An Introduction to Stata
Stata Workshop 2
ECON582: University of Kansas
August 30th, 2019

Contents

1 Introduction

Exploratory Data Analysis with One and Two Variables
Stata Workshop 2
ECON582: University of Kansas

Goal: To explore data with histograms and scatter plots.
In this workshop we will focus on creating and loading Stata data sets as well as basic exploratory commands.

A few highlights to review from the previous lab:
The Working Directory

1. 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.

2. Download the **WB Poverty Panel** data from Blackboard and save it to a folder of your choice. Set that folder as the working directory of your session. You can do this in one of two ways:

   (a) Use the menu bar and navigate to File $->$ Change Working Directory. This will pull up a window that you can then use to find the folder where you saved the data.

   (b) Issue a command to change the working directory by typing **cd** followed by the file path to your file.
   Typing **pwd** into the command line well tell you the current directory and can be used to verify that you are in the right place.

"Do-files" and comments

1. 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.

2 Lab Procedures

Exploring Poverty in the World
Which countries and regions have the highest level of poverty? We can begin to examine this with the WB Poverty Panel Dataset. This data was obtained using the World Bank API using the topic "Poverty". Procedures for how this dataset was constructed can be found in the Stata do file located on Blackboard called "01. WB Panel Creation".

The dataset comprises various measures of poverty on all countries observed by the World Bank. The time span ranges from 1960 to 2018. We want to visualize the distributions of these variables across time and across geography in this lab.

The variables included are:

Variable	*Description*
countrycode	Country Code
year	Year
countryname	Country Name
region	Region Code
regionname	Region Name
adminregion	Administrative Region Code
adminregionname	Administrative Region Name
incomelevel	Income Level Abbreviation
incomelevelname	Income Level Name
income100	Income share held by highest 20%
income80	Income share held by fourth 20%
income60	Income share held by third 20%
income40	Income share held by second 20%
income20	Income share held by lowest 20%
consumption_growth	Annualized average growth rate in per capita real survey mean consumption or income, total population (%)
consumption	Survey mean consumption or income per capita, total population (2011 PPP $ per day)
consumption_growth40	Annualized average growth rate in per capita real survey mean consumption or income, bottom 40% of population (%)
consumption40	Survey mean consumption or income per capita, bottom 40% of population (2011 PPP $ per day)
rural_poverty	Rural poverty gap at national poverty lines (%)

3 Overview and Exploratory Analysis

3.1 Useful Commands for Overview and Exploratory Analysis

- To get a visual representation of the data, it is often helpful to see the distribution using a histogram. **histogram varname, name(graph1, replace)** where varname is the variable of interest. As with most commands, there are many options available with histogram. One of the most useful options is name(graph1, replace). This stores a version of the graph in temporary memory to be used later and it causes the graph to be displayed in a window titled graph1. This allows multiple graph windows to be open at the same time. You use the replace option within the parentheses of the name option to overwrite the previous version of the graph.

- If you want to compare multiple histograms in one window you can combine them by typing **graph combine graph1 graph2, name(graphall, replace)**.

3.2 Overview and Exploratory Analysis Questions

1. After reading in the data, describe the distributions of foreign and domestic grosses. That is, say where most values are, note any outliers, and say whether the distribution is tightly packed around its mean or is spread out. Also, report the mean and standard deviation.

   Data Analysis Tip: The default histogram in Stata is a true histogram, where the areas of the bins sum to one. Often people want just the heights to sum to one. This is accomplished with the fraction option. Further, if you want the y-axis to simply count how many observations are in each bin, you can use the frequency option.

2. Which sentence best describes the distributions of domestic and foreign grosses?

   (a) Use the histogram function to graph the distribution of the consumption variable. Set the bins at 5 and choose to do fraction instead of density. Describe the shape of this distribution. Call this graph something of your choice using the name(graph1, replace) command described above.

   (b) Obtain a histogram of of consumption growth and describe the shape of the distribution. Combine this graph with the graph from part (a).

   (c) Use the histogram function to explore the different income distributions (income20, income40, and so forth). What do you see?

   (d) Try using the **over()** command to do a histogram over incomelevelname. Is this possible? How else might you achieve the result?

4 Analyzing the Data

4.1 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.

- The command for a box plot is **graph box contvar, name(graph1, replace)** or **graph box contvar, name(graph1, replace) over(catvar)**, where contvar represents the continuous variable that you are trying to graph. The option over(catvar) allows you to break the box plot down by different values of a categorical variable.

- To make a scatter plot, we will tell Stata that we want to do a twoway graph as a scatter: **graph twoway scatter varname1 varname2**.

- To find correlations in Stata, type **correlate varname1 varname2** which will show a matrix of the correlations between all the variables used as input (you can use as many as you would like). Note that the diagonal is always 1. Make sure you know why that is.

- In addition, you can create a scatterplot matrix which will create scatter plots of all the different outcomes by typing **graph matrix varname1 varname2**.

4.2 Questions

1. Use the scatterplot function to see what relationship there might be between the level of consumption and the growth of consumption. For data demonstration purposes, what variable makes the most sense on the x-axis?

2. Use the correlate and reg commands to test the relationship between consumption and consumption growth. What do these commands tell you?

3. Create boxplots for each of the income quintile variables over regionname. Why do these boxplots tell you about income inequality within countries? (hint: within the over command, you can put **label(angle(45))** and this will help make the categories readable.

4. Now use the hbox command instead of the box command to plot each of the income quintile variables over regionname. Is this a better way to display the data?

   Data Analysis Tip: Note that != is defined as "not equal to" and — is the "OR" operator.

   Data Analysis Tip: It is not acceptable to exclude outliers from analyses unless you have a scientific reason to do so (e.g., a data entry error, or maybe the outlying unit is not part of your target population). Hiding outliers is fudging data to get results you want. That is dishonest and unethical. When you see outliers, do analyses with and without them. When the results do not change much, report the results based on the full data set, and tell your audience that the results were not sensitive to the outliers. When the results do change substantially, report both sets of analyses: one with and one without the outliers. This honestly informs people that your conclusions are not on very solid ground, because particular data points affect the results greatly.

5. Feel free to explore relationships between the variables in ways that you might find interesting. Try different kinds of groupings across variables and see what patterns become evident.

6. Read about the wbopendata command (type **help wbopendata** into the command line in Stata. Try to load in particular parts of the World Development Indicators data based on the descriptions there and create a dataset of variables that interests you. You can use the code on Blackboard which created the dataset for this worksheet as a template if you wish (called WB Poverty Panel Do File).