Spring Stata Workshops
Data Wrangling
University of Kansas

Contents

1 Introduction

An Introduction to Stata
Stata Workshop 1
University of Kansas

Goal: to learn about data wrangling about Stata.
In this workshop we will focus on loading and managing data in Stata as well as basic exploratory commands.

Steps:

1. Open Stata on the KU virtual lab or use software installed on your laptop: Virtual Desktop Link.
2. In the command line Stata, run the command: ssc install wbopendata

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 well 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 Getting Data Into Stata

In order to conduct data analysis, you will need to load data into your Stata session. This lab will practicing loading data into Stata from a file stored on your local drive or a cloud-based storage system and also loading data into Stata from an API.

3.1 Option 1 - Load Data with an API

This part of the workshop will focus on an application programming interface (API). The World Bank has created an API for Stata to access the data housed in their World Development Indicators database. We will use that API to bring in several indicators of interest for all countries where that data is available. Link to World Bank API for Stata

This interface is accessible in Stata through the command *wbopendata*. In order to use this command, it must be downloaded through the Stata SSC using *ssc install wbopendata*. You should only need to run the install command once.

Use the wbopendata command to bring in change in inventory, manufacturing, manufacturing percent share of the economy, agriculture percent share of the economy, trade percent share of the economy, capital formation percent share of the economy, gross national income per capita, gross domestic product, gross domestic product growth rate, and gross capital formation growth rate.

Each indicator is accessible through a code and if you know that code then you can call individual indicators with the API. The API also has the ability to call indicators by topic, and we will practice using the API to pull in a set of indicators by topic as well as a set of indicators by indicator code.

3.1.1 Useful Commands for Questions

Stata has numerous commands which allow you to manipulate 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 gdp_growth year if countryname==“United States”) (line gdp_growth year if countryname==“Germany”)**. The twoway part allows you to do multiple plots. If you only want one line graph, you could do **line gdp_growth year if country==“United States”**.

3.1.2 Questions

1. Create a line graph for United States' GDP growth over time.
2. Create a line graph that combine the time series for United States' GDP growth and Germany's GDP growth over time.
3. Create a boxplot for GDP growth over region and income level.
4. Look at the relationship between Capital Formation and Gross Savings.
5. Examine how high income and upper middle income countries capital formation changed after the Great Recession relative to low and lower middle income countries.
6. Do the capital formation shares look similar for these groups of countries prior to 2009?

3.2 Option 2 - Import Data from a File

Load into Stata the stata_workshop dataset we created earlier in the lab. Steps:

1. Make sure that the excel file is saved 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 stata_workshop 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".

3.2.1 Questions

1. Stata displays missing values with dots, ".". How many values are missing from the variable gdp_growth? 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 growth rate 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 growth rate since 2010? Which country has the lowest GDP growth rate since 2010?

   First, note that Stata has a very useful command summarize that will give you basic summary statistics for a certain variables. summarize gdp_growth will give you the mean, std dev, min, max, and count of gdp_growth. 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 gdp_growth 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 countryname, summarize(gdp_growth)

(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 countryname, contents(mean gdp_growth sd gdp_growth N gdp_growth)
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 (gdp_growth) and then tell it how to break it out (by country) and which statistics you want (mean, standard deviation):
tabstat gdp_growth, by(countryname) 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 a gross savings share share over 30 percent in 2022? How many countries have gross savings share of under 30 percent in 2022? How many countries are missing data for their gross savings share in 2022?

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? Look at the Gross National Income per capita.

6. 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 International Economics that is relevant to the three main questions: 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?