

## EDUCATIONAL ATTAINMENT AND EARNINGS FUNCTIONS STATA COMMANDS

### STATA COMMANDS

2

#### INTRODUCTION TO STATA

Stata is an immensely powerful, resource-rich application and its manuals collectively run to several thousand pages. This document provides an alternative to the manuals that will enable you to undertake the regression exercises with the minimal investment of time.

#### Files

As with all statistical applications, you will be working with two types of file: a data file and output file.

##### Data files:

The data file is educational data.dta. As you do the exercises, the file will increase in size because you will occasionally add new variables to it, but you will never need to change its name. When you have added a new variable and wish to save the expanded file, do so with the existing name and overwrite the previous version. You cannot print this file directly because it is in a special format that no printer can recognize. If you wish to print out some of the data, you have to do this from within Stata.

##### Log files:

Output files are described in Stata as log files. You will create many of them, one for each Stata session. Assuming that you wish to preserve the output for each session, you should give each file a different name. To open a log file, click on the button with a scroll on the Stata menu bar. This is the fourth from the left. You are given a choice of type of log file at the time of opening it. The default is a "formatted log file" with extension

.smcl

. Avoid this, and choose instead a plain log file with extension .log

. Plain

.log files are ASCII (text) files and can be imported with no fuss into any word-processing package. Incidentally, having specified the name of your .log file and the directory (folder) in which it is to be located, it is natural to look for a tab with "open" or "OK" to click. There isn't one. You have to click on the button marked "save" to open the file. To close the file at the end of your session, click on the scroll icon again and choose the close option.

#### Stata windows

When you launch Stata, you will see four Windows: a command window, a results window, a variables window, and a review window.

##### Command window:

This one-line window is where you type in your instructions. You can save on typing in two ways. Instead of typing the name of a variable, you can click on its name in the variables window. Second, when you need to give a command that is similar to a previous one, you can do this by editing the previous command rather than starting from scratch. You do this by pressing the Page Up key as often as necessary to reach the previous command. You edit it and then

press the Enter key. It is easy to make mistakes when entering Stata commands and this will help to keep you sane.

Results window:

You would use the results window only if you do not intend to save your output and have not opened a log window. However, even if you have opened a log window, you need to be able to see the bottom of the results window. The reason is that if a command gives rise to more output than can be shown at once in the results window, the first window-full will be shown and the instruction `--more--` in blue letters will appear at the bottom of the results window. To see another window-full, press the space-bar, and keep doing this until all the output has been displayed. You will not be able to issue any more commands until you have done this. If the bottom of the results file is not visible, you will not be able to see the `--more--` instruction and you will think that Stata has hung up on you.

Variables window:

This contains a list of the variables in the data set.

Review window:

This lists your most recent commands.

Common commands

Here are a few commands that will be useful in the exercises:

- `reg`  
followed by a list of variable names. The first variable is regressed on the rest.

- `sum`  
followed by a list of variable names. This produces a table giving the mean, standard deviation, maximum and minimum for each variable listed.

- `tab`  
followed by one variable name. This produces a frequency distribution for the variable

- `tab`  
followed by two variable names. This produces a cross-tabulation with the first-named variable providing the rows and the second-named one providing the columns

- `gen`  
followed by an equation. This creates a new variable defined as the dependent variable of the equation.

Adding an if expression at the end of a command, for example

`if y>10`

, makes it selective as indicated. Most if conditions are straightforward, but there is one that is not: a condition which uses an `=` sign, like

`if y==10`, must repeat the `=` sign as shown. This is to distinguish between the use of `=` in equations defining variables and its use in tests for equality.

EXERCISES

#### Exercise 1 Simple regression analysis

You must download the data set in order to do this exercise.

Open a log file (see the description of the log window above) and name your log file EX1 or something similar. Stata will automatically add the extension .log if you have chosen the plain log file in the save as type box..

Open your data file.

Type in the instruction

```
reg S ASVABC
```

and press the Enter key. If you wish to do this exercise only, click on the log button, close your log file and exit from Stata. Import your log file into a word-processor (or the Windows notepad), finish the exercise, and print what you have done. You may continue with further exercises, in which case you should not close the log file until you have finished.

#### Exercise 2 Simple regression analysis

If you have just done Exercise 1 and the data file and a log file are still open, there is no need to close either. The output from Exercise 2 will be added to your log file.

```
reg EARNINGS S
```

#### Exercise 3: Multiple regression analysis

```
reg S ASVABC SM
```

```
reg S ASVABC SF
```

```
reg S ASVABC SM SF
```

Tip: After running the first regression, press the Page Up key, change SM to SF, and press the Enter key. For the third regression, press Page Up again and add SM to the list of variables. This will save some typing.

```
reg EARNINGS S EXP
```

```
reg S ASVABC SF
```

#### Exercise 4 Binary choice models

```
gen COLL=0
```

```
replace COLL=1 if S>12
```

```
reg COLL ASVABC MALE SM SF
```

```
logit COLL ASVABC MALE SM SF
```

```
probit COLL ASVABC MALE SM SF
```

```
sum ASVABC MALE SM SF
```

Calculate the marginal effects using a spreadsheet. (Stata does have a provision for calculating the marginal effects automatically, but you should do it manually at least once.)

#### Exercise 5 Sample selection bias

```
gen COLLYEAR = 0
```

```
replace COLLYEAR = S-12 if S>12
```

```
gen LGARNCL = LGARN if COLLYEAR>0

heckman LGARNCL COLLYEAR ASVABC EXP MALE ETHBLACK ETHHISP,

select (ASVABC MALE ETHBLACK ETHHISP SM SF SIBLINGS)
```