Stata: A (Somewhat) Brief Introduction
140.655 – Spring 2001

Tom Travison

What: Stata is a statistical computing package available for Windows and Macintosh platforms. It has several "nice" features, including a user-friendly graphical layout, spreadsheet-style data representation, easy transfer of output to various word processors, and efficiency (3 floppy disks). It is also easy to use, yet quite powerful. Minimum system requirements for PC include at least a 486 generation processor, 8MB RAM, and Windows NT/98/95/3.1 or later versions of DOS. Stata is also relatively inexpensive (approx. $100).

When: Stata will be useful for this course and beyond. During the course of the year, you will develop skills in data analysis using this computing package. Supplemental handouts will assist you in performing these operations using Stata. Best of all, the Stata license is perpetual – if you buy it, you own it forever.

Why: Stata is capable of performing all of the analytic methods introduced in the 620 series and most of the analysis you will perform in your career. It is also **command line** (as opposed to **menu**) driven, which allows you to fine-tune your operations and personalize your output.

Where/How: Each machine in the Hygiene building computer labs (W3017, W3025) is equipped with Stata; it is also installed on machines in the Hampton House basement lab. These labs may be crowded, however. User Support (W3014) has information about purchasing Stata. It is also easy to order by calling the Stata Corporation at 1-800-StataPC. Their "Gradplan" package includes Stata and a full set of manuals for $196. Be sure to order the INTERCOOLED version of Stata. If you purchase by credit card, the software and manuals may be picked up at User Support by the end of the next business day.

1. Introduction

This handout is intended to help you become familiar with Stata. Of course, we cannot tell you everything you will ever need to know in these few pages, but they should help you in establishing a firm foundation on which to build your knowledge.

This handout is organized into six sections. Each details a portion of the necessary steps and precautions you should take when using Stata. Because the Hygiene labs are (almost entirely) PC and Windows95-based, the discussion will be most applicable to the Windows95 compatible versions of Stata. However, Stata is fairly consistent across its various incarnations and any beginning user should benefit from the reading. Please pay particular attention to Section 2, which details the initial steps you should take upon opening Stata and the use of the two kinds of files associated with Stata, data (.dta) and log (.log) files, and to Section 6, which tells you how to close out Stata without losing your work.

2. Beginning in Stata

In the Hygiene labs, you open Stata by clicking on the "Start" button at the lower left corner of the screen and selecting the items "Programs," "Intercooled Stata," "Intercooled Stata" in the subsequent menus. At this point Stata will open and fill all or most of your screen. On another machine you will need to follow the path to the Intercooled Stata menu item.

Stata has a primary background screen and a set of internal windows serving different functions. The background screen has a menu at the top (File Edit Prefs Window Help). If you click on any of these words, a pull-down menu will appear listing a set of things you can do. Below the menu are a few buttons for things you may do especially often, including opening a log file (see below under Stata files) and switching between Stata's windows.

When you first open Stata, four internal windows open automatically. The "Command" window is where you enter **commands**. After performing any operation command, the result will show up in the "Results" window. The history of the commands you've entered is shown in the "Review" window. You can click on one of your former commands at any time to use it again. However, the history of what results Stata has given you is not automatically saved. The final window that opens when you first open Stata is the "Variables" window. This lists the variables currently in memory, along with whatever label you have attached to them.

Three other windows are available in Stata. The Log window will open if you decide to open a log file. Even if you close the log window the log file will continue to save whatever appears in the Stata Results window. The "Editor" (or "Browser") window can be opened by clicking on the relevant button under the main menu. The Editor window allows you to look at and modify the dataset currently in memory. The Browser allows you to read, BUT NOT TO CHANGE, the data. Finally, if you graph something in Stata, the "Graph" window will open. THE GRAPH WILL NOT BE SAVED IN YOUR LOG FILE, but you can save it separately using the menu.

So, what's the first thing we should do when we open Stata? What we need to worry about, as anyone who uses a computer regularly knows, is saving, saving, saving. THE FIRST THING WE SHOULD DO UPON OPENING Stata IS TO CREATE A LOG FILE. Click on the "Log…" button at the upper left corner of the screen. Stata will prompt you for the name of the Log file. You can browse around your folders and find one you've already created or make a new one.

Once you open a Log file, a white Log window will appear on the screen. You can close this window without closing the Log. All of the text generated by you or Stata will now appear in the Log file. Later, your Log can be opened and edited with any word processor or text editor. If you open Stata again and you want to use the same Log file, just choose that log file and "append" your new work to the old work. If you choose to "overwrite", all of your previous work will be lost.

The next thing we are concerned about is actually obtaining the data we want to work on. If this data already exists as a Stata file we can simply open it from the menu. Under the File menu, there is an Open option, just as in most Windows applications. Clicking on this option allows you to open Stata data (.dta) files, which you may then edit in the Editor. If you plan to change this data set, YOU SHOULD IMMEDIATELY SAVE IT UNDER A NEW NAME (under the File menu). You may also enter raw data using the Editor.

The easiest way to enter your raw data is in the Editor. Click on the button for Editor, and then click on the upper-leftmost white square. If you have ever used a spreadsheet the layout should be familiar. If not, it's fairly simple anyway. Each column will be one variable and each row is one observation. You can enter your data directly and name it by double-clicking on the gray box at the top of the column. The *name* is the short description you will use for your variable when you are giving Stata commands. You would also be wise to give your data a label, which is the longer description for it. This will appear in the Variables box (once you close the Editor) to the right of the name, and it will help you keep track of which variable is which. Also, when you make a graph with that variable, the axis will be labeled with your label.

3. Telling Stata What to Do

As mentioned above, Stata is command-driven . This means that for many operations, and for most if not all actual calculations, you will need to be bossy and type commands onto the screen. Some of these commands are quite intuitive or self-explanatory; others are not. All, however, follow the same basic format, or syntax:

command *variable(s)*, **options(***set***)**

Of course**,** you don't actually type the word **command** into Stata (It doesn't need to be **boldfaced**, either. We're just trying to make things easier for you to read). Instead of **command**, substitute the function you would like Stata to perform. Because you are giving Stata a **command** (the way you might yell "STOP!" to someone about to step out into moving traffic on Monument St.), the command word is usually a verb, like **compress, tabulate** or **type.**

Variables are exactly what their name implies. They represent different values that are tabulated for some (or all) of the observations in your data set. You (or someone you work with) will give them names. For example, if we are studying incidence of pneumonia in East Baltimore over the last two years, our observations might consist of tabulations of the number of visits to the Emergency Room at Johns Hopkins Hospital over that time period.

Then some *variables* might consist of each patient's *height, age,* and *weight* and whether or not a person is experiencing *pain* or has a *cough.* Some of the data might look like this:

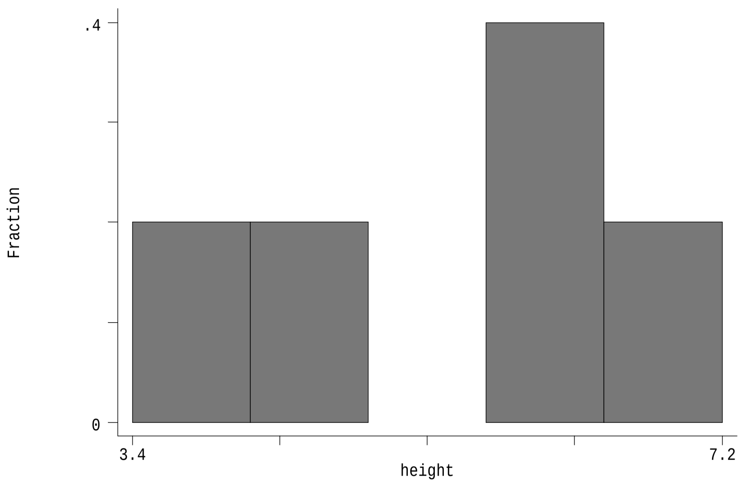
age	height	weight	pain	cough
18	4.1	109	1	0
11	4.3	97	1	1
22	7.2	222	1	0
25	4.9	101	1	0
16	6.3	190	1	0
37	6.5	212	1	0
76	5.8	156	0	1
19	6.3	176	1	1
40	6	187	0	1
12	3.4	84	0	1

If you want to work through this document in Stata, enter the data above into the Editor now. Label the variables as we described in the previous section.

Suppose in our pneumonia example we want to make a graph of the patients' *height.* We could type:

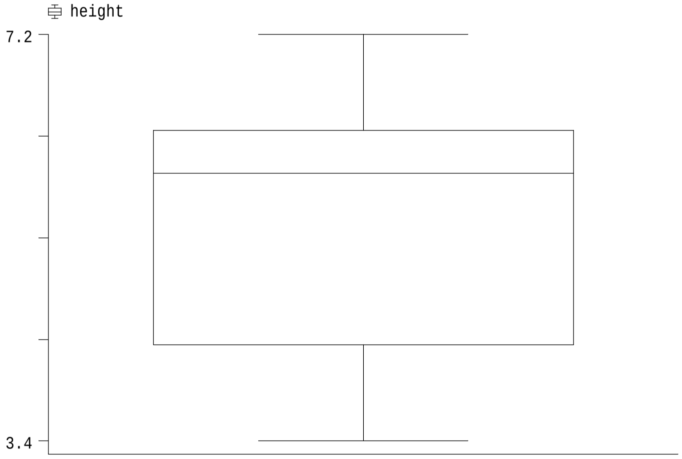
graph *height*

Stata would read the **graph command**, and would make a histogram (bar graph) of the *height variable:*



Options are, by definition, optional (with a few exceptions). Notice that they occur after a comma (,) in the command syntax. Suppose that we want the *height* graph to be a boxplot instead of a histogram. We would type:

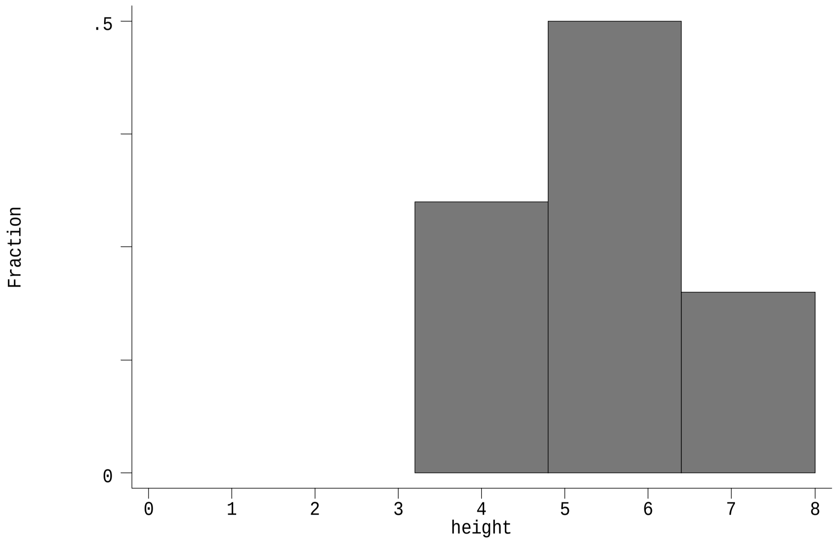
graph *height*, **box**



Graph is the **command.** *Height* is the *variable*, and **box** is an **option**. **Box** doesn't do anything by itself. If I simply type **box** into Stata, I would get an error message. **Box** is an option that modifies **graph** – it changes **graph's** behavior and makes it produce a boxplot instead of a histogram. This is the difference between **commands** and **options.** As we have said, **options** are (usually) optional, but **commands** are not.

The final piece of the puzzle, *set*, is in many cases created by you, the user. Suppose once again we wanted to make a bar graph of *height*, and we wanted to label the *x* (horizontal) axis at 0,1,2,3,4,5,6,7, and 8 feet. We could type

Graph *height*, **xlabel**(*0,1,2,3,4,5,6,7,8*)



Now **xlabel** is our option, and the *set* is the collection of numbers 0-8. You can think of the *set* as a user-defined portion of an **option**.

That's it! You need to learn individual **commands**, but everything has a similar structure. Keep in mind that each **command** has its own **options**. Use the Help menu!

4. More Graphing

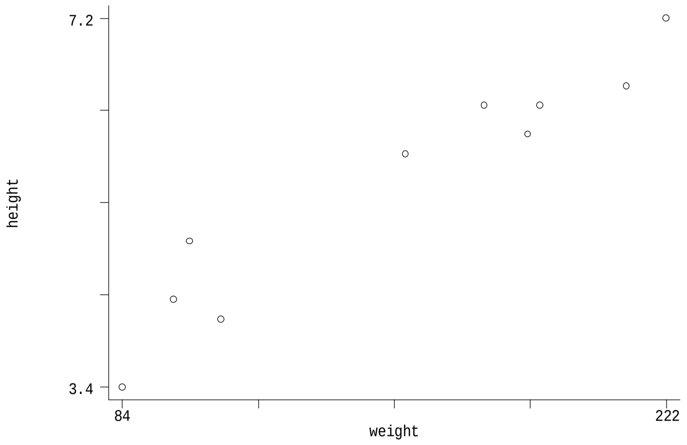
One command that you'll get very comfortable with is **graph.** One of the most powerful ways to observe relationships in data is to take a look at some pictures of it. Plotting the data is thus very important. In Stata, the basic format is

graph *varlist*

where *varlist* represents the variables you would like to graph. If I type

graph *height weight*

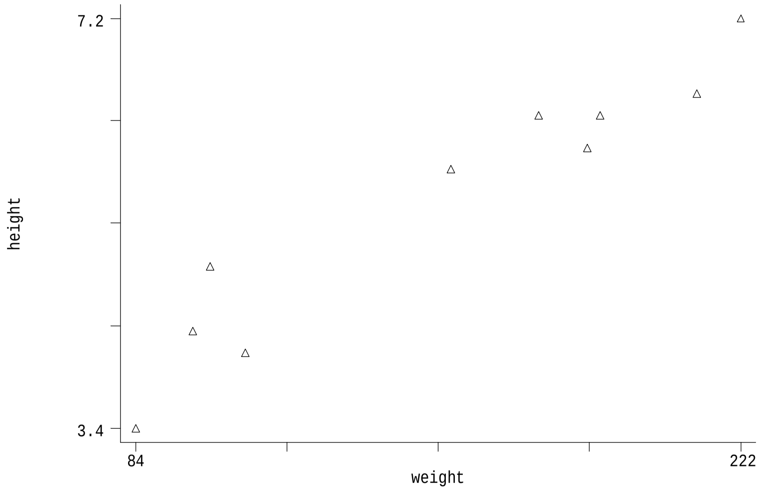
Stata will open its "Graph" window and show me a scatterplot of heights vs. weights (below), with *height* on the vertical axis and *weight* on the horizontal.



(As we saw earlier, if we specify only one variable Stata will instead give us a histogram, or bar graph).

You can also add options to change the way a graph looks:

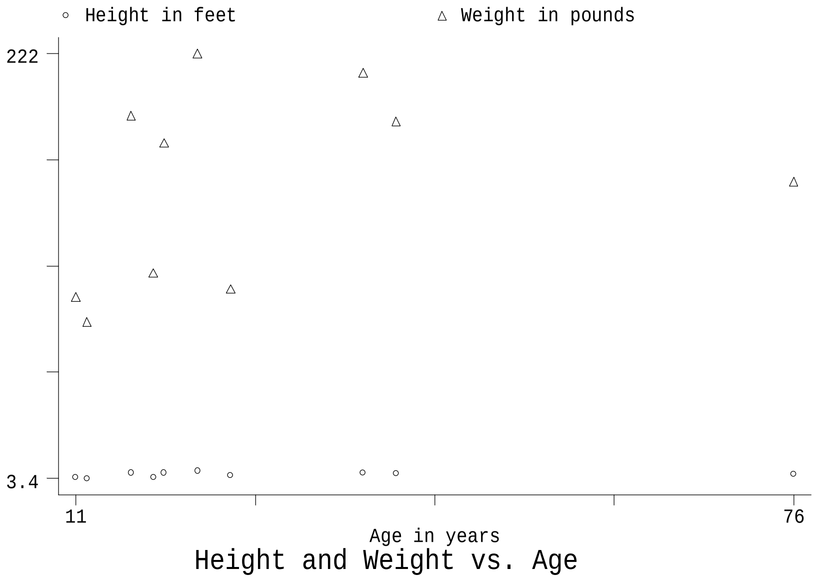
graph *height weight*, **symbol**(T)



This will produce the same graph as before except that each observation will be marked with a triangle. Instead of "(T)" you could also use (o) little circle, (0) bigger circle - this is the default, (i) invisible, (.) dot, (d) diamond, (p) plus sign, ([*var*]) the name of a variable, or ([_n]) the actual value of that observation.

You can also graph more than two variables at once.

graph *height weight age*, **symbol**(*oT*) **title**(*Height and Weight vs. Age)*

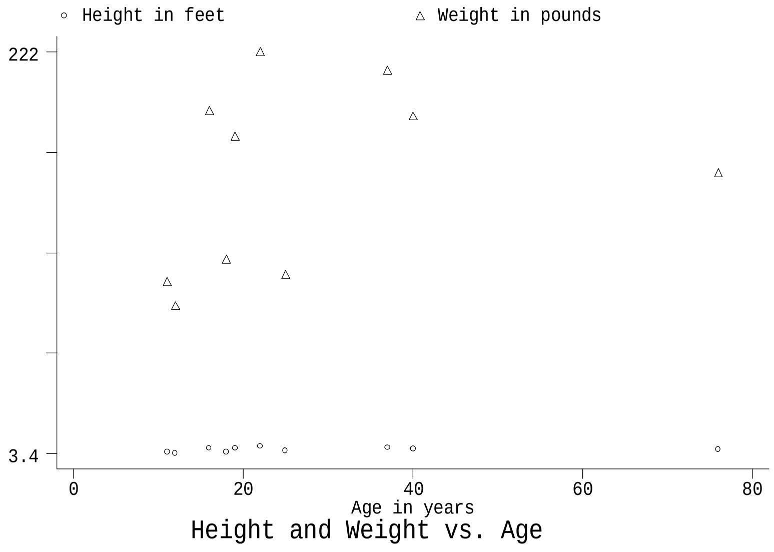


Now we have a much more interesting picture! This command builds on the same ideas as above. All the variables you list, except the last, are plotted against the last one you list. In this example, *age* will be the x-axis, and the other variables will all be superimposed over each other on the y-axis. The x-axis will automatically be labeled with whatever label you made for *age*. The option **symbol** has a *set* of parameters, corresponding to the first two variables you listed. Thus *height* will be shown as circles, while *weight* will be triangles on the graph. You can expand this with any number of variables. Just remember that the last variable you list will be your x-axis and each of the others will be plotted against it. There is also a title at the end of the syntax. Notice the text of the title below the graph itself.

Notice that both options are listed after the comma. The order of the options is irrelevant.

Another thing you can do to make your graphs look nicer is use the options for labeling the axes. If you add **xlabel** after the comma, Stata will label the X-axis with easy-to-read numbers. We saw this in the last section.

graph *height age weight*, **symbol**(*oT*) **title**(*Height and Age vs. Weight)* **xlabel**



The x-axis is now labeled with “nice” numbers that are easy for us mere humans to comprehend!

Once you have a graph you want to use, you can do any number of things. You can print it directly by clicking on “file” in the menu and then “print graph”. You can save it to look at later in Stata by clicking on “file” in the menu and then “save graph”. You can cut and paste the graph directly into a word processor by clicking on “edit” in the menu and then “paste” or “paste special” inside the word processor (the way we did to make this document).

5. Using Help

Stata's help menu is quite useful. If we click on the Help menu at the top of the screen, a dialog box appears . We can look up any topic by typing a word or words in this dialog box. If we type **graph** and hit the OK button, we are given a list of topics that refer to **graph**, some more relevant than others. But if on this menu we click on the **graph** pointer then we come to a section that has the syntax for graph, like this:

```
[by varlist:] graph [varlist] [weight] [if exp] [,options]
```

Sometimes the help can be a little intimidating! But the meanings are fairly straightforward. Anything inside the square brackets [ ] is optional; we do not need to use **by** or have **weights** to use graph. These are part of the command that will tailor it to your particular needs.

It is usually instructive to scroll to the bottom of a particular page in Help. Here there are examples which often illustrate the syntax better than abstract text can. You should get into the habit of using Help.

6. Closing Stata

To leave Stata you must exit the editor and save your data one last time. ALSO, BE SURE TO CLOSE YOUR LOG FILE. Click on the "Log…" button and choose the Close option. This will finish writing your log and prevent you from crashing Stata and losing the Log. Now your Log is saved for as long as you like.

Be sure that you have saved you Data (.dta) file as well. As mentioned in Section 2, you should continue to rename this file each time you change it, as you may want to retrace the steps your analysis at a later date.

This handout should provide you with a good start for Stata. Good luck!