

Research Methods for Social Sciences

Introduction to Stata 12.1

Alex Teytelboym

@t8el

Hilary 2013

Outline

1 Introduction to Stata

- Logging in and getting a grip
- Program appearance
- Stata syntax
- Task and self-study

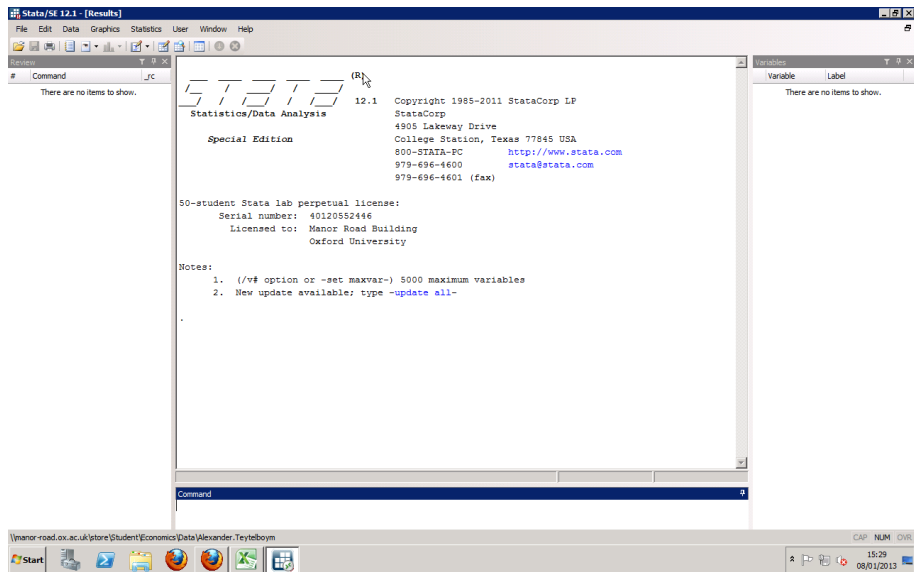
Logging in

- Access to the lab computers
 - ▶ Pick the **Manor-Road Teaching Lab** server
 - ▶ Username(s): d## (where ## is the number on your monitor)
 - ▶ Password: Forget.Me.Not

Starting Stata

- **Start -> All Programs -> Stata 12 -> StataSE 12 (64-bit)**
- Click **Connect** in the dialog box.
- **Start -> Mozilla Firefox**
- Go to **<http://t8el.tumblr.com/teaching>** and right-click on the **World Bank WDI data file** and choose "Save As..." to download this file to your **H:/ drive**.
- Let's ponder what Stata is actually good for: simple statistical analysis using a mixture of menu-based commands and intuitive syntax.
- Stata is the software of choice for many researchers and institutions.

This is what your screen should look like



The screenshot displays the Stata/SE 12.1 software interface. The main window shows the Stata logo and the following text:

```
Stata/SE 12.1
-----
Statistics/Data Analysis

Special Edition

Copyright 1985-2011 StataCorp LP
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC      http://www.stata.com
979-696-4600     stata@stata.com
979-696-4601 (fax)

50-student Stata lab perpetual license:
Serial number: 40120552446
Licensed to: Manor Road Building
Oxford University

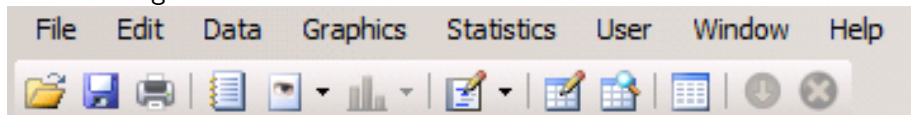
Notes:
1. (/v# option or -set maxvar-) 5000 maximum variables
2. New update available: type -update all-

.
```

The interface includes a menu bar (File, Edit, Data, Graphics, Statistics, User, Window, Help), a toolbar, a Command window (containing a '#' prompt and a cursor), and a Variables window (containing a table header with 'Variable' and 'Label' columns and the text 'There are no items to show.'). The Windows taskbar at the bottom shows the Start button, several application icons, the system tray with the date '08/01/2013' and time '15:29', and the system path '||manor-road.ox.ac.uk|store|Student|Economics|Data|Alexander.Teytelboym'.

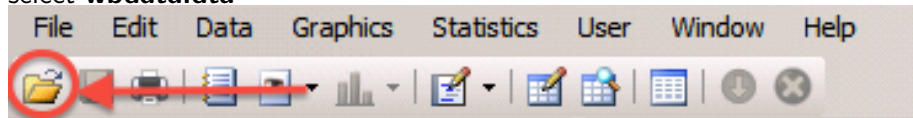
Menu bar and shortcuts

You can get a lot out of Stata by simply using the commands contained in the menu bar. For example, you can open a new Stata file by clicking on the folder sign.



Opening a file

Let's do precisely that. Click on the folder sign, go to your **H:/ drive** and select **wbdata.dta**



Command line

This is where you enter the commands for Stata to execute.

```
Command   
browse|
```

Let's try using it. Whenever it says `command` type the typewriter text into the Command line and hit Enter to execute the command.

```
browse
```


Logging our session

- We want to keep track of everything we have done in this session. First, let's define a directory...

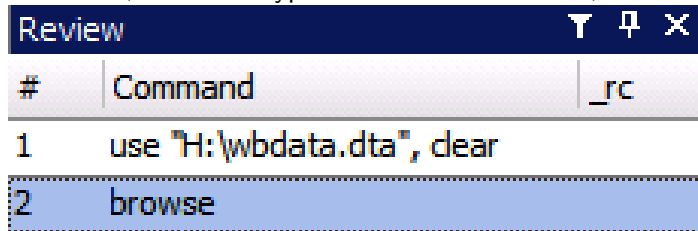
```
cd H:\
```

- ...and start a log file called class1

```
log using class1
```

Review

Every command you ask Stata to execute during a session will be stored in the Review box to the left. Our first command, for which we executed with the menu, was to open (use) the wldata.dta file. Our second command, which we typed in the command line, was browse.



```
Review [T] [M] [X]
# | Command | _rc
1 | use "H:\wldata.dta", clear
2 | browse
```

Everyone makes mistakes when they type code, but Stata will point them out every time. **Errors will be displayed in red.** Let's try this.

This is silly

```
unrecognised command: This
r(199);
```

Can you see the error in the Review box?

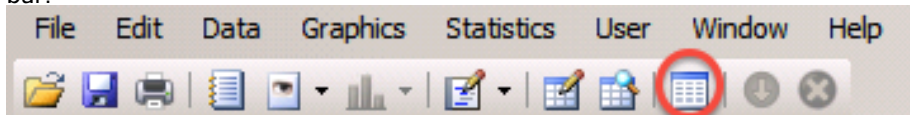
Variables

The list of variables and their labels appears on the right.

Variables	
Variable	Label
countryname	Country Name
countrycode	Country Code
year	
literacyratead...	Literacy rate, adult...
literacyratead...	Literacy rate, adult...
lifeexpectanc...	Life expectancy at ...
lifeexpectanc...	Life expectancy at ...

Variables

To look at the summary of all the variables click on the sign on the menu bar.



Data Entry

- Stata has its own data format. Stata files have **.dta** extension. Stata can import files from other formats (using the `insheet` command).
- You can enter data directly into Stata (especially if you want to play around or test various commands)

edit

- Like `browse`, this takes you to the data window, but you can now edit the data. But please restrain yourself for now!
- Data is ugly and patchy. That's just how life is, get used to it early on.

Defining the data

- To get an overview of the data

describe

- To rename variable *populationtotal* (“Population, total”) into *pop*

rename populationtotal pop

- Note that variable *year* doesn’t have a label. Let’s label *year* with “Survey year”. Can you see that the label has now appeared in the

label variable year "Survey year"

Cleaning up the data

- We have data for many years in this data set. Suppose we are only interested in year 2008.

```
keep if year==2008
```

- Note the double equality `==` sign. This is because it is a logical statement.
- Let's now look at some variables. Consider *adlit*.
- Let us see what countries we have the data for.

```
list adlit countryname
```

- Well, that was annoying. Stata listed all the countries.

Super useful tip: set more off

set more off

- You will never have to click `---more---` again!

Cleaning up the data

- Let's list the countries for which we have non-missing observations in *adlit*.

```
list adlit countryname if adlit!=.
```

- You can see there are only a few observations (for the other countries the observations are missing).
- Here !=. is a logical statement, which means “not equal to empty”.
- Let's drop all variables beginning with “literacy”.

```
drop literacy*
```

Summarizing data

- Let's look at male and female life expectancy

```
rename lifeexpectancyatbirthfemaleyears femlife  
rename lifeexpectancyatbirthmaleyears manlife  
summarize femlife manlife
```

- Suppose you only want to summarize the life expectancy for India and China.

```
summarize femlife manlife if countryname=="India" |  
countryname=="China"
```

- The `|` is command for the logical operator “or”.
- The country names are in “quotation marks”. This is because the *countryname* variable is a string variable (i.e. it is text). You can't do calculations with strings!

Sorting data

- Let's have a look at the HIV prevalence variable:

```
rename prevalenceofhivtotalofpopulation hiv
```

- Let's now sort all the countries by their HIV prevalence:

```
gsort hiv
```

- Actually, we want them in descending order:

```
gsort -hiv
```

- Now, let's look at the female life expectancy in 10 countries with the highest HIV rates:

```
list countryname femlife hiv in 1/10
```

Summary: mean, median, min, max

- What was the mean, median of GNI in our dataset?

```
su gni, detail
```

- How did I get away with that?
 - ▶ The variable name is *gnipercapitaatlasmethodcurrentus*. But this is the only variable with begins with *gni* so there can no confusion for Stata.
 - ▶ The command name is summarize. Stata allows us to abbreviate it.

Super useful tip: shortcuts

`help summarize`

- Under the heading Syntax, it says summarize. The underlined part is the shortcut. Shortcuts mean less time typing and more time checking Twitter.

Generating new variables

- So what are the largest economies in our sample? We have their GNI per capita and their populations, so we should be able to calculate the economy size.

```
generate totalgni=gni*pop
```

- Let's have a look at the result using commands we have already learned

```
gsort -totalgni  
list countryname
```

- Why wasn't *totalgni* calculated for all the countries?

Generating new variables

- Average GNI in 2008 was, in fact, \$8688. Why isn't this average the same as what we find in the data?
- Are most countries half as rich as that?
- Let's group countries by their income. First, let's generate a variable for the grouping (note the shortcut `generate`)

```
g quartile=.  
replace quartile=1 if gni<=2172  
replace quartile=2 if gni>2172 & gni<=4344  
replace quartile=3 if gni>4344 & gni<=6516  
replace quartile=4 if gni>6516 & !missing(gni)
```

- The `&` is command for the logical operator “and”.
- `!missing(gni)` means the value of *gni* is not missing. Otherwise Stata will replace all the missing values with of *quartile* 4.

Labeling values

- Let's create a label called *qualabel* for each value

```
label define qualabel 1 "Bottom" 2 "Lower Middle" 3 "Upper  
Middle" 4 "Top"
```

- And attach label *qualabel* to values in variable *quartile*.

```
label values quartile qualabel
```

- Let's have a look at the country groups

```
tab quartile
```


Generating new variables

- What is the average *unemployment* in our income groups?
- Let's generate a variable called *aveue* which will show that.
- We need to use an extension to the `generate` command, called `egen`

`bysort quartile: egen aveue=mean(unemployment)`

- Read this as: “Sort by *quartile* and for every *quartile* calculate the mean unemployment and put this into *aveue*”.

Graphs: scatter

- Picture can be better than a thousand words. Is there a relationship between female and male life expectancy?

```
scatter femlife manlife
```

- But pictures in Stata can take many arguments, which makes syntax horrendous, so I prefer to use the menu for that (or do them in Excel!).

Graphs: histogram

- Picture can be better than a thousand words. Let's look at the mean *unemployment over quartile*.

graph bar (mean) unemployment, over(quartile)

- An interesting relationship? Probably not.

Closing our session log

This has been a good session. To finish the log:

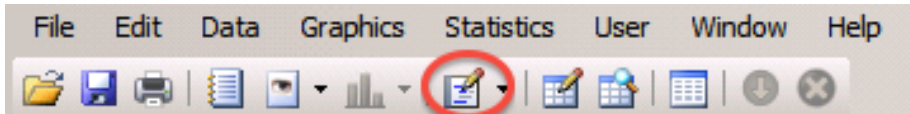
```
log close
```

To view the log:

```
view class1.smcl
```

Do files

- This hasn't been very satisfying so far because we typed commands in line by line. What if you wanted to execute a lot of code together?



- Download the **Class 1 do file** from <http://t8el.tumblr.com/teaching> into your H:/ drive.
- All the commands from this class are here: you can simply **do** the do file all over again.

do "Class 1.do"

- Or you can play with the bits of code with by selecting it and pressing the do button on the top!
- To comment text (which won't run and break the do file) use `*` or `/*` to comment blocks of text `*/`

Task

- This should run without any errors from a single do file called task.do starting with opening the wldata.dta again.
- ① Generate a variable called *diff* which is the differences between male and female life expectancy in 2007.
- ② Drop all the countries for which *diff* wasn't calculated.
- ③ Generate a variable, called *bin*, which takes value 1 if the *difference* is negative and 0 if the *difference* is positive.
- ④ Label the values "Positive" and "Negative".
- ⑤ List the countries for which the *difference* is positive.
- ⑥ Plot *diff* against HIV rates.
- Bonus: List the countries which have no missing observation for any variable.

Further study

- The best learning resource for Stata (as well as SPSS, SAS and R) can be found here: <http://www.ats.ucla.edu/stat/stata>
- **stackoverflow.com** will have many answers once you have got the hang of Stata.

Data sources

- Get your own World Bank data by clicking on DATABANK here:
<http://data.worldbank.org/data-catalog/world-development-indicators>
- UNDP website with the raw data on HDI etc. If you are feeling advanced in this class you can create your own table with as many countries and indicators as you would like.
<http://hdr.undp.org/en/statistics/data/>
- Markus Eberhardt's website is a treasure:
<http://sites.google.com/site/medevecon/development-economics/devecondata>
- Keep up to date with the latest data at:
<http://devecondata.blogspot.com/>