Hour 1: Introduction to Stata, basic operation

1. General introduction to Stata

1. What is Stata?

Stata is a general purpose statistics and econometrics software, developed by www.stata.com. Stata is a simple to use yet powerful software package.

2. Some main features:

1. data construction and manipulation
2. regression analysis
3. statistical inference
4. graphic presentation
5. user programmable advanced operation (macro programming .ado files)

3. Version, license.

- Latest version, Stata 10 with multiple threads capability. Stata MP
- For this course we use Stata student version 9.

4. Limitations.

- General purpose, for specific problem the user needs to write own program in macro style, which is less elegant comparing to other programming languages.
- Block operation mode. Every command is run on all observations. The conditional or logical expressions are evaluated on every observation. Comparing to SAS, Stata runs one command on all observations, and move to next command. In SAS, all commands are run on one observation, before move on to another observation. So SAS is disk intensive.
- Memory and CPU intensive program. The data is loaded in to system memory, which has the physical limitation of how large the datasets one can operate on. For 32bit windows system, the maximum addressable memory is 3 GB, and windows system and processes will occupy ca 1.2-1.5 GB. So the actual maximum workplace for Stata is somewhere 1.5 GB.
- No loop (in ordinary mode) and no array functions.

5. Resources.

- www.stata.com
- The Stata Reference Manual
- The Stata Journal

- Survival Analysis with Stata: Module EC968 at http://www.iser.essex.ac.uk/teaching/degree/stephenj/ec968/
- A SAS User's Guide to Stata http://www.cpc.unc.edu/services/computer/presentations/sas_to_stata/
- Google, etc

2. Basic operation

1. The user interface.

- Review, command history window,
- Variable list window,
- Results, output window,
- Command, command input window.
- Popup windows for help, and graphics.
- Menu system

2. Operational mode

- Interactive mode. Run command one by one in command window.
- Batch mode: do files. Write do files to facilitate run.

3. Variables and functions

- string,
- numerical,
- built-in (index variable _n _N, saved results, missing [. is the largest])
- functions

4. Expressions

- assignment expression: generate x=1, replace x=2
- logical expression: &, |, ~=, !=
- conditional expression: if (In Stata, if comes at the end)
- estimation commands: regress, logit
- general syntax: command *target* if in, options
- abbreviation
- comment /* */ or from * to the end of line
- carriage return / line feed
- case sensitive, commands all in small cases

5. Datasets

- Stata datasets,
- ASCII files,
- format conversion, using StatTransfer
- input,
- output,

- join, split, merge

6. Save results (datasets, logs, figures)

3. Interactive mode

- set environmental variables
- set memory 100m
- set matsize 500
- discard, clear
- Open sample data file using File menu. Open *sample_data1.dta*. This is an abstraction of real data about wage and demographic info and occupational pension choice.
- have a look of data and variable format: describe
- have a look of first 10 observations: list in 1/10
- do a summarizing statistics: summarize
- do a detailed summarizing statistics on wage: summarize wage
- do a frequency statistics on occupational pension choice: tabulate op
- generate a dummy variable sex: gen sex=1 if gender=”male”; replace sex=0 if gender=”female”
- generate a new variable log wage: gen lnwage=ln(wage)
- summarizing log wage: summarize lnwage
- save datasets using File Menu

4. Batch mode (do file) exercise1.do

Write a do file to do all those above, plus:

- define the delimiter: #delimit ;
- create a log file
- sort data set by id
- join this data set with another one *sample_data2.dta*, by id
- drop observations with missing value on wage or log wage or tax gain
- generate dummy variables for education attainment: <=9, 10-12, >= 13
- rescale tax gain (since there are negative values, we cannot use log function, we can e.g. rescale by 1/1000)
- do a OLS regression of Mincer wage equation $\ln w = \beta_0 + \beta_1 sex + \beta_2 edu + \varepsilon$
- predict log wage using estimates: predict lnw_pred
- do a summarizing statistics of the result, compare with the original lnwage
- do a logit estimation using maximum likelihood, dependent variable is the op choice, and independent variables are the log wage, sex, edu, tax gain.