An Introduction to Stata
Stata Workshop 1
University of Kansas

Contents

1 Introduction

An Introduction to Stata
Stata Workshop 1
University of Kansas

Goal: To become more familiar with the statistical software package Stata.
In this workshop we will focus on creating and loading Stata data sets as well as basic exploratory commands.

Steps:

1. Open Stata on the KU virtual lab or use software installed on your laptop.
2. Notice that you could program in a command line (as Stata is a programming language). HOWEVER: you should always put your code and commands in a do file. Open the Do-file Editor.

2 General features

2.1 Working Directory

The Working Directory
It is always important to know where the files that you are using are saved on the computer. This is so both you and Stata can access the correct files. Let's see how this would work for this problem.
For a data analysis project, your file system should be something similar to the following:

- Data
- Working Data
- Do Files
- Literature
- Output (the LaTeX file should be located in output for access to graphs)

Save the data for this project in a working directory and set the folder containing data as the working directory. The command **cd** will 'change directory' in Stata, and you need to add the file path after that command.

Typing **pwd** into the command line will tell you the current directory and can be used to verify that you are in the correct folder.

2.2 Do Files

"Do-files" and comments

All the commands you enter into the Command Line for the lab can (and should) be put into a "do-file" to allow replication and access at a later date. You should save a "do-file" in the folder for do files and number them in the order that you would use them to transform and analyze the data.

You should heavily comment your do files to explain what you are doing and record the process that you follow.

3 Creating new data sets

Sometimes you need to create your own data sets from scratch (e.g., when analyzing data you have collected). This lab familiarizes you with some of the ways to create Stata data sets. One of these is to use Stata's Data Editor and the other is to import data from a text file using the insheet command.

3.1 Option 1 - Use Stata's Data Editor

Data analysis tip: This is one of the times when it is easier to use a Stata menu option rather than simply typing the commands into the Command line.
Click on the Data Editor (Edit) button or navigate to Data − > Data Editor − > Data Editor (Edit). This brings up a blank spreadsheet that will become the data set. To keep things simple, we will use observations (rows) for this part of the lab.

Here are the data for the eleven most populous countries in 2009:

country	population
China	1,320
India	1,160
United States	307
Indonesia	240
Brazil	199
Pakistan	181
Bangladesh	154
Nigeria	149
Russia	140
Japan	127
Mexico	111

Enter the country names in the first column and enter the population in the second column. Once data is entered, you will note that the columns were automatically given a name, var1 and var2. This is a meaningless name for a variable, so we should rename it. Double click on the box containing var1 and change the variable name to country; do the same for var2 naming it population.

Data analysis tip: When you create data sets for your own research, give your variables descriptive labels. It is easier to interpret analyses when the output has descriptive labels than when the output has labels like var1, var2, var3, etc. Descriptive labels also make your data set comprehensible to others who may want to use it. Finally, if you use the data set in future analyses, you will not have to spend lots of time trying to decipher uninformative variable names.

Note that the color of the values for country and population are different. Stata distinguishes between variables containing only numbers (numeric) and variables that contain letters, symbols and other characters (string).

Now that the data has been entered, return to the main Stata window. Note that now we have 2 variables in memory and that the results window has a lot of different commands in it (e.g., **replace var1 = 1 in 8**; **rename var1 country**). Everything you did in the Data Editor could have been done by typing commands into the Command line. However, as mentioned, this is an instance when it is easier to use the menu rather than the command line.

3.2 Option 2 - Import Data from a File

Another way to get data into Stata is to read it in from a text file with the insheet command. This requires two steps in this case: making the dataset, then reading it in.

1. Make the dataset

    - Open up a text editor and copy and paste the above data into it, including the headers.
    - We want to make the file a comma-delimited or comma-separated file, which means that each column of data is separated by a comma. The numbers have a comma as the thousands indicator and we do not want that, so remove the commas from each number. (See note below).
    - Now replace the space between each country and number with a comma.

    The first two lines should be:
    country,population
    China,1320

Make sure the headers are also separated by a comma, and save the file as "countries.txt" in the same directory that all of your work will be stored. Remember you can always check the Present Working Directory by typing pwd.

Note: With comma-delimited files, you can also put items in quotations that you want to be treated as one column if you want to preserve commas. Thus the first line could also be represented as China,"1,320" if we wanted the comma to be preserved.

Data analysis tip: When creating your own data, you often have to work a lot with text editors/Excel. Using the "Find and Replace" option found in these programs can save time. First replacing the commas in the numbers with blanks, then replacing the spaces with commas would make formatting this dataset very easy.

2. Read in the dataset
The general syntax to read the data in is: **insheet using filename**, where filename is the filename (with extension) of your data. Use help insheet to see more about the options that can be used with the command. Use this dataset to work through the following illustrative questions.

3.3 Questions:

1. How many countries have a population greater than 250 million?
   With ten cases it is straightforward to look at the data and get an accurate count. But with a longer dataset, counting the incidences of each number "by hand" would be cumbersome. In such settings, you can make life easier by sorting the numbers in increasing order, then count the incidences. We should do this in Stata just to get familiar with this command: **sort population**.

   The data is now sorted in ascending order by population.
   Typing browse into the command line will open the data and let you look at the sorted data.

   Data analysis tip: Stata is case sensitive. The commands **sort Number** and **sort number** are commands referencing two separate variables, so an error will occur if you do not use the same case as your variable name. Also, **Sort** is not a command while **sort** is. You can try out the four combinations (SN, Sn, sN, sn) and you will see that only one is valid.

2. You can also count conditions in variables using the **count if** command and then a condition like >100: e.g. **count if population>100**. Rather than using sort, use **count if** to count the number of countries whose population is greater than 250 million.

Data analysis tip: There is a more advanced form of sort called gsort which allows the user to specify the order (ascending vs. descending) that each variable should be sorted on. This can be navigated to by Data − > Sort − > Ascending and Descending Sort.

4 Loading General Data Sets

Load the Penn World Tables dataset provided into Stata. Steps:

1. Download excel file and save it somewhere accessible on your computer.
2. In Stata, go to file and import; you should choose the .xls or .xlsx option.
3. Use the browse function to find the location on your computer where you saved the PWT file. Make sure you indicate that you want the 'first row as variable names'.

When you get a data set, the first thing to do is to figure out how many variables and how many units of observation you have to play with. This is pretty easy in Stata through the **describe** command. The describe command will list information about the dataset and its size as well as information about each variable. If you just want the information about the dataset, use **describe, short**. Alternatively, if you just want information about a specific variable, type **describe varname**.

Let's get into some data analyses. Remember to save all of your commands in a "do-file".

4.1 Useful Commands for Questions

Stata has numerous commands which allow you to change and explore the data in ways which can help you better understand it and analyze it. Below are some commands which may be helpful to think about the questions which follow.

- To save data in Stata, make sure you are in the correct working directory. If not, use the **cd** command to change the working directory to the proper location for your data. Once you are in the correct working directory, the command to save is simply **save filename**. Filename here should include the extension (.txt, .dta, .xlsx, .csv, etc).
- It is often the case that you have data over many years and you want to look at only one year. Save a temp file (**save temp.dta** and then keep the year that you want, i.e. **keep if year==2007**.
- Graphing is often the best way to relay the story the data can tell. Stata has many graphing capabilities using a variety of different commands. One particularly useful command is **graph twoway** followed by a type of plot such as **(line y-variable x-variable)**. An example would be **graph twoway (line gdppc year if country=="United States") (line gdppc year if country=="Germany")**. The twoway part allows you to do multiple plots. If you only want one line graph, you could do **line gdppc year if country=="United States"**.

4.2 Questions

1. Stata displays missing values with dots, ".". How many values are missing from the variable gdppc? What percentage of the dataset is missing?

   Data analysis tip: Sometimes you want to be able to add notes into your do-file that are not commands. It is very helpful to leave yourself comments as you go. To create a comment in Stata, simply begin a line of text with an asterisk (*) or a double backslash.

   Data analysis tip: It is common for some data to be missing in a file. Unfortunately, there is no universally accepted way of representing missing values. Some software packages, like Stata, use a dot or period. Other packages use an "NA" for not available. Some data producers, often federal agencies, use extreme values of a variable (e.g., -99) to indicate missing values. Using extreme values is bad practice: how does the user know if the value is correct as written or if it is a dummy entry to denote missingness? When you get a data set from someone, learn how they code missing data before doing any further analyses.

2. Of all the countries in the dataset, which has the highest GDP per Capita in 2007? Which has the lowest in 2007? (hint: Save the file as temp.dta in your working directory. Keep only year 2007 and then sort gdppc).

3. Which country has the highest average GDP per Capita over the timeframe of the dataset? Which country has the lowest GDP per Capita over the timeframe of the dataset?

   First, note that Stata has a very useful command summarize that will give you basic summary statistics for a certain variables. summarize gdppc will give you the mean, std dev, min, max, and count of gdppc. In order to find similar stats for each country, it is good to use tables.

   There are 3 basic ways that you can create such a table. Each row in the tables that we will create here will list the mean, std dev and frequency of gdppc by country.

   (a) tabulate By default, tabulate calculates frequencies. So typing tabulate country will tell you how many times the country name appears in the dataset (try this). In order for it to give you different information (mean, std dev, freq), you simply use the summarize option: tabulate country, summarize(gdppc)

   (b) table The table command is very useful in that it gives you great flexibility in deciding how you want your table to be organized and what information you want. Like

tabulate, you tell it the variable(s) by which you want the table to be organized, and then tell it which statistics you wish to have calculated for those variables, with the contents() option: table country, contents(mean gdppc sd gdppc N gdppc)
Why is N different for each country?

(c) tabstat The tabstat command works a little differently. You first tell it the variables for which you are interested in getting summary statistics (gdppc) and then tell it how to break it out (by country) and which statistics you want (mean, standard deviation): tabstat gdppc, by(country) stats(mean sd N)

You can verify that each one of these three tables produces the same values for the mean and standard deviation.

4. How many countries have an investment share over 30 percent? How many countries have a consumer share over 70 percent?

5. Suppose you are from Tunisia and your friend is from Bangladesh. Your friend tells you that Bangladesh has a higher GDP than Tunisia does and is a wealthier country. What is another perspective that could be taken? Are the people of Bangladesh on average wealthier than the people of Tunisia?

6. Create a line graph of the variable gdppc for Tunisia and Bangladesh over time. Can you get the label to use the country names?

7. Recreate the graphs provided in the do file. Think about how these require the creation of new variables and how you might use the existing data to generate new variables that help you answer interesting questions. The code provided creates two new variables: decade and gdp. Use them to make the box and whisker plots by decade. Try to generate a new variable of your own and use it to create an interesting graph.

8. Explore the data to answer at least one question that interests you. Think about how you could use this data to begin to write a paper about a topic in Development Economics that is relevant to some of the salient questions in Development Economics: why are there differences in income level across countries? Why are their differences in growth rates across countries? Will these differences continue to persist or will there be convergence of GDP per Capita over time?