Center for Teaching, Research and Learning
Research Support Group
American University, Washington, D.C.
Hurst Hall 203
rsg@american.edu
(202) 885-3862

Introduction to STATA

WORKSHOP OBJECTIVE:

This workshop is designed to give a basic understanding of how STATA works and how to use simple statistical methods to analyze data.

LEARNING GOALS:

1. Understand the basics of the STATA interface
2. Managing data
3. Computing descriptive statistics
4. Creating basic graphs
5. Running simple linear regressions

SCENARIO:

We are interested in studying the relationship between the amount of corruption in a country and the quality of their economy.

DATA: For this question, we are lucky as data already exists. We will use the “World Development Indicators” which are compiled annually by The World Bank. We are using data from a past year, and we have simplified the data set, but you can find the full data set here: http://data.worldbank.org/data-catalog/world-development-indicators. We are also lucky, because the World Bank saves their data in a way that makes it easy to open. Often when you find data set (for example, old Census data), it can take a lot of effort to transform the data into a usable form. And, of course, there is never a guarantee that someone else has created the data set you want in advance; often the first step in a research project is spending a long time looking up values to create the data set you need.

Within the WDI data set, we will focus on the variables “corrup” and “gdppc.” The first is an index of perceived corruption within a country’s government. Citizens are surveyed, and asked questions like “Would you need to plan bribes to start a local business?” Note that corrupt is “reverse coded” so that high scores equal low corruption. The second variable is Gross Domestic Product per Capita, which indicates how well the economy is doing, scaled for the population size of a country (for example, it is not very interesting to note that Botswana produces less than Brazil overall, but it might be interesting to note that Botswana produced more *per citizen* than Brazil in 2014).
Note that, in using these variables from the WDI data set, we have given over the responsibility of determining what “corruption” and “quality of economy” means to the World Bank. We have also

allowed dynamic concepts to become simple numbers. However, this allows us to move quickly to the later stages of analysis, which has its advantages.

I. Understanding the basics

- In this section, we introduce the STATA interface and few basic but very helpful commands. **the Stata Interface.**

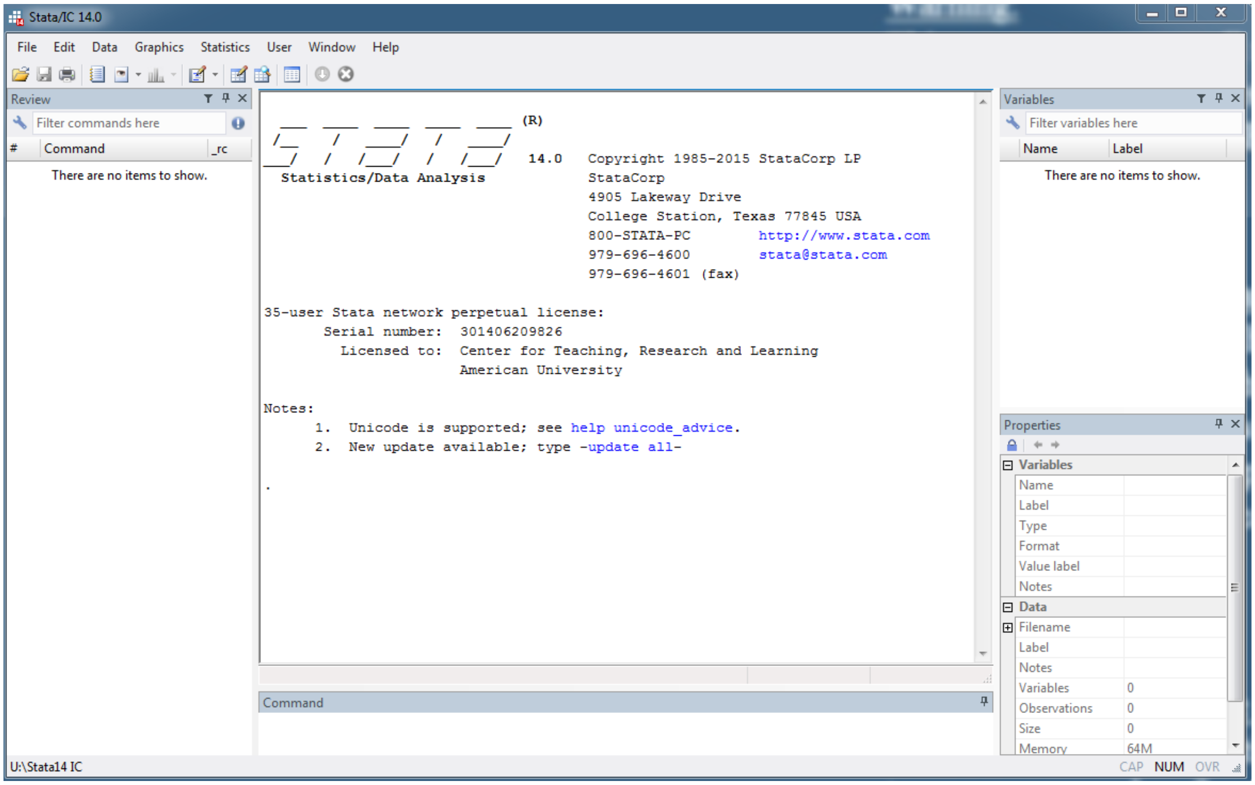
When you start Stata, you will see the default Stata interface. The interface is made up of several small windows: command, results, variables, properties, and review. This view is customizable and can be altered. You can resize, close, or change the location of these windows using *Edit/Preferences* menu.

➢ Results Panel

The large central panel is the *Results* Panel. Outputs of your analysis appear in this window. Only graphics will appear in a separate window. It can display only a limited number of lines at a time.

➢ Variables Panel

The *variables* panel, on the top right, lists all variables in the active dataset. This includes the variable name, its label, and other information. Double click on a variable, and it will appear in the command window.



There are no items to show.

```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   14.0   Copyright 1985-2015 StataCorp LP
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
                                      College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

35-user Stata network perpetual license:
       Serial number:  301406209826
         Licensed to:  Center for Teaching, Research and Learning
                       American University

Notes:
      1.  Unicode is supported; see help unicode_advice.
      2.  New update available; type -update all-

.
```

There are no items to show.

⊟ Variables
Name
Label
Type
Format
Value label
Notes
⊟ Data
⊞ Filename
Label
Notes
Variables 0
Observations 0
Size 0
Memory 64M

U:\Stata14 IC

- ➢ Properties Panel

The *Properties* panel immediately below the variables panel displays basic features of your variables and dataset including variable type, number of observations, etc.

- ➢ Review Panel

The *Review* panel logs all commands (from the command window) as they are entered. This allows you to keep track of every command you have run and run a command again if necessary. Click on an old command from the *Review* panel, and it will appear again in the command window.

- ➢ Command Panel

The window labeled *Command* is where you type the syntax to run. One way to execute Stata commands is to type them directly into the command window and hit “Enter.”

- Do-File
  We recommend that you create a do file before you start working in STATA. A Do-file allows you to record and save all of your command lines so that you can repeat those steps in the future.

  - ➢ **Do-file Editor, New Do File**
  - ➢ **Save Do File to the desktop or a USB drive.**

II. Managing Data

In this section, we cover opening and saving data and creating and labeling new variables.

- Opening Data
  - ➢ **File, Open/Import**

Example path: *C:\Users\<username>\Desktop\Workshop Data\cs.dta*
You may save copy of data as cs.dta in to practice with later.
Copy command line into Do-file

III. Descriptive Statistics

In this section, we will demonstrate how to obtain information on descriptive statistics for one or more variables.

- A full description of all variables in the data can be obtained with the *describe* command:

- ➢ ***describe***

To generate a table of basic descriptive statistics, such as the number of observations, mean, standard deviation, minimum, and maximum use the command

```
sum corrup gdppc
```

STATA displays each variable name, along with its descriptive statistics in table form. To insert these results into another document, highlight the table, right-click, and select any of *copy* options. The results can now be pasted into another document.

Other useful descriptive commands: ***codebook, tabulate or tab***

IV. Recode and Generate a new variable

Often, we are interested how the percentage change in one variable affected another variable. In those cases, we usually take the log of that variable. As an example, we will generate the log of Gross Domestic Product per capita *(gdppc).*

```
gen log_gdppc=log(gdppc)
```

Making a frequency table of a continuous variable is not very useful. If we want a good frequency table, we should recode. Look at the description for the variable (*corrup*) you want to recode.

```
tab corrup
codebook corrup
```

Because of the way corruption is measured we want to reverse it, so that higher values indicate more corruption.

```
gen corrup2 = 10 - corrup
```

We will recode this variable into three categories. Additionally, we will generate a new variable for the recoded data and provide labels for the new values.

```
recode corrup2 (0/3.3 = 1 "Low") (3.31/6.66 = 2 "Medium") (6.67/10 = 3 "High"), gen (corrup_categ)
tab corrup_categ
```

We can also assign a label to the variable itself to help us remember what it is.

```
label variable corrup_categ "Corruption Index in 3 categories"
tab corrup_categ
```

IV. Creating Basic Graph

In this section, we demonstrate how to generate a histogram and a scatter plot.
A histogram is used to show the frequency distribution of a variable.
Let’s generate a histogram using the variables we just created

```
hist gdppc
hist gdppc, freq title ("GDP per Capita Frequency") note ("Source: World Bank")
```

In addition to the histogram, we also titled the graph and noted the data source.

Suppose you are curious about the relationship between GDP per capita (gdppc) and the Corruption Index of a country (corrup2). Let’s create a scatter plot to help visualize the relationship between these two variables. We will also include a trend line and title the graph to help us understand what we are seeing.

```
twoway (scatter gdppc corrup2)
twoway (scatter gdppc corrup2) (lfit gdppc corrup2) title(“GDP per captia and
Corruption Relationship”) note("Source: World Bank")
```

As we can see, there is a negative relationship between these two variables.

We also might want to look at the average GDP per capita for each of our categorical levels of corruption. For this, we can use a bar graph.

```
graph bar (mean) gdppc, over(corrupcat)
```

V. Bivariate Analysis

T-test

- A t-test is the most basic way to compare two groups together, based on their scores on a single variable. While we divided the **Corruption Index (corrupt_categ)** into three groups, you can only compare two of those groups at a time using a t-test.
- Let’s compare the medium and high corruption countries (group 2 and 3)

```
ttest gdppc if inlist(corrup_categ,2,3), by(corrup_categ)
```

- In the Output, look for the “t”, which gives the value of the t-test, degrees of freedom, and the “p-value” or “Pr(T,t)”, which tells you the probability of getting that result. Remember that, in general, you are looking for a probability value less than .05.

Correlation

- The correlation coefficient is a measure of association between two variables. The sign indicates inverse versus directly proportional relationships. 0 would be a non-relationship, while a value of 1 would be a perfect positive relationship, and -1 would be a perfect inverse relationship.
- Let’s look at the correlation between **GDP per capita and Corruption Index**

```
pwcorr gdppc corrup2
```

- Your output will be a 2x2 table, but only one cell is relevant. Use either cell that shows the relation between your two variables.

VI. Running Linear Regressions

In this section, we show you how to run a simple regression in order to understand the relationship between GDP per capita and Corruption Index.

- ***regress** gdppc corrup2*

STATA displays the regression results in table form. The important information provided is the Adj R-squared, coef. and t/P>|t|.

VI. A few tricks that will make you more efficient

- Press Page Up to retrieve your last command; press it more than once to retrieve earlier commands
- Click on a command in the Review window and it will be pasted into the Command window for editing. Double click on a command and it will be executed again
- Click on a variable name in the Variables window and it will be pasted into the Command window at the current location of the cursor
- Press q or click on the circled-X button at the top to interrupt a command in progress (the button turns red when something is running)
- Use the Properties window to learn about your data set, the individual variables it contains, and how much memory Stata is using.

Help Command

We can use the *help* command to search for help with specific commands. For example, to search for help about the *summarize* command, type (in the command window):

- ***help** summarize*