Impact Evaluation and Applied Econometrics

Spring term 2016
Markus Olapade

Assignment 1:
Stata Introduction

In this session we will get to now STATA commands that are necessary to prepare data for econometric analysis. Further, we will see how to write well structured and reproducable STATA code.

For each exercise the required commands are provided in brackets. Use also the STATA help menue to get to know the syntax of the particular commands. To accss the help menue, type the command`help xxx` into the command window. In place of `xxx` your need to put the command you are interested in. The help menue explains all commands in detail and provides examples on how to use the command.

Please, see Cameron und Trivedi (2009), "Microeconometrics Using Stata", Stata Press, for a detailed introduction to STATA.

Further online ressources can be found at:

http://www.ats.ucla.edu/stat/stata

http://leuven.economists.nl/stata/stataintro.pdf

1) Entering Data, Getting an Overview

Create a folder on your hard drive where you save both files wms1998.dta and stata1.do. We will use this folder as our working folder in STATA and all files produced during this session will be saved there.

The file wms1998.dta contains the data that we will use during this session. The file stata1.do is a do-file for the Stata Introduction. A do-file is a text file containing STATA- "Code".

Start STATA.

(a) Open the do-file stata1.do using the do-file Editor (`doedit`). Type all the following commands in the do-file. How can you run the commands using the do-file?

(b) Change the STATA working folder (`cd`). We want to use the folder that you created for this session and where you saved the files wms1998.dta and stata1.do as the working folder.

(c) Open/start a log-file in order to document your work (`log`).

(d) Load the data wms1998.dta (`use`).

(e) How many observations does the data set contain (`describe`)?

(f) According to which variables (key variables) are the observations sorted (`describe`)?

(g) Sort the data according to households and individuals (`sort`).

(h) Which individuals were interviewed from the first household (`browse`)? (The first household is the household appearing in the first line according to the current sorting.)

(i) Generate a global macro- variable, `hh_key`, which contains the key variables and in addition the variables `hhld_id` (`global`).

(j) Generate a local macro-variable, `ind_key`, which contains the variable `indiv_id` in addition to the variables included in `hh_key` (`local`). Where is the difference between a local and a golbal macro variable? Make sure you now how to use these macros.

2) Data Management

The data set has been manipulated, in order to give you an idea what you need to pay attention for.

(a) Consider the mean-, minimal-, and maximum values of all variables (`summarize`). Are there any problems?

(b) The variable `read_wri` tells us whether the respondent can read and write. the command `summarize` Befehl showed that the maximum value of the variable is 9. Tabulate the variable `read_wri` with and without value lables (`tabulate`).

(c) Apparently mistakes have been made at data entry since `read_wri`=3 does not make sense. Look at all variables for the individuals who have `read_wri`=3 (`list`) and make sure that there are not more problematic values for these individuals. Debate whether you would erase this individual or replace and missing values it has for other variables (`replace`). Stata recognizes missing values if there is a cell that contains "." steht.

(d) The variable read_wri has values 1 for "yes" and 2 for "no". Since we want to use it as indicator dummy variable, we need to set "no" to 0 and "yes" to 1 (`replace`). Set "Not stated" to missing (`mvdecode`).

(e) Look at the values of variable `age` with and without lables (`tabulate`). What problem can occur if we leave values `97 or larger` and `not reported` the way they currently are?

(f) Often at encoding (i.e data entry from questionnaire to digital format) extreme values such as "99" or "98" are used to indicate that information is missing or not applicable. Use `mvdecode` in order to change `not reported` to a missing value in the age variable. Use a missing value that is recognizable by STATA and not a numeric value. Tabulate the variable `age` and look at the missing values (`tabulate ,missing`).

(g) Delete all individuals aged younger than 16 or older than 97 (`drop`). the sign "|" canbe used in STATA to indicate "or".

(h) Tabulate the variable `sex` (`tabulate`).

(i) The variable `sex` will also be used as a dummy. Set the values `sex` =0 for men and `sex` =1 for females (`replace`). You need to also adjust the value lables (`lable values`).

(j) In order to be able to properly interpret the indicator for gender you should change the name of hte variables to `female` (`rename`).

(k) The variable names `age` and `female` are unambiguous but you can add a value lable to these variables (`label variable`). Set the lable of variable `age` to `age` and for the variable `female` set the label to `female indicator`.

In the following exercise we will generate variables that are required for the later analysis.

3) Generating variables using `generate` and `egenerate`

In Stata, variables can be generate either by using commands `generate` or `egenerate`. We use the command `generate`, when a variable is created that contains information from thet same row. On the other hand, we use `egenerate` if informationen from one column is being used in the new variable. Especially when using `egenerate` one should be sure that the individuals are correctly sorted. After generating a variable use `tabulate` and/or `summarize` to verify whether the new content is what you wanted.

(a) What is the share of households with female household head (`tab`)? Generate a variable that informs on female household head (`generate`).

(b) Give a running number to all individuals in the data set (`generate` and `_n`).

(c) Givea running number to all households in the data set (`generate` and `replace`).

(d) Generate a variable which contains for each individual the average age of the household to which the individual belongs (`egenerate, by`).

(e) Generate a variable that contains the household size of the individual's household and also calculate the average household size in the sample .

(f) Close the log-file and be sure where it is aved and how you can open it again.