Introduction to Stata for MRes/PhD Students

2010/2011

Exercise 1: Introduction to STATA

This class will provide you with an introduction to Stata, a computer package for manipulating and analyzing data. This will give you some first-hand experience of analyzing real data, and should help with understanding the concepts and techniques we will be looking at in lectures.

The course involves "learning by doing". You will learn how to use Stata by doing a number of different exercises. For the first and second exercise you will look at the relationship of height and wages, descriptively in the first and running more formal regressions in the second class.

1. Getting started

To open Stata, go to the windows **Start** button, then **Programs**, then **Statistics**, then **Stata**

Stata consists of a number of boxes:
Stata command is where you type the commands
Stata results is where you will see the main outputs
Review lists previous commands (if you click on a command with the mouse, it will appear in the Stata command box)
Variables lists the variables in a data set once you have opened one
In a separate window you can open the do-file editor.
Do-File Editor is where you collect the commands that are consecutively executed (i.e. one line after the other)

2. Changing directory

When you open Stata it will default to a specified directory (displayed in the bottom left hand corner above the windows start button). To change to the directory you want to work in, use the following set of commands:

cd c:	changes to c: directory
cd temp	changes from c: to c:\temp
cd c:\temp	changes to c:\temp in one go
cd ..	changes back (from c:\temp) to c: directory
mkdir stata	Makes a new folder within the c:\directory called stata

3. The first task

To keep all material for this course in one place you should create a folder on your H: space (the network space provided by the LSE which you can access from any computer at the LSE as well as from home).

a) Create a folder on your H: space (e.g. H:\ECStata)

b) Download the data for the first exercise from my website: http://personal.lse.ac.uk/lembcke/teaching.html.
c) Open Stata and open a new do file by clicking on the appropriate button or using the keyboard shortcut (ctrl+8)

4. Opening the data

Once Stata is up and running you should open the data set. You can do this either by clicking on the appropriate button/menu item or (and this is what you should do!) write down the appropriate command in the do-file editor (or the command window).

With a few exceptions, commands in Stata follow the same syntax (brackets denote optional elements).

[prefix :] command [varlist] [=exp] [if] [in] [weight] [using filename] [, options]

This looks rather confusing, but will become clearer over time. The central part of any command is the command itself. Sometimes a **prefix** precedes the command. We will make use of prefixes later in the course.

A **varlist** in Stata denotes a list of variable names (i.e. one or more variable names). The variable list can have different uses, it might specify the variables we want to use in a regression, or a variable that we want to create or variables for which we want summary statistics.

The variable list can be followed by an expression, for example when we generate a new variable **=exp** is the value that we assign to the new variable (which might be a transformation of existing data).

Both **if** and **in** are qualifiers that restrict the data a command applies to.

Sadly the command to open a data set is one of the exceptions and all you need to do is to use the **use** command and specify the filename (and the location if you haven't used the **cd** command to change to the appropriate folder)

use H:\ECStata\data-height

Stata assumes that the file is Stata data file (*.dta). Also note that if you have a space either in the folder or the name of the file, you need to use quotation marks around the whole path-filename expression.

use "H:\ECStata 2010\data-height"

The reason is that Stata interprets a space as separator of different parts of a command.

d) Employ the **use** command to open the data for this exercise.

5. A first descriptive look

Once you read the data into Stata you can see variables appearing in the lower left hand corner.

e) Use the **describe** command to read out general information about the data set. How many variables does it have? What information is available? How many observations are there?
f) Have a look at the data using either the **list** command or **browse** (opens a new window). What is the gender of the 10^{th} person in the sample? What is his/her height?
g) Find the average earnings and the standard deviation of earnings using the **summarize** command. Is the estimate for the standard deviation biased or unbiased? Find out by finding the formula used from the manual (use the **help** command to get there).
h) Can you use the **summarize** command to find the median earnings as well? If so, what are the median earnings?
i) For discrete variables **tabulate** provides simple frequencies. What is the number of part-time employees in the sample?
j) But **tabulate** has some more powerful capabilities. Use **tabulate** to create a cross-tabulation for gender and part-time work. How many men and women are working part-time? What share of men and women works part-time?
k) Find the average earnings for men and women working part- and full-time.

6. Non-Stata data

A lot of data sets are readily available in Stata format, but given the multitude of Statistical packages it is often the case that we can download data in some basic format that can be read by all statistical packages. The most common version is a delimiter separated file (e.g. separated by a comma or a tab)

l) Download and open the Excel data set "G7 less Germany pwt 90-2000.xls". Save the data as a comma separated file (csv).
m) Use **insheet** to import the data into Stata.
n) List the first 10 observations using the **in** qualifier.
o) List the observations for 1995 using the **if** qualifier.
p) List all the observations for the UK.

7. Data manipulation (the good kind)

Before we get to the nitty gritty business of data manipulation let's tidy up the data set a bit.

q) Rename the variable countryisocode to something shorter.
r) Label the variable ki, it is the share of real gdp invested
s) Save the data as a Stata data set using the **save** command.