

Econ 527: Stata Basics

Haimin Zhang

Introduction

Stata is one of the most popular statistical packages (other popular ones include SAS and SPSS). It is commonly used by economist for its powerful regression tools, up-to-date statistical models, and ability to produce publication-quality graphs. You can also easily find many procedures written by researchers for various tasks. This tutorial introduces the very basic features of Stata.

Syntax

The syntax of Stata is very straightforward. Notice that Stata is case sensitive. Things in [] are optional.

`[by varlist:] command [varlist] [=exp] [if exp] [in range] [weight] [using filename] [, options]`

command: It is the only required element, which is usually (but not always) an action verb, and is often followed by the names of one or more variables. Commands can usually be abbreviated. Usually in the help file, the letters that are underlined of a command are required. For example: regress indicates that the "regress" command can be abbreviated to "reg".

=exp: Commands used to generate new variables.

if exp and **in range:** They are used to restrict the command to a subset. "if" is used for logical condition. "in" is used for specify a range, for example in 1/10 will restrict the command's action to the first 10 observations. Type "help numlist" to learn more about lists of numbers.

weight: Some commands allow the use of weights, type help weights to learn more.

using filename: The keyword using introduces a file name; this can be a file in your computer, on the network, or on the internet.

options: Most commands have options that are specified following a comma. It is usually helpful to type "help command" to obtain the full list of available options.

by varlist: It is a very powerful feature. It tells Stata to repeat the command for each group of observations defined by distinct values of the variables in the list.

Load Data

Just as most of other Stata commands, you can load data by typing in the command in the command window, or you can simply use mouse to select the task you want to perform. For different type of data, the command is different.

`.dta` file is Stata dataset format, therefore it can be directly opened in Stata. For example, if the full location of your file is "E:\Users\Haimin\Documents\Research\example.dta", type

```
use "E:\Users\Haimin\Documents\Research\example.dta" , clear
```

Or you can simply click "File → Open...", then choose the `.dta` file you want to open.

However, when you obtain data from different sources, they don't usually come in Stata format. The most common type is plain `.txt` file, or comma separated format `.csv`. Sometime it also comes as Excel format `.xls`. My personal favorite way of doing it is by the help of Excel. After converting the data in Excel into `.csv` file, type

```
insheet using "E:\Users\Haimin\Documents\Research\example.csv" , comma
```

Or you can click "File → Import → ASCII data created by a spreadsheet", and then select the file you want to import. This way can apply to importing `.txt` file, `.raw` file as well.

For Stata 11, which is the one installed at the Arts Computer Lab (Buchanan B) as well as Econ Graduate Lab (Buchanan Tower 10th floor), there is even an easier way: open the Data Editor in Stata (you can click the icon on the menu bar, or click "Window → Data Editor"), then copy and paste the data you want from Excel.

Useful commands:

`clear`

This command is used to clear the memory, can be dataset, matrix. You have to clear the memory before loading in a new dataset.

`set memory`

It is usually used before loading data. For really big dataset, you would want to allocate more memory to Stata.

`drop` and `keep`

They both are used to eliminate variables or observations. It can be followed by **if**, **in**.

Manipulate Data and See the Pattern

`generate`

Generate new variables flowing the desired expression. +, -, *, /, ^, sqrt()

egen

It is the extension of generate command. It can generate new variables using functions in Stata. For full list, type “help egen”.

replace

Replace the content of existing variables.

tabulate

It is used to describe the frequency of each value for given variables. It could be one way or two way. One way tabulate is very useful in generating dummy variables. For example if you have a variable named “gender”, and value 1 indicates female, value 2 indicates male, you can type

```
tab gender, gen(male)
```

You’ll get two dummy variables indicating female or male status.

summarize

It is used to generate summary statistics. It gives number of observation, mean, standard deviation, min and max.

Logical Operators

& (and) |(or) !(not) ~(not)

> (greater than) < (less than) >= (greater than or equal) <= (less than or equal)

== (equal) != (not equal) ~= (not equal)

Notice that the logical operator “==” is different from the expression “=”. For example, if you want to generate a new variable only for male, suppose you have a dummy variable named “male” that takes value 1 for male, you type (pay attention to the number of equal sign)

```
gen variable=3^2/15 if male==1
```

Analyze the Data

regress

It is the most basic and commonly used command in Stata.

ivregress

It is used to perform IV regression.

graph twoway

It has a lot different graph types and options. Check help file.

Finding Help

The most efficient way of learning Stata is using help file that Stata carries on its own, and searching on Google.

help

It is used when you know the name of the command, but not sure about the syntax or options.

findit

It is the best way to search for information on a topic across all sources

search

It searches a keyword database and the Internet.

And of course, don't forget [Google](#). It is super useful for finding tutorial, examples, commands, etc.

Two Important Files

1. Do file

Do file saves the commands you perform. It is useful if you want to do the analysis more than once and for future reference. You can use any text editor to create a do file, just save as “.do” file. However, it is most convenient to use Stata built in Do-file Editor. You can start the editor by clicking the icon or choose “Window → Do-file Editor”. To run the do file, simply click “run” button in the editor, or you can type

do filename

2. Log file

Log file is used to record all the operations you perform, including the command you type, the output, even the error information. Basically what you see in the output window would all be recorded in a log file. It is very useful to record the result for later use. You can open and close a log file using the icon on the menu bar, or you can type

log using filename[, option]

log close