Introduction to Stata - Session 1

Siv-Elisabeth Skjelbred

ECON 3150/4150, UiO

January 20, 2016

Preparation

Before we start

- Sit in teams of two
- Download the file auto.dta from the course homepage
- Save the file in a new folder "statacourse" in your home directory (e.g. in your Documents folder)
- Go to kiosk.uio.no (using Internet Explorer) and log on using your UiO user name
- Navigate to Analyse
- Open an available Stata version (the newest available)

Aim with stata sessions

Challenges:

- We will start learning a tool that you don't know what you need for yet.
- We have only three double hours, which means that learning STATA requires effort on your own time in addition.

Advantages:

- Stata is partly intuitively built so it is not as hard as it looks.
- Many things can be figured out through trial and error inside Stata.
- There is a ton of help on Google.

Aim:

- Equip you for using Stata for solving seminar exercises.
- Make you interested in learning more so that you can use Stata in later work.

Outline of the course

Session 1 The basics of stata, reading data, stata workflow.

Session 2 Working with data, do-files.

Session 3 Merging and reshaping data sets, drawing graphs.

Outline of this session

- What do we want? Why Stata?
- Quick start: Your first interactive session
- The basics of Stata
- Reading data

Tasks we want to perform

1. Data management
   - Create a new data set.
   - Merge different data sets.
   - Label and structure variables.
2. Data manipulation
   - Create new variables from existing.
   - Sort observations.
   - Change order of variables.
3. Data analysis
   - Graphs, tables, ...
   - Summarize separately, mean, count variation, ...
   - Summarize jointly, correlations, regressions, inference, ...

Why not use spreadsheet

Excel may be useful for presenting data, inputting data and does allow you to do data management, manipulations and many types of analysis but:

- it is easy to make typographical errors and there are no protection against it. Difficult to check formulaes
- it is impossible to backtrack data manipulation. Provide no audit trail so others cannot easily control your work.
- possible truncation data values, (data that is truncated (norsk: avkortet) at top and/or bottom.
- it is cumbersome when dealing with a large number of observations.

.. and once you get used to the graphs from Stata you will think the graphs from excel look horrible.

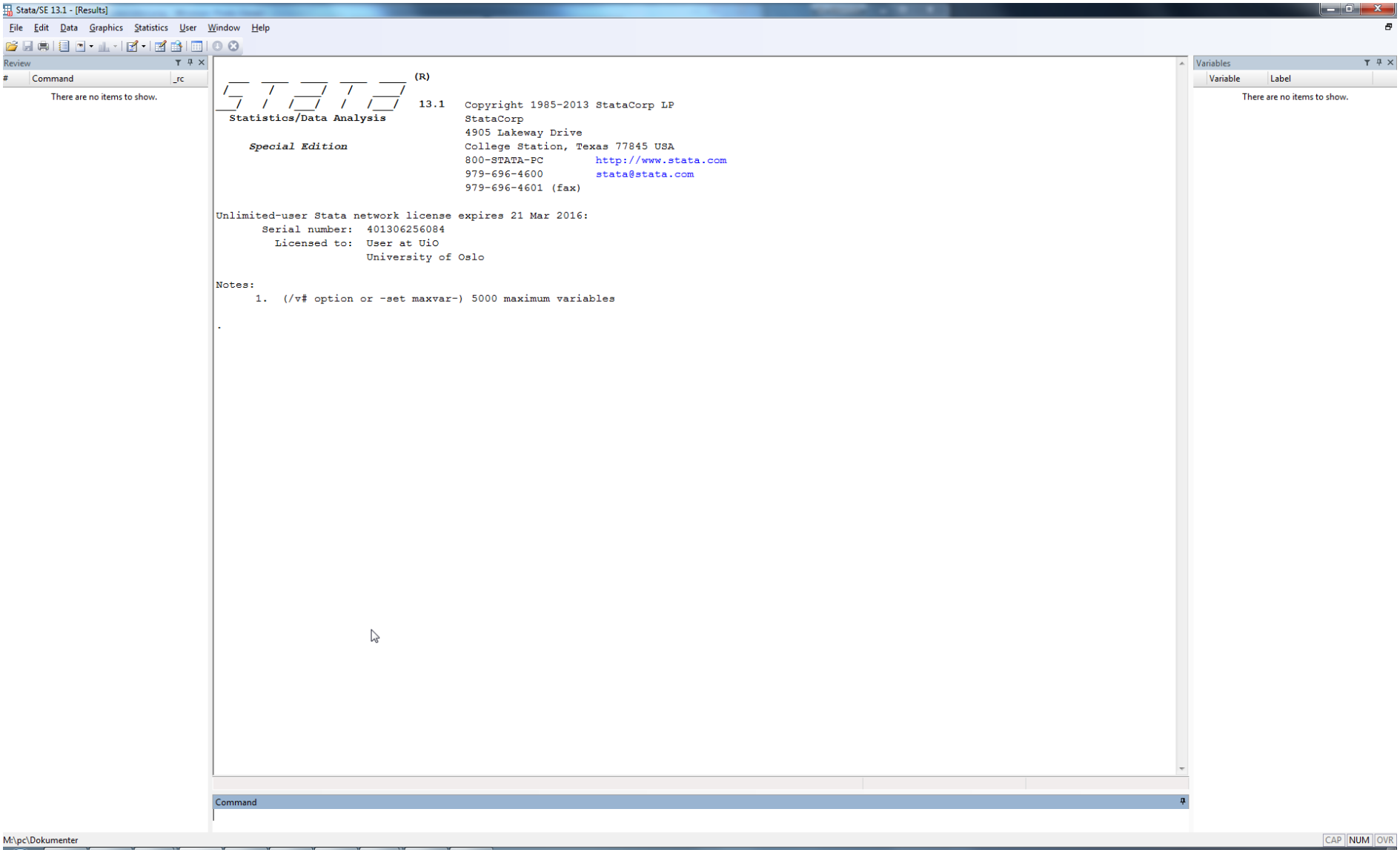
Why use Stata (or other statistical softwares)

- They allow you to log everything you do.
- You can save the actual steps you have performed so that you can control, correct and rerun your steps.
- Stata offers lots of help, tutorials and discussions available on the web.
- Stata offers lots of ready-made programs for what you may want to do so you don't have to know programming.
- Efficient in runtime and programming time.

Stata is one of the most common tools in economics and social sciences.
Alternative statistical softwares:

- R: free and popular.
- MatLab: Popular in dynamic macro, very efficient at matrix operations.
- SPSS: popular in political sciences.
- ...

Stata 101 - user interface



Working in Stata

You can either:

- Find the desired alternative in the menu
- Write the command associated with the desired alternative in the command window.
- Example: Change working directory:
  - Go to File/Change Working directory and navigate to your statacourse folder
  - Write cd "File-path"

Loading data

We can either:

- Load existing Stata data (a .dta file). → Go to File/Open
- Load data from other sources
  - Load data from excel (a .xsl file) → Go to File/Import Excel Spread sheet. Find file and mark of "Import first row as variable names." Or alternatively copy and paste to data editor.
  - Load a comma-separated file (a .csv file) → go to File/Import/Text Data. Find file. In window choose delimiter Comma (or the correct delimiter) and "OK".
- Generate a new data set.
- Use online data sets without downloading:
  - use "link adress", clear -

Loading data (2)

The following data sets are 'stored' in Stata

```
. sysuse dir
auto.dta        census.dta      network1.dta    sp500.dta       voter.dta
auto2.dta       citytemp.dta    network1a.dta   surface.dta     xtline1.dta
autornd.dta     citytemp4.dta   nlsw88.dta      tsline1.dta
bplong.dta      educ99gdp.dta   nlswide1.dta    tsline2.dta
bpwide.dta      gnp96.dta       pop2000.dta     uslifeexp.dta
cancer.dta      lifeexp.dta     sandstone.dta   uslifeexp2.dta
```

These data sets can be loaded by the command -sysuse 'filename'.
F.ex: -sysuse auto.dta, clear- loads the 1978 automobile data set.

Inputting data

We can input data by hand. Alternatives:

- Write in a do-file
- Input in the edit window
- Input in Excel and copy to Stata

Exercise

- Download and open the auto.csv data
- Insert the data to the left into Stata
- Download and open the auto.dta data

year	bh12
2001	37.7
2002	40.4
2003	43.8
2004	47.8
2005	53.9
2006	61.8
2007	69.3
2008	74.7
2009	77.1
2010	78.8
2011	79.5
2012	80.2

Study dataset

Being in the auto.dta dataset:

- The right window (variables window) shows the variables in the data set.
- Three alternatives to visually inspect the data: (browse)
  - Go to Data, Data Editor, Data Editor (Browse).The information is stored in columns$\rightarrow$ (variables) and rows$\downarrow$ (observations).
  - Press the browse data button
  - Write -browse - in the command window
- In a similar manner you can edit the data: (using the command -edit-:)
  - ONLY do this if you are constructing a new data set, or
  - if you know EXACTLY what you're doing
  - ALWAYS log your sessions of you edit something so you can backtrack.

Note: Stata uses dot as decimal indicator and comma may be used to separate the thousands.

List

- List displays the values of variables.
- If no variables are specified the values of all the variables are displayed.

Stata 101 - Study dataset

- Go to Data/Describe/describe data in memory. Press "OK".
- Data types:
  - integer - only whole numbers.
  - float: a fractional (floating point) number.
  - String: A sequence of characters.
  - Byte: Is the smallest integer type.
- The variable labels tell you what each variable measures (and in what units).

Missing values

- Missing values are shown as a dot '.' in the data set, and it represented by a large positive value.
- Thus income $>$ 100 includes those with income larger than 100 and those with missing values!!
- Most statistical commands in Stata deal with missing values by disregarding observations with one or more missing values.

Summary Statistics

To look at summary statistics (mean, obs, std. dev): go to Statistics/Summaries/Summary.../Summary Statistics. "OK". Use the auto.dta and find:

- What is the price of the five cars with a missing value for rep78 (that is rep78==.)
- Get the summary statistic for the variable price
- Get the summary statistic for the variable price if mpg is less than 21.
- Find the percentile and the median for mpg.
- Get summary statistics on mpg separately by foreign status.

Working in Stata

So far we have worked Stata using the menus.

- This is useful for graphs, typically saves time.
- It is useful to learn commands and explore what Stata can do.

Over time you should rather use commands:

- Over time it is easier to explore new things through help files, manuals or online.

Stata syntax

With a few exceptions, the basic language syntax in Stata is:
command [varlist] [if] [, options]
where [..] indicate optional elements. Example:

- *summarize* provides summary results for all variables
- *summarize price* provides summary results for only the variable price.
- *summarize price if foreign==1* provides summary results of price for only foreign cars.

Each alternative can be also be inserted through the summary statistic window.

Commands

To get help on a command in Stata simply write:

- help "command"

which will open a window that explains the full syntax of the command and often includes examples.

If you don't know the name of your command but know what you want to do you can search for commands by:

- findit "keyword(s)"

which will search the keynote database and the Internet and pop-up a window with the search results. If you only want to search files, type hsearch instead.

Not all packages of commands are by default installed by stata. To install a new package write

- ssc install "package name"

Core commands

Table: Add caption

Task	Commands
getting help	help, findit, lookfor
using Sata data	use, save, append, merge
looking at data	describe, list, tabulate, summarize
preparing data	generate, replace, rename, egen, encode, sort, by, reshape, collapse, keep, drop
saving output	log
"calculator"	display

Command efficiency

- There is no need to type the complete command or variable name. You may abbreviate commands and variables as long as Stata may not become confused about what you mean i.e. the shortest string of characters that uniquely identifies the variable suffices.
- List of variables can be selected using wildcards:
  - * = zero or more characters here
  - ? = one character here
  - - = range of variables.

Ex: If you have the variables year2000, year2005, year2010 then:

- year* selects all the variables
- year200? selects year 2000 and year 2005
- y*0 selects year2000 and year2010.

NOTE: Stata is case-sensitive

Efficiency (2)

Useful keyboard commands:

PgUp	Retrieves previous command
cursors	Back and forward to go back and forth inside your command
Home/End	To get to beginning/end of your command
ESC	Delete all written in command window
Ctrl + Del	Delete to the end of line

Exercise

- Find the commands for summary statistics.
- What are their abbreviations?

Stata memory

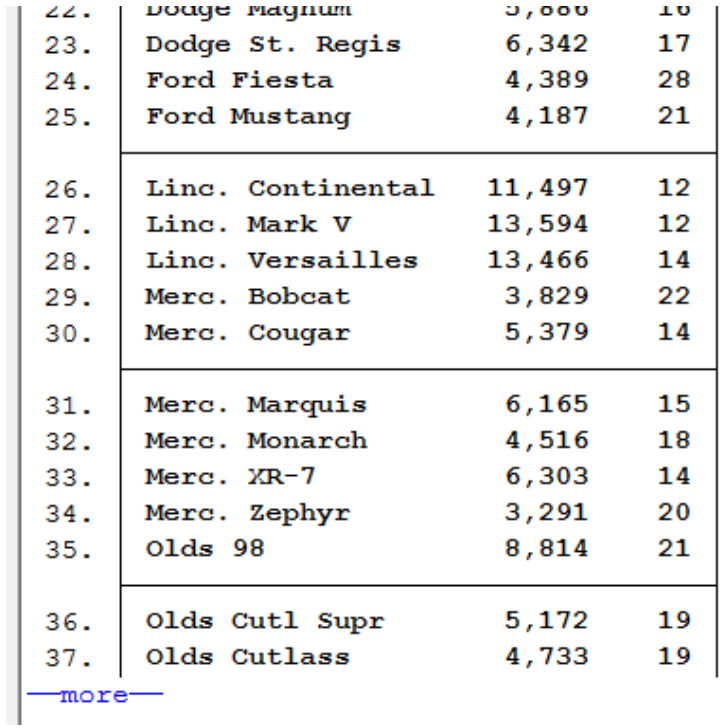
Useful commands:

- *clear* - removes data and value labels from memory
- *clear results* eliminates stored results from memory
- *clear all* remove all data, value labels, matrices, scalars, constrains, clusters, stored results... from memory.

Versions of Stata newer than 12 have automatic memory management so you don't need to think about setting memory size.

Long output

Sometimes your command will produce output longer than the window. So it will look like this:



22.	Dodge Magnum	5,886	16
23.	Dodge St. Regis	6,342	17
24.	Ford Fiesta	4,389	28
25.	Ford Mustang	4,187	21
26.	Linc. Continental	11,497	12
27.	Linc. Mark V	13,594	12
28.	Linc. Versailles	13,466	14
29.	Merc. Bobcat	3,829	22
30.	Merc. Cougar	5,379	14
31.	Merc. Marquis	6,165	15
32.	Merc. Monarch	4,516	18
33.	Merc. XR-7	6,303	14
34.	Merc. Zephyr	3,291	20
35.	Olds 98	8,814	21
36.	Olds Cutl Supr	5,172	19
37.	Olds Cutlass	4,733	19

—more—

- Pressing ⟨ Enter ⟩ show next line
- Pressing ⟨Space⟩: show next screen of output
- Typing ⟨q ⟩: breaks (i.e. ask Stata to stop what it is doing)
- If you want stata to start showing all output instead of just what can fit on one screen use command *set more off*

What you should have learned

- To load and inspect data sets.
- Stata's command syntax
- Some useful commands: help, list, summarize, display

Exercises

Explore the following commands:

Summarize	label	drop	keep
Tabulate	describe	list	count
Sort	egen/gen	regress	rename