#### Sustainability of Public Policy Lecture 1 Introduction STATA

Rossella Iraci Capuccinello

#### Getting started in STATA

#### Start STATA

- Simply click on icon
- Stata should look like this:
  - Buttons/menu
  - $\circ$  Review window
  - Results window
  - Command
  - entry window
  - $\circ$  Variables window

# • To exit type:

exit



StataSE 10.lnk

# Getting help

- Lots of on-line help
- Click on *help* on menu
- Type *help xxx* for help on the "xxx" command

		*		
Command     Command     Help     Pelp	/	statistics/Data Analysis 10	0.0 Copyright 1984-2008 StataCorp 4905 Lakeway Drive College Station, Texas 77845   800-STATA-PC 978-666-4600 stata@stat 979-696-4601 (fax)	ISA / stata.com a.com
		Single-user Stata for Windows per Serial number: 8191052217 Licensed to: Giuseppe M	rpetual license: 76 Migali	
		/iewer (#1) [Help Contents]		
		Advice Contents What's New News	₽~	
		Тор		
		Category listings		
		Basics language syntax, expressions and fun	ctions,	
		Data management inputting, editing, creating new var	iables,	
< ]	)	Statistics summary statistics, tables, estimati	on,	
Variables		Graphics scatterplots, bar charts,		
Name Label	Type Format	Programming and matrices do-files, ado-files, Mata, matrices		
		<u>Help file listings</u>		
		Language syntax advice on what to type		
		Manual datasets download datasets from the Reference	manual s	
		Copyrights		
				.::

# Click and point in v10

- Use the menu bar to click and point to most commands
- Then fill in the boxes in the resulting dialog box
- Click on tabs for further options



### Important features

#### • NOTE

- Always use lowercase in STATA
- ren \*, lower
- More
  - When you see --more-- press the spacebar! Or type set more off
- Break
  - To stop scrolling output, hit the red cross (Ctrl-Break)
- Not enough memory
  - set mem XXXm (resizes to allow XXX mb)
  - set matsize XXX (max matrix to XXX square)
- Type set to check all Stata settings

### Loading up some data

- You will usually want to open some dataset
  - Stata expects datasets to be rectangular with columns being variables and rows being obs
- Several ways of getting data into STATA:
   use myfile (or click *file open* on the menu bar) (opens a stata format file called myfile.dta)
  - use var1 var2 var3 using myfile in 1/1000 if var4==1

(loads var1, var2, var3 for the first 1000 obs if var4=1)

insheet using myfile.csv (or .txt)

imports *csv* file which Excel can read (or "*text*" file) clear clear memory

#### Data

- Huge variety of datasets from data archives
  - <u>UKDA</u>, <u>ICPSR</u>, and around the world:
    - <a>www.esds.ac.uk/international/access/map.asp</a>
  - Research centres
    - NBER <u>www.nber.org</u>
  - Government depts and international agencies
    - US <u>Census Bureau</u> and <u>World Bank</u>
  - And journal websites (like <u>AER</u>)
    - Most major journals make datasets and code easily available

#### Stat-Transfer

🔞 🖉 🍻 👩 3 Micr... 🔹 🏣 2 Stata 🔹 🛅 2 Win... 🔹 🖉 Welcom...

- Use STAT-TRANSFER to convert data.
- Clicksattransfer.Ink
- Stat-transfer is "point and click".
- Just tell it the file name and format
- and the format you want it in.
- Click "transfer".

🛃 start

Image: Second	text in latex equation	<b>P</b> •
File Edit View Favorites Tools Help		
Google 🕞 🚽 🔽 Go 🕼 🥵 😴 🗸 😰 Bookmarks 🗕 🥸 21 blocked 🛛 🖑 Check 🗢 🔨 AutoLink 👻 🔚 AutoElli 🍙 Send to 🗸 🥖		🔘 Settings 🕶
👷 🛠 🕮 - 🍘 IPTV Technical Trial at Lanca 🕡 Welcome to Tabbed Brow 🗙	🏠 🔹 🔝 👘 📼 🔂 Page	• 💮 Tools • »

Stat/Transfer

🧭 WinEdt...

📱 summ

🔁 slide\_ec..

🖂 RE: NP

🔁 sftp:

Input FiceSupe: Excel

Output File Type: Stata/SE

ransfer Variables Observations Options (1) Options (2) Options (3) Options (4) Log About

• ?

Egit

▼ Browse…

Help

😝 Internei

💐 100% 🛛 🕶

EN 🔇 💽 🗞 🧐 1

Version 10
 Use Options(4) to change version

Browse...

#### Practicising

- Import Stata's own demos sysuse
  - E.g. sysuse auto
- Many datasets available at specific sites
   E.g. STATA's own site has all the demo data
- Use the *webuse* command to load the files directly into stata without copying locally *webuse auto* /\* gets *auto* from www.stata.com\*/

#### Web resources

- STATA is web-aware
  - update /\*updates v10.0 from www.stata.com\*/
- <u>Statalist</u> is an email listserv discussion group
- The Stata Journal is a refereed journal
- SSC Boston College <u>STATA Archive</u>
  - Files can be downloaded in Stata using *ssc* 
    - E.g. ssc install outreg2

Installs the *outreg2* ado file that makes tables pretty

- findit finds ado files from the web
- lookfor finds variables in your data set searching for strings among varnames and labels

### Keeping track of output

• STATA allows you to scroll back your screen

- But better to open a log file: it saves commands as well as output
- Click on *file, log, begin* . Or type

log using myoutput my commands..... list age

log close

 You can turn logging off and back on log off then log on when ready to resume log command allows the *replace* and *append*.

### Saving output

Oefault is a .smcl file extension (to "view")

You might prefer to give your own, say, .log, .txt

 then you get an ASCII file that anything can edit
 you can translate files to and from .smcl format
 click on file, log, translate and fill in the dialog box

#### It doesn't save graphs

- Copy graphs (use cut and paste or use menus)
- Output Content of the current directory, to change it type
  - o cd"C:\Documents\Course Ecmtrx\myfolder"
  - Is show the files in the directory

#### STATA's command file editor

- STATA has an editor that allows you to create *do* files
  - Enter cmds 1 per line-
  - Save the commands in a "do" file
  - Highlight commands and click the button with page (or page with text) and right arrow to "run" (or "do") commands.



#### Command .do files

- More complicated ideas can be implemented as a sequence of commands. For example:
   *list var1 var2 var3 in 1/10 if var4>=0*
  - Lists the first 10 rows of var1 to var3 for which var4≥0
- Collect commands in the editor, save .do file.
- Then type *do mycommands.do, nostop* echoes to screen, and keeps going after error

executes "silently"

• Or *run mycommands.do* 

#### Saving commands

- It is ALWAYS good practice to use a .*do* file

   easy to develop ideas and correct mistakes
- Logging your output is a good way of developing a .do file
  - since it saves the commands as well as output
- Or you can just log the commands
  - Type *cmdlog using* xxx.txt
- Then save the file for subsequent analysis
   save newfile
   save, replace take care it overwrites

# Using the data editor

- Open a datafile (eg.lfs\_3rdquart08.dta)
- Click on the icon
- Or type *edit*
- You can edit datapoints!
- Or just browse the datafile

File Edit	Data Graphics Statistics User Window	Help													
8 - 6	🗛 • 🛛 🗗 • 🔟 • 🛅 🖸		0												
eview	7				х										
A Comma	nd			_	Π		_	<u> </u>							
use "C:	\Documents\Teaching\Course Ecmtrx\Econom	etrics La	ncs\lfs_qua	rtjs01	8.d		🔲 Data Edi	tor							X
edit						Statistics/Da	Preserve	Bestore Sort		>>> (	Hide	Delete			
						Special Eo			semo[1] =	10123011					
									austa	week	uten	anta	s d d	wavefield	
							1	101230110101	quota 1	WEEK	wiyr 7	yr ur 3	1	wavinu	
							- 2	101730240102	1	1	7	3	2	4	-
						Single-user Sta Serial n	3	101830310202	1	1	8	3	3	1	F
						Licens	4	101830310203	1	1	8	3	3	1	-
							5	103730110101	1	3	7	3	1	1	F
						Notes:	6	103730110102	1	3	7	3	1	1	ŕ.
						1. (/m#	7	104830310101	1	4	8	3	3	1	ŕ.
						2. (/v# 3. New u	8	104830310102	1	4	8	3	3	1	ř.
							9	106730310101	1	6	7	3	3	1	ř.
						. use "C:∖Docum	10	106730310102	1	6	7	3	3	1	Ē
						. edit	11	107830110103	1	7	8	3	1	1	Ē.
						– preserve	12	108830110102	_1	8	8	3	1	1	Ē
							13	109830210102	7	9	8	3	2	1	
							14	110730110103	1	10	7	3	1	1	
				_			15	110830110101	1	10	8	3	1	1	
16					_		16	110830110102	1	10	8	3	1	1	
				J	>		17	111730110106	1	11	7	3	1	1	L
riables					x		18	112830310103	1	12	8	3	3	1	L.
lame	Label	Туре	Format		^		19	112830310104	1	12	8	3	3	1	L.
aseno	case identifier	str14	%14s				20	202830210101	2	2	8	3	2	1	1
uuta eek	week no. when interview took place	byte	%8.0g				21	202830210102	2	2	8	3	2	1	+
1yr	year that address first entered survey	byte	%8.0g				22	203830110102	2	3	8	3	1	1	-
rtr dd	address number on interviewer addres	byte byte	%8.0g %8.0g		_		25	203830210102	2	3	8	3	2	1	H
vavfnd	wave at which household was first found	byte	%8.0g				24	203830310102	2	3	8	3	3		-
hld	household number	byte byte	%8.0g				25	204/30110101	2	4	/	3	1		F.
efdte	reference date	str8	%8s				20	204630210101	2	4	•	2	2	1	-
efwkd	reference week day	byte	%8.0g				27	205630410102	2		7	2	4	1	F.
erwkm efwkv	reference week month reference week vear	byte int	%8.0g %8.0a				20	206730210101	2	6	7	3	2		F.
cthr	actual hours excluding overtime	do	%10				30	206830210102		6	,	3	2		F
ge geeul	age age band	byte byte	%8.0g %8.0g			Command	31	208830410101	2	8	R	3	4	1	F
geoui ges	age bands	byte	%8.0g				32	208830410102	2	8	8	3	4	1	F
ofl16	age oldest child in family under 16	byte	%8.0g				33	209730110101	2	9	7	3	1	1	F
ori19 ohl16	age oldest child in ramily under 19 age oldest child in hhld under 16	byte byte	%8.0g %8.0a				34	209730210102	2	9	7	3	2	1	F
ohl19	age oldest child in hhld under 19	byte	%8.0g				35	210830110101	2	10	8	3	1	1	Ē.
acthr	basic actual hours in main job gross pay estimate-band	byte	%8.0g %4c				20	210020210101	-	**	0		7	1	~
andg2	gross pay estimate-band (2nd job)	str4	ют5 %4s				5								<u>1</u>
andn	net pay estimate-band	str4	%45												
andn2 ountry	net pay estimate band (2nd job)	str4 byte	%4s %8.0c												
ourse	type of course	byte	%8.0g												
y01	country of birth	int	%8.0g		¥										
data					-										

#### Basic data reporting

- describe (or press F3 key)
  - -Lists the variable names and labels *describe using myfile*
  - Lists the variable names etc WITHOUT loading the data into memory (useful if the data is too big to fit)
- codebook (you can also use inspect)
  - Tells you about the means, labels, missing values etc

#### First look at the data

- *summ x1 x2* (or *summarize* or *sum* or *su*)
  - Gives you the means, std devs etc for x1 and x2
- corr x1 x2 in 1/100 if x4<0 (,cov)</li>
  - correlation coeffs (or covariances) for selected data
  - *pwcorr x1 x2 x3* does all pairwise corr coeffs
- tab x1 x2 (or tabulate)
  - gives a crosstab of x1 vs x2
  - use only if x1 and x2 are integers

#### Tabulating

- tab x1 x2 if x4==0, sum(x3)
  - gives the means of x3 for each cell of the x1 vs x2 crosstab for obs where x4=0 (note ==)
- tab x1 x2, missing
  - Includes the missing values
- tab x1 x2, nolabel
  - Uses numeric codes instead of labels
- tab x1 x2, col
  - Gives % of column instead of count
- table degree71 ethnic, c(mean age) row col – Customises the table

#### Data manipulation

Data can be renamed, recoded, and transformed:

gen logwgr=log(grsswk) gen agesq=age^2 ^ raises to the power genid = n*n* is obs # in STATA gen ylagged=y[ n-1] *replace rate=rate/100 rescale the var* replace age=25 if age==250 rename agesq agesquare

### Sorting data

- *sort* and *count* 
  - sort id
    - sorts data by person id
  - $count if id == id[\_n-1]$ 
    - counts how many unique separate personids
    - \_*n-1* is the previous observation
  - duplicates report id
    - same function count if

# Extended generate (egen)

#### egen

- Useful when you need a new variable that is the mean, median, etc. of another variable
  - for all observations or for groups of observations.
- Also useful when you need to simply number groups of observations based on some classification variables.
- Great when you have panel data

#### egen examples

#### egen sumvar1 = sum(var1)

creates sumvar1 as sum of values of var1

#### egen meanvar1= mean(var1), by(var3)

creates meanvar1 as mean of all values of var1

#### egen counter = count(id), by(company)

creates count as the number of companies with nonmissingid's

#### egen groupid = group(month year)

assigns a number to each month/year group

## Handling string variables

#### encode

- Use encode when the original var is a character var (e.g. gender is "m" or "f")
- encode command does not produce dummy variables, it just assigns numbers to each group defined by the character variable.
- In this example, gender was the original character var and sex is new numeric var: encode gender, gen(sex)
- *decode* does the opposite
- destring

#### **Dummy variables**

- gen wales\_resid=0
   replace wales\_resid=1 if uresmc==17
- gen region17=(uresmc==17)
- to create a series of dummies from a categorical var
  - tab uresmc, gen(dresid)
  - xi: sum i.uresmc
- recode uresmc (1/16=1) (18 19=2) (17=3) (20=4), gen(state\_resid)
- gen engwales\_resid= 0 if state\_resid==3
- replace engwales\_resid=1 if state\_resid==1

# Labelling

- Tedious to do for lots of variables
  - but then your output will be intelligibly labelled
  - other people will be able to understand

### Using STATA as a calculator

- *display* command
  - -dis 22/7
  - -disp log(250)
  - -di exp(3.6)
  - -*di chiprob(2,6.45)* (i.e. 2 df, deviance 6.45)
    - returns 0.398 (i.e. its significant at 5% level)
  - -display\_N
    - Returns the sample size
      - (\_N is the number of the last obs)

#### Data selection

• Organise your data with various commands:

keep if \_n<=1000 \_n is the obs number
drop x1</pre>

*drop if x2 ~=1* 

keeps only the first 1000 obs, drops x1, and drops all the observations where the variable  $x2\neq 1$  (~= is "not equal to")

#### Syntax to remember

"greater or equal", >= means & "and", means "or" means "set equal to" means = "is it equal to?" means "not equal" (or use != ) ~= means missing value means

For example keep if x1>= 1 & x2

keep if x1>= 1 & x2<=3 | x2==7 & x3 ~= .

### Functions

- Lots of functions are possible.
- See *help functions* 
  - Obvious ones like
    - log(), abs(), int(), round(), sqrt(), min(), max(), sum()
  - And many very specialised ones.
  - Statistical functions
    - distributions
  - String functions
    - Converting strings to numbers and vice versa
  - Date functions
    - Converting dates to numbers and vice versa
  - And lots more

### Merging data - 1

- file1 has *id x1 x2 x3*, file2 has *id x4 x4 x5*.
- You can merge using "key" in BOTH files (*id*)
- But you need to *sort* both files first.

use file1 if *personid* doesn't exist already gen id = nsort id sorts *file1* according to *id* variable save, replace use file 2 gen id =\_n sort id sorts *file2* according to *id* variable merge id using file1 drop if merge~=3 drops obs with any missing info save file3

### Merging data - 2

- For each row (*id*) all vars in *file1* added to corresponding row of *file2* (if there is one).
- *merge* creates a new variable, *merge* 
  - which =1 for those obs only in *file1*, =2 for those only in *file2*, and =3 for those in **both**.
- So the syntax above drops those obs that don't have data in both files

– and saves the result containing x1-x6 in *file3* 

• *append* to add more obs on the same vars.

#### Collapsing data (use with care)

- collapse converts the data in memory into a dataset of means (or sums, medians, etc.)
- This is useful when you want to provide summary info at a higher level of aggregation
  - For example, suppose a dataset contains data on individuals – say their region and whether u/e
  - To find the average u/e rates across reg type:
     collapse unemp, by(region)

leaves 1 obs for each reg and mean u/e rate.

### **Reshaping files**

- Data may be "long" but thin
  - Eg each record is a household member
  - But there are few vars say wage and hours
- Data may be "wide" but short
  - each record is a household and has lots of vars
  - (eg w1 w2 w3 hours1 hours2 hours3)

reshape long inc ue, i(id) j(year) wide to long
reshape wide inc ue, i(id) j(year) back to wide
Handy for merging data together and for panel
data

#### Some useful websites

- Stata's own resources for learning STATA
  - Stata website, Stata library, Statalist archive
  - <u>http://www.stata.com/links/resources1.html</u>
- Michigan's <u>web-based guide</u> (for SA)
- UCLA resources
  - <u>http://www.ats.ucla.edu/stat/stata/</u>
- ESDS "<u>Stata for LFS</u>"
- <u>Princeton</u>; <u>Illinois</u>; <u>Gruhn</u>