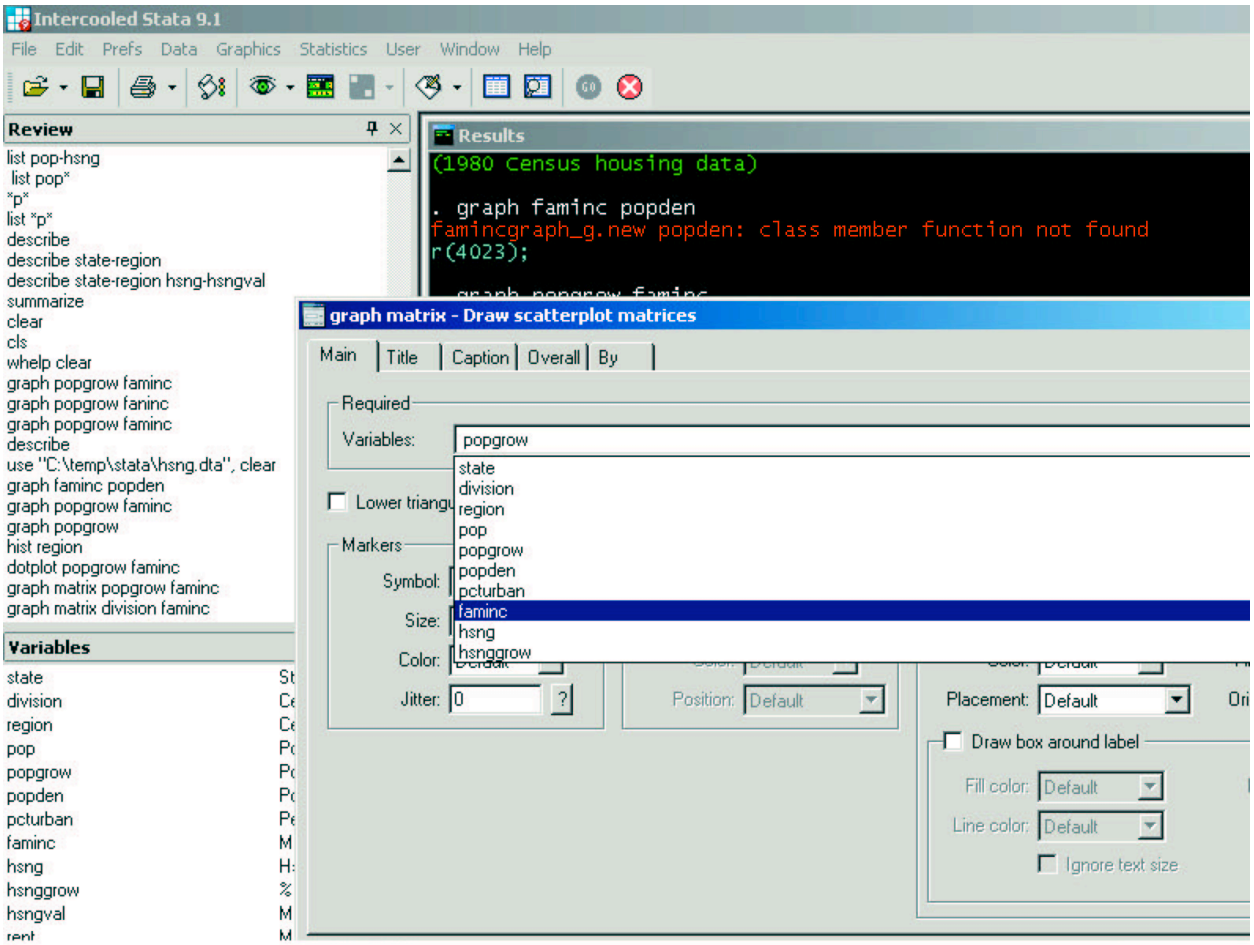

The histogram on the variable *region*, as drawn by Stata.

Other graphs can be created in a similar fashion to the histogram. Either use the drop-down menus to select the graph you want, or enter the corresponding command in the command window.

For example, to create a matrix scatter plot on variables *popgrow* and *faminc*, either type *graph matrix popgrow faminc* or go up to the menus and select Graphics➔Scatterplot Matrix.

You'll note that the scatter plot matrix is different than the histogram graph that we did previously, because it graphs two variables on each other instead of the characteristics of just one graph, as histogram did.

As such, you have to select two variable names from the drop-down menu in the variable box if you choose to use the pop-up menu option.




## Copying Graphs to Other Programs for Publication

To copy a graph to another program (such as Microsoft Word) for publication, right-click on the graph popup window, and select "copy". You can then paste it in to the other program through the Edit popup menu.

# Statistics: Correlations, Regressions and Hypothesis Testing

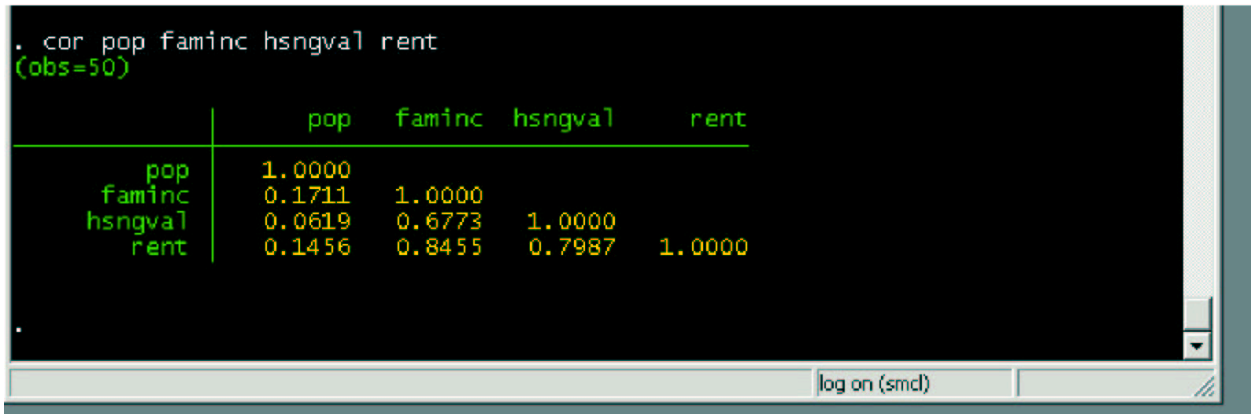
## Correlations

Pearson's correlations are easy to calculate in Stata. There are two methods:

Type *cor pop faminc hsngval rent* for correlation among these variables; or

Use the drop-down menu: Statistics➔Summaries, Tables & Tests➔Summary Statistics➔Correlations & Covariences. If you use the menu method, keep in mind that you must select several variables from the drop-down menu.

Regardless, the output should look like this:




## Regressions

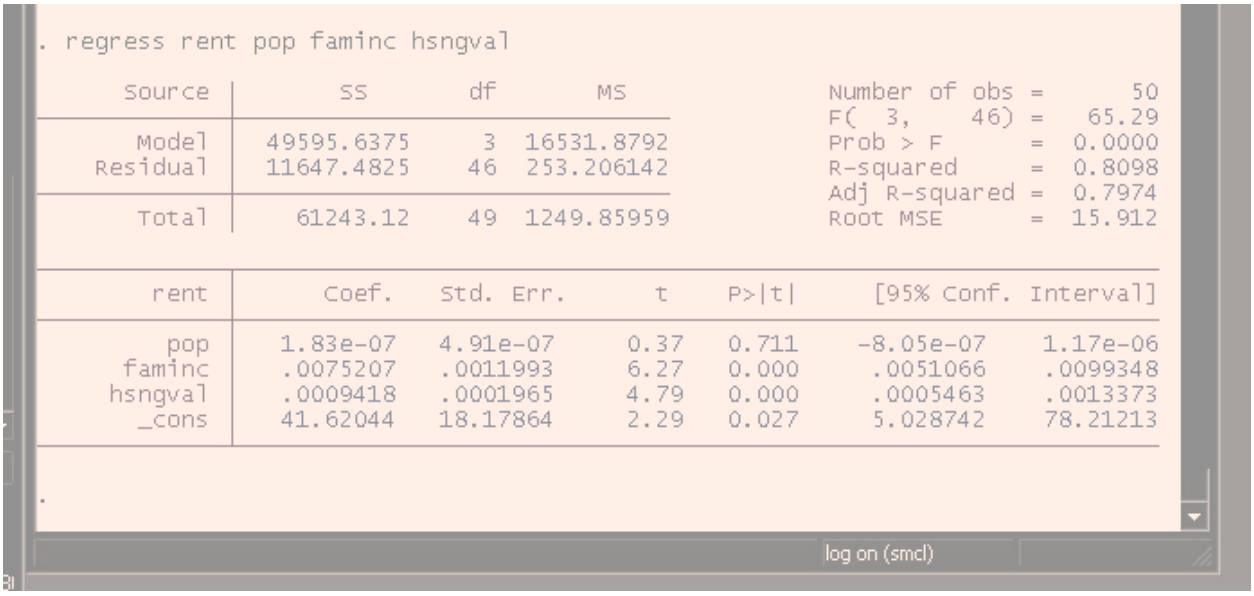
Simple linear regressions are easy to run in Stata. For this example, we will be examining how rent is affected by population, family income, and housing unit value. As when we ran correlations above, there are two ways of doing this: through the command window, or through the drop-down menus:

Type regress rent pop faminc hsngval in the command window; or

Use the drop-down menus to select Statistics➔Linear models and related➔Linear regression.

Remember that linear regressions use one dependent variable with two or more independent (explanatory) variables. In the command line entry, the first variable name is always dependent and subsequent variables are independent. In the drop-down menu option, you must select them by hand.

Regardless of the method you used to derive it, your regression results should look like this:



```
. regress rent pop faminc hsngval

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  3,    46) =   65.29
       Model |  49595.6375     3  16531.8792            Prob > F      =  0.0000
    Residual |  11647.4825    46  253.206142            R-squared     =  0.8098
-------------+------------------------------           Adj R-squared =  0.7974
       Total |   61243.12     49  1249.85959            Root MSE      =  15.912

------------------------------------------------------------------------------
        rent |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         pop |   1.83e-07   4.91e-07     0.37   0.711    -8.05e-07    1.17e-06
      faminc |   .0075207   .0011993     6.27   0.000     .0051066    .0099348
     hsngval |   .0009418   .0001965     4.79   0.000     .0005463    .0013373
       _cons |   41.62044   18.17864     2.29   0.027     5.028742    78.21213
------------------------------------------------------------------------------

.

log on (smcl)
```

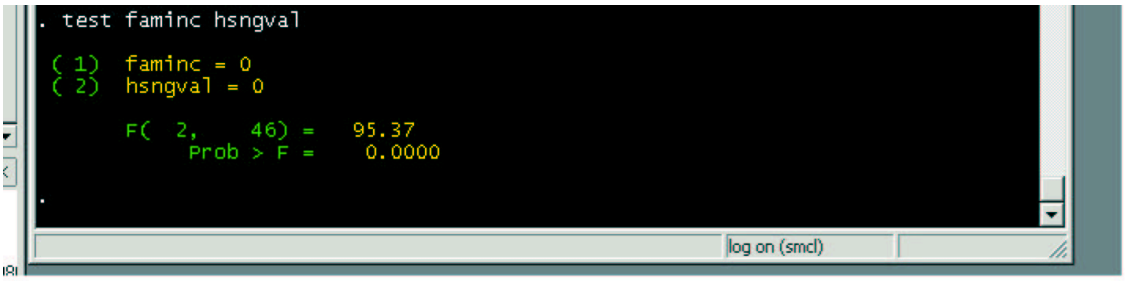
## Hypothesis Testing

Finally, if you wish to do some hypothesis testing on coefficients derived from your regression results, you can do so using either the command window or the drop-down method:

Type *test faminc hsngva*; or

Use the drop-down menus to navigate to:
Statistics➔Linear models and related➔ANOVA➔Test linear hypothesis after anova

Your output should look like this:



```
. test faminc hsngval

 ( 1)  faminc = 0
 ( 2)  hsngval = 0

       F(  2,    46) =   95.37
            Prob > F =    0.0000

.

log on (smcl)
```

Note that these tests can only be performed after a regression has been run.

## Quitting Stata

To end your Stata session, it is always a good idea to save your log file by entering *log close*. Stata will acknowledge that the log is closed. Once closed, this log will serve as a record of your work for the day.

## Getting Stata Help

Stata has a robust help program. It can be accessed either through the command window by typing *whelp* followed by the command you need help with, or through the drop-down *Help* menu.

Appendix 2 has a list of helpful Stata commands that can be entered from the command window; alternately, you may find them through the appropriate drop-down menus.

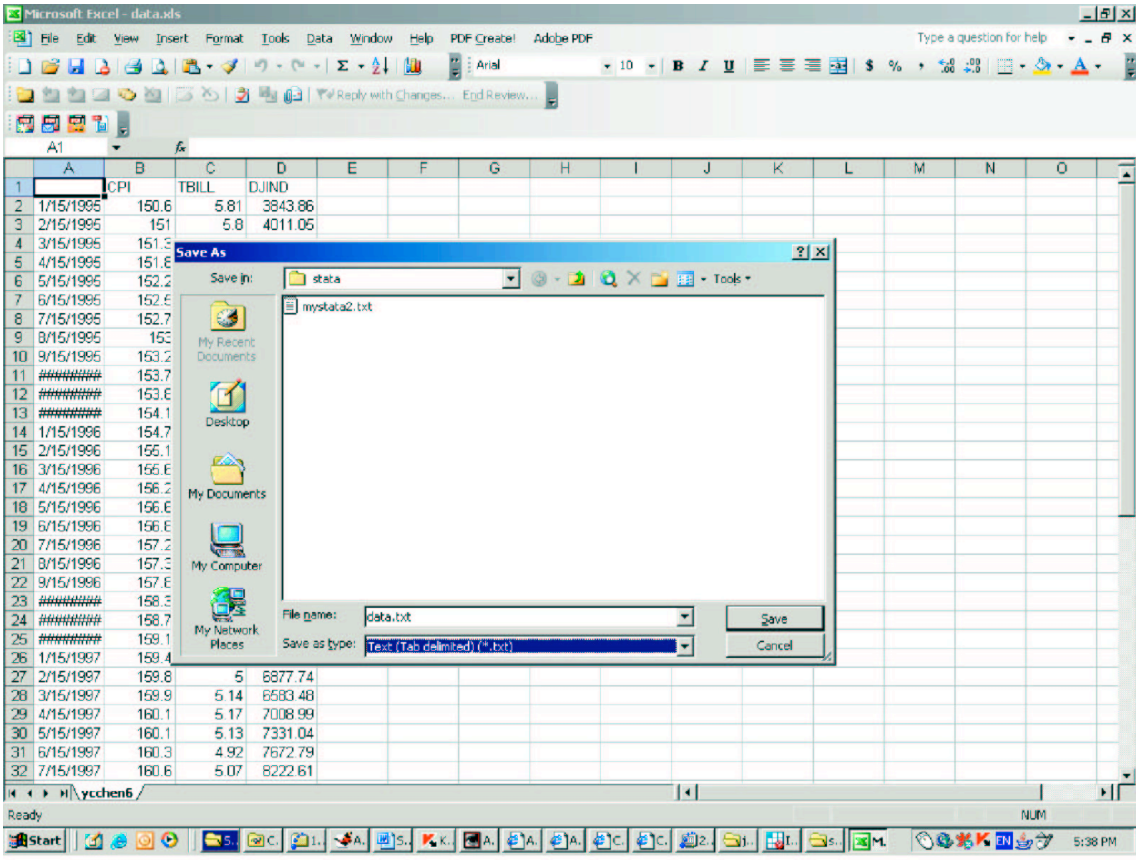
Finally, CSSCR has consultants with Stata knowledge and, if needed, a full complement of Stata manuals that are available for check-out.

# Appendix 1: Converting Data to Stata

It's not uncommon to find data that isn't in Stata when it is first found. Fortunately, there is an easy way to convert data into Stata, if it is in tab-delimited format.

For our example, use the data entitled *data.xls* which is in the c:\temp\stata directory on your CSSCR computer:

1. Press Start➔Run➔c:\temp\stata
2. Double click on *data.xls* to open it in Excel
3. In Excel, use the drop-down menu to select File➔Save As
4. Select either .txt (tab delimited), or .csv (comma delimited)
5. Select the directory you want to save it to and press save.




It is extremely important to remember the path, because from Stata, you have to enter the proper path in the *insheet* command in the command window. For this example, I took an excel spreadsheet in the directory *c:/temp/stata* called *data.xls* and saved it to the same directory as: *c:/temp/stata/text.txt*

To bring it into Stata, use the *insheet* command, and type:

```
Insheet "c:\temp\stata\text.txt",clear names
```

This tells Stata that you are bringing a spreadsheet (in tab-delimited format) into Stata for use as the Stata sheet. From now on, when you save the data, it will save as a Stata data (.dta) file.

*Happy Fun Shortcut*
Talk to a CSSCR consultant, and tell them to convert the data to Stata using StatTransfer, a program they have access to at CSSCR.

# Appendix 2: Some Additional Useful Commands

**(For use from the command window)**

Below you will find some additional commands that many beginning Stata users find useful. In some of them, there is syntax that is missing, so if the command doesn't work the first time type *whelp COMMAND* where COMMAND is the command you need additional help with.

You'll note that these commands don't have corresponding drop-down menus. This is to provoke the user to use the command line more often.

**Reading and Writing Data**

| *Command* | *Description* |
|---|---|
| use filename | Open a Stata file (if including directory use use quote marks as in "c:/temp/data.dta" |
| save filename | Save as a Stata file |
| compress | Compress a Stata data file |
| insheet using filename | Read ASCII data in a tab delimited format |

**Editing Data:**

| *Command* | *Description* |
|---|---|
| generate newvar = expression | Create a new variable named newvar |
| replace oldvar = newvalue | Assign a new value of newvalue to the variable oldvar |
| recode varname | Change the values of the variable varname |
| reshape | Convert the data between wide and long formats |

**Descriptive Procedures**

| *Command* | *Description* |
|---|---|
| describe | Display the properties of variables |
| ds | Compact version of describe command |
| list | List observations for the variables |
| summarize | Calculate major descriptive statistics |
| inspect | Display more information on variables |
| by catvar: summarize | Summarize data by categorical variable catvar |
| tabulate | Create one- or two-way tables |
| table | Create multi-way cross-tables |
| correlate | Calculate Pearson's correlations |

**Statistical Procedures**

| *Command* | *Description* |
|---|---|
| regress y x1 x2 x3 | Run linear regression (enter dependent variable y, followed by independent variables x1, x2 and x3, by default the constant is included) |
| rreg | Robust regression |
| anova | Perform ANOVA analysis |
| logit | Logit analysis |
| probit | Probit Analysis |
| glm | Generalized linear models |
| predict | Calculate predictions or residuals after estimation |
| test | Coefficient test for the last estimated model |
| ttest | Perform t tests on equality of means |
| lrtest | Perform Likelihood-ratio tests after estimation |

**Graphics**

| *Command* | *Description* |
|---|---|
| histogram var | Create a histogram for var if var is continuous |
| histogram var, discrete | Create a histogram for var if var is discrete |
| graph matrix y x | Draw a scatterplot of y against x |
| graph matrix y x1 x2 x3, | Create two-way scatterplot matrix for y, x1, x2 and x3 |
| graph box var | Box plot |

Help / Search

| Command | Description |
|---|---|
| help/whelp stata_command | Display the description and the syntax on Stata commands (use help or whelp when you know the name of the Stata command) |
| search topic | Find the Stata command you are looking for (use search when you don't know the specific Stata command) |
| lookfor string | Search for string that contains in the labels of or the names of variables. |

**Miscellaneous Commands**

| *Command* | *Description* |
|---|---|
| sort var | Sort the data according to the variable specified |
| rename oldname newname | Replace the existing variable name oldname with newname |
| order/move | Change the order of the variables listed in the dataset. |
| drop | Remove variables or observations from the memory. |
| keep | Keep the variables or observations. |
| label | Create labels for the variables. |
| do filename | Run a do file |
| exit | Quit Stata (no changes have made in the data set). |
| exit, clear | Quit Stata without saving changes on the data set. |
| Log using filename | Open/create a log file for the current session. |