

**EDUCATIONAL
ATTAINMENT
AND
EARNINGS
FUNCTIONS

STATA COMMANDS**

**c.dougherty@lse.ac.uk
September 2005**

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 *EAEF* 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 files.

Data files: There are 22 *EAEF* data files, as explained at the beginning of the *EAEF* manual. You will be working with only one, for example EAEF07.dta (the extension .dta indicates that it is a data file). 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 a data set in order to do this exercise. Instructions are given at the beginning of the *EAEF* manual. You should download only once.

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

From now on the instructions to open the data file and to open a log file will be omitted. 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 t tests of regression coefficients and confidence intervals

Use the regression results from Exercises 1 and 2.

Exercise 4 F test of the goodness of fit of a regression equation

Use the regression result from Exercise 1.

Exercise 5 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.

Exercise 6 Multiple regression analysis

```
reg EARNINGS S EXP
```

Exercise 7 Multiple regression analysis

```
reg S ASVABC SF
predict ES, resid
reg SM ASVABC SF
predict ESM, resid
scatter ES ESM
reg ES ESM
```

Note: The graph produced by the `scatter` command does not go to the log file and has to be printed separately. If you want a graph for the log file, use `plot` instead of `scatter`. The scatter diagram that results is less accurate but it gives the general idea.

Exercise 8 Explanation of the size of a standard error

Regression results are provided.

Exercise 9 Correlated explanatory variables

```
reg S SM SF ASVAB2 ASVAB3 ASVAB4
reg S SM SF ASVABC
cor ASVAB2 ASVAB3 ASVAB4
```

Exercise 10 Correlated explanatory variables: use of a restriction

```
reg SIBLINGS SM SF
cor SM SF
gen SP = SM + SF
reg SIBLINGS SP
```

Exercise 11 F tests of the goodness of fit in a multiple regression model

Use the regression results for Exercises 5 and 6.

Exercise 12 Nonlinear regression analysis

```
gen LGEARN = ln(EARNINGS)
reg LGEARN S EXP
```

Save the data set, with the same name as before, after defining *LGEARN*. If you do not, you will have to define *LGEARN* again the next time you wish to use it.

Exercise 13 Nonlinear regression analysis

```
gen LGWT85 = ln(WEIGHT85)
gen LGHEIGHT = ln(HEIGHT)
reg LGWT85 LGHEIGHT
```

Exercise 14 Box–Cox test

```
sum LGEARN
```

In the next instruction, replace *Z* with the mean of *LGEARN* from the sum command:

```
gen EARNSTAR = EARNINGS/exp(Z)
gen LGEARNST = ln(EARNSTAR)
reg EARNSTAR S EXP
reg LGEARNST S EXP
```

You now have all you need for the Box–Cox test.

Exercise 15 Regression analysis with a dummy variable

```
reg S ASVABC SM SF MALE
```

Exercise 16 Regression analysis with a dummy variable

```
reg LGEARN S EXP MALE
```

Exercise 17 Regression analysis with multiple categories of dummy variable

```
reg S ASVABC MALE SM SF ETHBLACK ETHHISP
```

Exercise 18 Regression analysis with multiple categories of dummy variable

```
reg LG EARN S EXP MALE ETHBLACK ETHHISP
```

Exercise 19 F test of the explanatory power of a group of dummy variables

Use the regression results from Exercises 15–18.

Exercise 20 Evaluation of the effect of changing the omitted category in a regression with dummy variables

```
reg S ASVABC MALE SM SF ETHWHITE ETHHISP
```

Exercise 21 Evaluation of the effect of changing the omitted category in a regression with dummy variables

```
reg LG EARN S EXP MALE ETHWHITE ETHHISP
```

Exercise 22 Dummy variable trap

```
reg LG EARN S EXP MALE FEMALE
```

Exercise 23 Slope dummy variable

```
gen MALEASVC = MALE*ASVABC
reg S ASVABC SM SF ETHBLACK ETHHISP MALE MALEASVC
```

Exercise 24 Slope dummy variable

```
gen MALES = MALE*S
reg LG EARN S EXP ETHBLACK ETHHISP MALE MALES
```

Exercise 25 Interactive dummy variable

```
gen MALEBLAC = MALE*ETHBLACK
gen MALEHISP = MALE*ETHHISP
reg S ASVABC SM SF MALE ETHBLACK ETHHISP MALEBLAC MALEHISP
```

Exercise 26 Chow test

```
reg S ASVABC ETHBLACK ETHHISP SM SF
reg S ASVABC ETHBLACK ETHHISP SM SF if MALE==1
reg S ASVABC ETHBLACK ETHHISP SM SF if MALE==0
```

Note: The `if` condition really does need a double = sign.

Exercise 27 Chow test

```
reg LG EARN S EXP ETHBLACK ETHHISP
reg LG EARN S EXP ETHBLACK ETHHISP if MALE==1
reg LG EARN S EXP ETHBLACK ETHHISP if MALE==0
```

Exercise 28 Comparison of a Chow test with an *F* test using a full set of dummy variables

```
gen MALESM = MALE*SM
gen MALESF = MALE*SF
reg S ASVABC ETHBLACK ETHHISP SM SF MALE MALEASVC MALEBLAC
    MALEHISP MALESM MALESF
reg S ETHBLACK ETHHISP ASVABC SM SF
cor MALE MALEASVC MALEBLAC MALEHISP MALESM MALESF
```

Exercise 29 Comparison of a Chow test with *F* test using a full set of dummy variables

```
gen MALEEXP = MALE*EXP
reg LG EARN S EXP ETHBLACK ETHHISP MALE MALES MALEEXP
    MALEBLAC MALEHISP
reg LG EARN S EXP ETHBLACK ETHHISP
cor MALE MALES MALEEXP MALEBLAC MALEHISP
```

Exercise 30 Oaxaca decomposition of the difference in male and female earnings

```
reg LG EARN S EXP ETHBLACK ETHHISP if MALE==1
sum LG EARN S EXP ETHBLACK ETHHISP if MALE==1
reg LG EARN S EXP ETHBLACK ETHHISP if MALE==0
sum LG EARN S EXP ETHBLACK ETHHISP if MALE==0
```

Exercise 31 Omitted variable bias

```
reg S ASVABC SM
reg S ASVABC
reg S SM
```

Exercise 32 Omitted variable bias

```
reg LG EARN S EXP
reg LG EARN S
reg LG EARN EXP
cor S EXP
```

Exercise 33 Omitted variable bias

```
reg LG EARN S EXP MALE ETHBLACK ETHHISP
reg LG EARN S EXP MALE ETHBLACK ETHHISP ASVABC
```

Exercise 34 Redundant explanatory variable

```
reg LGFEARN S ASVABC EXP MALE ETHBLACK ETHHISP
reg LGFEARN S ASVABC EXP MALE ETHBLACK ETHHISP SIBLINGS
cor SIBLINGS S ASVABC EXP MALE ETHBLACK ETHHISP
```

Exercise 35 Proxy variables

```
gen PWE = AGE - S - 5
reg LGFEARN S ASVABC MALE ETHBLACK ETHHISP
reg LGFEARN S ASVABC MALE ETHBLACK ETHHISP PWE
reg LGFEARN S ASVABC MALE ETHBLACK ETHHISP EXP
```

Variation:

```
reg LGFEARN S ASVABC MALE ETHBLACK ETHHISP if MALE==1
reg LGFEARN S ASVABC MALE ETHBLACK ETHHISP PWE if MALE==1
reg LGFEARN S ASVABC MALE ETHBLACK ETHHISP EXP if MALE==1
reg LGFEARN S ASVABC MALE ETHBLACK ETHHISP if MALE==0
reg LGFEARN S ASVABC MALE ETHBLACK ETHHISP PWE if MALE==0
reg LGFEARN S ASVABC MALE ETHBLACK ETHHISP EXP if MALE==0
```

Exercise 36 Test of a restriction

```
gen EXPBEF = EXP - TENURE
reg LGFEARN EXP S ASVABC MALE ETHBLACK ETHHISP
reg LGFEARN EXPBEF TENURE S ASVABC MALE ETHBLACK ETHHISP
reg LGFEARN EXP S ASVABC MALE ETHBLACK ETHHISP if MALE==1
reg LGFEARN EXPBEF TENURE S ASVABC MALE ETHBLACK ETHHISP
    if MALE==1
reg LGFEARN EXP S ASVABC MALE ETHBLACK ETHHISP if MALE==0
reg LGFEARN EXPBEF TENURE S ASVABC MALE ETHBLACK ETHHISP
    if MALE==0
```

Exercise 37 t test of a restriction

```
reg LGFEARN EXP S ASVABC MALE ETHBLACK ETHHISP TENURE
reg LGFEARN EXP S ASVABC MALE ETHBLACK ETHHISP TENURE
    if MALE==1
reg LGFEARN EXP S ASVABC MALE ETHBLACK ETHHISP TENURE
    if MALE==0
```

Exercise 38 Heteroscedasticity

```
sort S
reg EARNINGS S ASVABC EXP MALE ETHBLACK ETHHISP in 1/214
reg EARNINGS S ASVABC EXP MALE ETHBLACK ETHHISP in 357/570
reg LGFEARN S ASVABC EXP MALE ETHBLACK ETHHISP in 1/214
reg LGFEARN S ASVABC EXP MALE ETHBLACK ETHHISP in 357/570
```

Exercise 39 Measurement error, instrumental variables estimation and Hausman test

The instrumental variables estimation command in Stata is `ivreg`. It should be followed by the dependent variable, then the explanatory variables not requiring instrumentation, then, in parentheses, the variables requiring instrumentation, followed by an = sign and a list of the instrument(s). Thus the command for an IV regression of *LGEARN* on *S*, *ASVABC*, *MALE*, *ETHBLACK*, and *ETHHISP*, with *SM* instrumenting for *ASVABC*, is

```
ivreg LGEARN S EXP MALE ETHBLACK ETHHISP (ASVABC=SM)
```

To perform the Hausman test, continue with the following commands:

```
estimates store name1
reg LGEARN S ASVABC EXP MALE ETHBLACK ETHHISP
estimates store name2
hausman name1 name2, constant
```

name1 and *name2* are names to be supplied by you to identify the IV and OLS regressions in question.

To use *SM*, *SF*, *SIBLINGS*, and *LIBRARY* as joint instruments for *ASVABC*, the first command should be

```
ivreg LGEARN S EXP MALE ETHBLACK ETHHISP (ASVABC=SM SF
SIBLINGS LIBRARY)
```

To perform the Hausman test, continue with the next three commands as before, but use different names for *name1* and *name2*.

Exercise 40 Simultaneous equations estimation

```
reg ASVABC S
ivreg ASVABC (S=SM)
```

Exercise 41 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 42 Sample selection bias

```
g COLLYEAR = 0
replace COLLYEAR = S-12 if S>12
g LGARNCL = LGARN if COLLYEAR>0
heckman LGARNCL COLLYEAR ASVABC EXP MALE ETHBLACK ETHHISP,
select(ASVABC MALE ETHBLACK ETHHISP SM SF SIBLINGS)
```