Introduction to STATA

Adrian Rohit Dass Institute of Health Policy, Management, and Evaluation Canadian Centre for Health Economics University of Toronto

September 24th, 2021

Outline

- Why use STATA?
- Reading/Cleaning data
- Regression Analysis
- Post-estimation Diagnostic Checks
- Other Topics in STATA
- Applied Example
- STATA Resources

Learning Curves of Various Software Packages



Source: https://sites.google.com/a/nyu.edu/statistical-software-guide/summary

Summary of Various Statistical Software Packages

Software	Interface*	Learning Curve	Data Manipulation	Statistical Analysis	Graphics	Specialties
SPSS	Menus & Syntax	Gradual	Moderate	Moderate Scope Low Versatility	Good	Custom Tables, ANOVA & Multivariate Analysis
Stata	Menus & Syntax	Moderate	Strong	Broad Scope Medium Versatility	Good	Panel Data, Survey Data Analysis & Multiple Imputation
SAS	Syntax	Steep	Very Strong	Very Broad Scope High Versatility	Very Good	Large Datasets, Reporting, Password Encryption & Components for Specific Fields
R	Syntax	Steep	Very Strong	Very Broad Scope High Versatility	Excellent	Packages for Graphics, Web Scraping, Machine Learning & Predictive Modeling
MATLAB	Syntax	Steep	Very Strong	Limited Scope High Versatility	Excellent	Simulations, Multidimensional Data, Image & Signal Processing

* The primary interface is bolded in the case of multiple interface types available.

Source: https://sites.google.com/a/nyu.edu/statistical-software-guide/summary

Why STATA?

- Moderate learning curve
- Widely used in economics and other social sciences
- Feature rich for analyzing various types of data (survey data, panel data, etc.)
- Wide array of free, user-written routines to expand the scope of STATA's capabilities
- Support for export of regression results to tables through packages such as "estout" (STATA 16 or older) or Tables feature (STATA 17)

Reading/Cleaning data

STATA Basics

- Contains a menu and syntax based interface
- Prior programming experience is not required, but can be helpful (especially with the syntax based .do files)
- Case sensitive, so be careful:
 I.e.
 - regress y x results will result in a successful OLS estimation (if everything else is right)
 - Regress y x results <u>will</u> in an error message



Starting a Log File

This should generally be your *first* step when using Stata

- Menu:
 - − File \rightarrow Log \rightarrow Begin:

Log		Begin
Import Export	*	Append Close Suspend
Example Datasets		Resume
Page Setup Print	ዕ ജ P ▶	View Translate

- Stata will prompt you to name the file. Pick a creative name (E.g. logfile1), then click ok
- At this point, Stata will record everything you do (importing data, running commands, regression output, etc)
- Syntax:
 - log using filename [, append replace [text|smcl] name(logname)]

Importing Data into Stata

• Menu

− File \rightarrow Import \rightarrow Choose appropriate option:

Import		Excel spreadsheet (*.xls;*.xlsx)
Export	•	Text data (delimited, *.csv,)
Example Datasets		Text data in fixed format Text data in fixed format with a dictionary
Page Setup Print	<mark>ዕ</mark> жР ▶	Unformatted text data SAS XPORT ODBC data source XML data

- .csv (Comma Separated) is a common option, but .xls (Microsoft Excel Format) and other formats are compatible too
- Syntax
 - import excel [using] filename [, import excel options]
 - For .csv files, command changes to import delimited

Importing Data into STATA (Microsoft Excel (.xls)

0	0			N. C. C.	Impo	rt Ex	kcel						
Exce	l file												
/Us	ers/	adrianr	ohitdas	s/Docur	ments/Si	ata	Tutor	ial/H1	TWT2	2.xls	Br	rowse	
Worl	kshe	et:					Cell	range					
She	eet1	A1:E21				\$	A1:8	21					
lr □ lr	npor npor	t first r t all da (showir	ow as v ta as st ng rows	ariable rings 2-21 o	names f 21)		Varia	ble ca	ase:	pres	erve		\$
rrev	ICTV.												
Prev	obs	height	weight	Gender	Age								
2	obs 1	height 5	weight 140	Gender 0	Age 13								1
2	obs 1 2	height 5 9	weight 140 157	Gender 0 0	Age 13 15								
2 3 4	obs 1 2 3	height 5 9 13	weight 140 157 205	Gender 0 0 0	Age 13 15 18								
2 3 4 5	obs 1 2 3 4	height 5 9 13 12	weight 140 157 205 198	Gender 0 0 0 0	Age 13 15 18 NA								
2 3 4 5 6	obs 1 2 3 4 5	height 5 9 13 12 10	weight 140 157 205 198 162	Gender 0 0 0 0 0	Age 13 15 18 NA 20								
2 3 4 5 6 7	obs 1 2 3 4 5 6	height 5 9 13 12 10 11	weight 140 157 205 198 162 174	Gender 0 0 0 0 0 0 1	Age 13 15 18 NA 20 25								
2 3 4 5 6 7 8	obs 1 2 3 4 5 6 7	height 5 9 13 12 10 11 8	weight 140 157 205 198 162 174 150	Gender 0 0 0 0 0 1 1	Age 13 15 18 NA 20 25 24								
2 3 4 5 6 7 8 9	obs 1 2 3 4 5 6 7 8	height 5 9 13 12 10 11 8 9	weight 140 157 205 198 162 174 150 165	Gender 0 0 0 0 0 1 1 1 1	Age 13 15 18 NA 20 25 24 13								
2 3 4 5 6 7 8 9 10	obs 1 2 3 4 5 6 7 8 9	height 5 9 13 12 10 11 8 9 10	weight 140 157 205 198 162 174 150 165 170	Gender 0 0 0 0 0 1 1 1 1 1	Age 13 15 18 NA 20 25 24 13 15								

Once happy with settings, click ok

000		Stata 13.1			E
Open Save	Print	Log Viewer Graph Do-file Editor Data Editor Data Browser	Break	Q• Search H	elp
Review	Q	Results	Q	Variables	
Review Command 1 import		Results	Q	Variables Q< Enter filter tes Name Obs height weight Gender Age Age Properties Variables Name Label Type Format Value Label Notes Variables Observations Size Memory Sorted by	▼ Q t here I I I
		A adrianrohitdass) (Documents) (Stata Tutorial) (September 13, 2014) =			

Starting off

Type describe to obtain some useful information about your dataset:

Contains dat	a						
obs:	20						
vars:	5						
size:	140						
	storage	display	value		 	 	
variable nam	ne type	format	label	variable label			
obs	byte	%10.0g		obs	 	 	
height	byte	%10.0g		height			
weight	int	%10.0g		weight			
Gender	byte	%10.0g		Gender			
Age	str2	%9s		Age			

Sorted by:

Note: dataset has changed since last saved

To look at your data, type browse

000



Filter Variables Properties Snapshots

003 height weight forder Age 1 1 5 148 Male 13 2 2 3 13 205 Male 13 3 3 33 205 Male 13 10 10 0 bit		obs[1]	1									
1 1 5 140 91/2 13 2 9 157 91/2 13 3 3 13 20 0/1 13 4 12 198 Mit 14 14 14 4 12 198 Mit 14 16 16 05	1	obs	height	weight	Gender	Age			Variables			
2 9 137 fale 15 3 3 13 285 Male 15 4 12 138 Rate 27 6 12 27 7 8 159 Feale 24 6 11 174 Feale 25 7 8 159 Feale 24 6 6 11 14 16 9 18 179 Feale 15 6 6 14 171 7 8 159 Feale 13 13 130 16 76 7 7 8 16 7 7 8 16 7 7 8 16 13 130 16 18 14 12 18 16 18 14 12 18 16 13 130 155 Feale 18 14 12 18 16 13 135 18 12 18 14 12 18 14 14 14 14 14 14 14 14 14	1	1	5	140	Male	13			Q* Enter filt	er text	here	
3 3 13 285 Mate 18 05 4 12 198 Mate 20 05 05 5 5 18 152 Mate 20 05 06 6 6 11 17 7 8 159 Ferate 25 05 06 7 7 8 159 Ferate 15 06 06 06 10 18 12 130 Ferate 28 Black text is for numeric variables 06 06 11 11 108 150 Ferate 28 0 06 06 13 13 108 155 Ferate 18 0 Nate 20 14 12 188 130 155 Ferate 18 0 Nate 20 13 13 130 155 Ferate 18 Blue text is labeled Nate 06 13 13 130 155 Mate 20 Nate 06	2	2	9	157	Male	15			Name	ur turit	Label	
4 12 198 Male M	3	3	13	205	Male	18			obs.		ohs	
5 10 152 Male 20 Magin Magin Magin Magin 6 6 11 174 Feeale 25 Feeale 26 Gender Gender 8 9 155 Feeale 13 14 12 186 Feeale 15 Black text is for numeric variables Mage Age Age 13 13 10 156 Rele 20 numeric variables Mage Age Mage Mage <th>4</th> <th>4</th> <th>12</th> <th>198</th> <th>Male</th> <th>NA</th> <th></th> <th></th> <th>A height</th> <th></th> <th>height</th> <th></th>	4	4	12	198	Male	NA			A height		height	
6 6 11 134 Feate 25 7 7 8 159 Feate 24 8 9 165 Feate 15 9 9 10 170 Feate 15 11 11 170 Feate 15 Black text is for numeric variables 13 11 11 177 Feate 16 Properties 13 13 106 Feate 16 Properties Properties 14 12 188 Mate 19 135 Feate 19 13 13 165 Feate 19 17 18 155 Feate 19 13 15 190 Feate 10 Numeric variables Numeric variables Numeric variables 14 14 15 190 Feate 10 Numeric variables Numeric variables 13 13 165 Red 10 Numeric variables Numeric variables 14 14 15 100	5	5	10	162	Male	20			weight		weight	
7 7 8 359 Preate 24 Age 9 9 18 179 Preate 13 Age 13 10 12 189 Feate 13 Total Age 13 11 11 110 179 Feate 28 Preate 28 Preate 15 14 14 12 189 Male 15 Male 15 Properties 16 Age 15 15 9 155 Feate 18 Blue text is labeled numeric variables Properties Image: State Image: State <td< th=""><th>6</th><th>6</th><th>11</th><th>174</th><th>Female</th><th>25</th><th></th><th></th><th>G Gender</th><th></th><th>Cender</th><th></th></td<>	6	6	11	174	Female	25			G Gender		Cender	
8 8 9 10 125 Feate 13 18 10 112 120 Feate 20 numeric variables 13 13 100 Feate 20 numeric variables 13 13 100 Feate 20 numeric variables 14 12 180 Mate 18 18 15 Feate 20 14 14 121 180 Mate 18 18 19 135 Feate 19 15 15 8 100 Feate 20 numeric variables Image: Control of the state Image: Contro of the state <	7	7	8	150	Female	24			Age		Age	
3 3 10 110 110 110 110 111 <th>8</th> <th>8</th> <th>q</th> <th>165</th> <th>Female</th> <th>13</th> <th></th> <th></th> <th>- Age</th> <th></th> <th>nge -</th> <th></th>	8	8	q	165	Female	13			- Age		nge -	
1 1	9	9	10	170	Female	15						
11 11 <td< th=""><th>10</th><th>10</th><th>10</th><th>100</th><th>Fonalo</th><th>10</th><th></th><th></th><th></th><th></th><th></th><th></th></td<>	10	10	10	100	Fonalo	10						
11 11 <td< th=""><th>10</th><th>10</th><th>12</th><th>100</th><th>Female</th><th>10</th><th>1</th><th>Black text is for</th><th></th><th></th><th></th><th></th></td<>	10	10	12	100	Female	10	1	Black text is for				
12 12 9 162 Nate 22 13 13 165 Hate 12 18 14 12 180 Hate 15 15 15 8 160 Fenale 18 16 15 Fenale 19 17 17 10 165 Fenale 28 19 13 165 Hate 20 numeric variables Name obs 19 13 155 Hate 20 numeric variables Name obs 28 28 11 155 Mate 20 numeric variables Name obs 19 13 165 Hate 20 Name obs 100 10 11 155 Mate 20 Name obs 100 10 100 100 100 100 100 100 10 100 100 100 100 100 100 19 13 165 Fenale 100 <	11	11	11	1/0	remate	20	n	umeric variables				
13 10 1655 Male 22 14 14 12 188 Male 15 15 15 8 160 Fenale 18 16 16 9 155 Fenale 20 18 15 190 Fenale 20 Properties Image: Constant of the const	12	12	9	162	Male			unienc variables				
14 14 12 180 Male 15 15 15 8 160 Ferale 18 16 16 9 155 Ferale 19 17 17 10 165 Ferale 20 18 18 15 190 Ferale 20 19 19 13 155 Male Male 20 20 11 155 Male Male 16 15 190 Ferale 20 Name Dis 19 19 13 155 Male 20 Name Dis 10 155 Male 20 Numeric variables Name Dis 10 155 Red text is for character variables Variables S 11 155 Red text is for character variables Size 140 100 (called string variables in Stata) Size 140 101 Interviewer Interviewer Interviewer 101 Interviewer Interviewer	13	13	10	165	Male	22						
15 16 160 Fenale 18 16 16 9 155 Fenale 19 17 17 18 155 190 Fenale 20 18 15 190 Fenale 20 Blue text is labeled numeric variables Name 05 20 20 11 155 Male 20 numeric variables Type byte Format ¥10.0g 16	14	14	12	180	Male	15						
16 16 9 155 Feeale 19 17 17 10 165 Feeale 20 18 15 190 Feeale NA Blue text is labeled 19 13 185 Male Na Blue text is labeled 10 155 Male 20 numeric variables Name obs 10 155 Male 20 numeric variables Type byte 10 155 Male 20 numeric variables Yotatables Notes 10 155 Male 20 numeric variables Yotatables 100 10 155 Male 20 numeric variables Yotatables 100 10 155 Male 20 Notes Yotatables 100 10 155 Male 20 Notes Yotatables 100 10 155 160 155 100 100 100 10 155 160 155 100 100 100	15	15	8	160	Female	18						
17 10 165 Fenale 20 18 15 190 Fenale NA 19 13 185 Male State 20 20 11 155 Nale 20 20 11 155 20 20 11 20 20 11 20 20 11 20 20 11 20 20 11 20 20 20 20 21 20 20 20 20 20 210 20 201 20 210 20 211 20 211 20 212 20 213 20 214 20 215 20 216 20 217 20 218 20 219 20 210 20 210 20 210 20 211 20 212 20 213 20 214 20 214 20 <th>16</th> <th>16</th> <th>9</th> <th>155</th> <th>Female</th> <th>19</th> <th></th> <th></th> <th></th> <th></th> <th></th> <th></th>	16	16	9	155	Female	19						
18 15 190 Female NA 19 13 185 Male Name Obs 20 20 11 155 Male 20 numeric variables 7ype 0 0 0 0 0 0 10 155 Male 20 0 0 0 0 10 155 Male 20 0 0 0 0 0 10 155 Male 20 0	17	17	10	165	Female	20			Properties			(I) =
19 13 185 Male Name obs 28 28 11 155 Male 20 numeric variables 7 7 9 9 10 7 9 10 19 13 155 Male 20 numeric variables Type byte 19 10 10 10 10 10 10 10 19 10 10 10 10 10 10 10 10 10 <th>18</th> <th>18</th> <th>15</th> <th>190</th> <th>Female</th> <th>NA</th> <th></th> <th></th> <th>Variables</th> <th></th> <th></th> <th></th>	18	18	15	190	Female	NA			Variables			
20 11 155 Male 20 numeric variables Format Label Notes Value Label Notes Value Label Notes Value Value Called string variables in Stata) Notes Variables Size 140 Memory 64M Sorted by	19	19	13	185	Male	18		Blue text is labeled	Name		obs	
Type byte Format %10.0g Value label Notes VData Label Variables Sorted by Sorted by Sorted by	20	20	11	155	Male	20		numerie verieblee	Label		obs	
Format \$\$10.0g Value Label Notes VData Value Solution Called string variables in Stata) Sorted by Sorted by Gamma Solution								numeric variables	Type		byte	
Notes Notes Value Label Notes Notes Notes Notes Notes Notes Notes									Format		%10.0g	
Image: Section of the section of t									Value Las	Del		
Red text is for character variables (called string variables in Stata) Variables 5 Size 140 Memory 64M Sorted by							_		▼Data			
Image:									► Filename			
Notes Notes Variables 5 Observations 20 Size 140 Memory 64M Sorted by 64M									Label			
Red text is for character variables Variables S (called string variables in Stata) 0bservations 20 Size 140 Memory 64M Sorted by 1									Notes	- II		
Red text is for character variables Observations 20 Size 140 Memory 64M Sorted by Sorted by					1	- I.			Variables		5	
(called string variables in Stata) Size 140 Memory 64M Sorted by Image: Size 140						Red te	ext is fo	r character variables	Observat	ions	20	
Called String variables in Stata) Memory 64M Sorted by						111-	مناسعه ام		Size		140	
Sorted by						calle	ea string	g variables in Stata)	Memory		64M	
									Sorted by	1		
						1	t k		8			

Convert Character variable to Numeric

Make use of Stata's destring command:

destring [varlist] , {generate(newvarlist)|replace}
[destring_options]

Eg:

destring age, replace ignore(NA)

Sorting the Observations and Variables

- Sorting changes the order in which the observations appear. We can sort numbers, letters, etc.
- Example (ascending): sort x
 - Note: Use gsort for descending or create a negative version of x and sort
- Ordering changes the order variables in dataset appear.
- Example: order x y z

Changing Existing variables: rename

- Command: rename
- changes the name of an existing variable

Example, rename variable 'ZGMFX10A' as 'height' rename ZGMFX10A height

Working with Labels

label give descriptions to variables or data sets

- To label the dataset in memory:
- label data "National Population Health Survey"
- To label a variable:
- label var healthstat "Self-Reported Health Status"
- To label different numeric values the variable may take:
- label define vlhealthstat 1 "Excellent" 2 "Very Good" 3 "Good" 4 "Fair" 5 "Poor"
- label values healthstat vlhealthstat

Obtaining basic summary statistics

• Summarize command: Use to obtain basic summary statistics of 1 or more variables (mean, standard deviation, min, max, etc.)

summarize [varlist] [if] [in] [weight] [, options]

. summarize weight height

Variable	Obs	Mean	Std. Dev.	Min	Max
weight	20	169.4	16.32692	140	205
height	20	10.35	2.207046	5	15

• Correlate command: Creates a matrix of correlation or covariance coefficients for 2 or more variables

correlate [varlist] [if] [in] [weight] [, correlate_options]

. correlate height weight
(obs=20)

	height	weight
height weight	1.0000 0.8620	1.0000

tabulate

- command: tabulate
- Calculates and displays frequencies for one or two variables
- Syntax:
- tabulate varname [if] [in] [weight] [, options]

KEYSEX	Freq.	Percent	Cum.
Male	4,599	51.19	51.19
Female	4,385	48.81	100.00
Total	8,984	100.00	

. tab KEYSEX

More detailed descriptives

• Use tabstat command

tabstat varlist [if] [in] [weight] [, options]

tabstat earnings, s(sum)

variable	sum
earnings	6.7

 The example above calculates the sum of the variable, but you could specify other statistics as well (min, max, range, etc.). If you don't specify a particular statistic at the end, then *tabstat* will generate the mean

Changing Existing variables: replace

- Command 'replace' changes the contents of an existing variable
- Syntax:

replace oldvar = exp [if exp] [in range]

- replace can be using in many circumstances, including
- Creating binary and categorical variables
- Fixing values

Ex: Replace responses coded as "no response" (-1 in this case) with missing values

replace variable = . if variable == -1

Creating a new variable: generate

- command: generate
- Syntax:
- **generate** newvar = exp [if exp] [in range]
- Example:
- generate age_sq=age*age
- Notes:

Can type generate or gen for short

Create a Binary Variable

- To create a binary variable (0 / 1):
- Generate a variable equal to 0 for all observations
- Replace it to be 1 for selected observations

- Example, create a binary variable for people with income over \$80,000:
 - gen highinc=0
 replace highinc=1 if hh_inc>80000

Exploring Missing Values

- Missing values are given by "." in STATA
- To count the number of missing values in all variables in dataset, use user-written command tabmiss
 - To install, type <u>findit tabmiss</u> in command window
 - To use, type tabmiss
- Important Note: you can use "findit" to install other user written commands, as well as help files for commands in STATA
- Can also use **tab** var, m (one variable)

Saving data

If you've imported data into STATA from a spreadsheet, text file, etc., you may want to save it as a STATA dataset.

- This is particularly useful for large datasets, as STATA can generally read its own datasets faster than importing raw data
- Menu: go File → Save (will give you an option to replace the data if it already exists)
- Syntax: save [filename] [, save_options]

Graphing/Plotting Data

• Two-way scatter plot

twoway scatter yvar xvar

• Two-way line plot

twoway line yvar xvar

Two-way scatter plot with linear prediction from regression of y on x

twoway (scatter yvar xvar) (lfit yvar xvar)

 Two-way scatter plot with linear prediction from regression of y on x with 95% CI

twoway (scatter yvar xvar) (lfitci yvar xvar)

Regression Analysis

Fitting a Linear Model To The Data

General notation:

regress depvar [indepvars] [if] [in] [weight] [, options]

Where:

Y is our *dependent* variable X is our *independent* variable(s) Note: You may type "reg" instead of "regress"

Fitting a Linear Model To The Data

Stata Output:

. reg weight h	eight		Follc nota (reg	ows tion <i>Y X</i>)				
Source	SS	df		MS		Number of obs	=	20
Model	2762 76056	1	2762	76056		F(1, 18)	-	52.07
Pasidual	1201 02044	10	72 2	700600		P-coupred	_	0.7421
Residuar	1301.03944	10	12.2	/99000		Adi B-squared	_	0.7431
Total	5064.8	19	266.	568421		Root MSE	=	8.5018
weight	Coef.	Std.	Err.	t	P> t	[95% Conf.	In	terval]
height _cons	6.377093 103.3971	.883 9.:	7324 3421	7.22 11.07	0.000 0.000	4.520441 83.77006	8 1	.233746 23.0241

Post Estimation

Post Estimation

• Obtaining residuals

predict residuals, residuals

NB: The "residuals" after predict is just the name you want to give to the residuals. You can change this if you want to

Obtaining fitted values
 predict fittedvalues, xb

Residual Diagnostic and Heteroskedasticity testing

- OLS regression assumes homoskedasticity for valid hypothesis testing. We can test for this after running a regression
- Examine residual pattern from the residual plot

rvfplot, yline(0)

Heteroskedasticity test
 estat hettest

RVF Plot



Test for Heteroskedasticity

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of VOL
chi2(1) = 171.05
Prob > chi2 = 0.0000

Reject the null (no heteroskedasticity) in favour of the alternative (there is heteroskedasticity of some form).

Linearity testing

- OLS assumes a linear relationship between the Y and X's. We can test for this after a regression:
- Command:

acprplot var, lowess

ACPRPLOT Stata



Testing for multicollinearity

OLS regression assumption: independent variables are not too strongly *collinear*

Detection:

Correlation matrix

correlate varlist (before regression)

Variance Inflation Factor
 vif (after regression)

Specification testing

- To see if there is omitted variables from the model, or if our model is miss-specified
- Syntax: estat ovtest

. estat ovtest

Ramsey RESET test using powers of the fitted values of crime Ho: model has no omitted variables F(3, 44) = 6.45Prob > F = 0.0010

Standard Errors

- Heteroskedasticity-robust standard errors
 regress y x₁ x₂...x_n, vce(robust)
- Cluster robust standard errors

– regress y x₁ x₂...x_n, vce(cluster *clusterid*)

• Bootstrapped standard errors

– regress y x₁ x₂...x_n, vce(bootstrap)

Storing Estimation Results

 STATA can store the results of your regression via the estimates command:

estimates store name

- This can be very useful in analyzing regression results after running multiple models
- estout package (needs to be installed) can be used to create tables from the regression results that can be exported from STATA. To install, type: ssc install estout, replace

http://repec.org/bocode/e/estout/esttab.html

Other Topics in STATA

Regression commands for other types of outcome variables

- Binary outcomes: probit or logit
 (help probit; help probit postestimation)
 (help logit; help logit postestimation)
- Ordered discrete outcomes: oprobit or ologit (help oprobit; help oprobit postestimation)
 (help ologit; help ologit postestimation)
- Categorical outcomes: mprobit or mlogit (help mprobit; help mprobit postestimation)
 (help mlogit; help mlogit postestimation)

Panel Data Econometrics

• Pooled Linear Regression

regress depvar [indepvars] [if] [in] [weight] [, options]

Random Effects

xtreg depvar [indepvars] [if] [in] [, re RE_options]

• Fixed Effects

xtreg depvar [indepvars] [if] [in] [weight] , fe [FE_options]

Working With Do-Files

Motivation

Why bother?

- We can ovoid tediously running the same set of commands over and over again through the menu/command window
- Creates a document listing *all* the commands we've run
- 3) Increases our productivity with STATA!

How to get to do file editor:

• File \rightarrow New \rightarrow Do-file

File Edit View Data	Graphics St	tatistics Use	r Windo	
New	Do-file		жN	
Open	жо	Project	企業N	
Open Recent Open Recent Do-files	*	Tab	жт	

• Or "Do-file Editor" button at top (depending on which version of STATA you have)





Clean_panel

• •	0					
P	E			1	91%	•
Open	Save	Print	Find	Show	Zoom	

	1	2
	N	
D	Ē	
	-	

Open	Save Print Find Show Zoom Do
	Clean_panel +
	lear mport excel "/Users/adrianrohitdass/Documents/Stata Tutorial/HTWT1 copy 2.xls", sheet("Sheet1") firstrow
	<pre>/Rename Variables ename R0000100 PUBID ename R0536300 KEYSEX ename R0536401 KEYBDATE_M ename R0536401 KEYBDATE_Y ename R0536401 KEYBDATE_Y ename R1482500 KEYPACE_ETHNICITY ename R1482500 smoke_1998 ename R2189400 smoke_1998 ename R3563300 income_1998 ename R3568600 smoke_2000 ename R3884900 income_2000 ename R5464100 income_2000 ename R5464100 income_2000 ename T6650500 VERSION_R15 ////////////////////////////////////</pre>
	eshape long smoke_ income_, i(PUBID) j(year)
	/Run regression
	eg weight height

Applied Example

- Analysis of Health Expenditure Data in Jones et al. (2013) *Chapter Three*
- The data covers the medical expenditures of US citizens aged 65 years and older who qualify for health care under Medicare.
 - Outcome of interest is total annual health care expenditures (measured in US dollars).
 - Other key variables are age, gender, household income, supplementary insurance status (insurance beyond Medicare), physical and activity limitations and the total number of chronic conditions.
- Data can be downloaded from here (mus03data.dta): <u>https://www.stata-press.com/data/musr.html</u>

Code for Applied Example

cd "/Users/adrianr/Desktop/STATA Example" /*Set working directory - Change as appropriate*/

log using "mylogfile.smcl", replace /*Create log file - extra "replace" argument saves over log file if it already exists*/

clear /*Clear memory in STATA*/

use "mus03data.dta" /*Load data in STATA*/

describe /*Describe data*/

browse /*Open window to look at dataset*/

table posexp /*Frequency table of posexp*/

drop if posexp ==0 /*Sample restriction*/

Normal Regression
regress totexp female income suppins phylim actlim totchr /*OLS Regression*/
eststo reg1/*Store regression results*/

regress totexp age female income suppins phylim actlim totchr eststo reg2

esttab reg1 reg2 using "myresults.csv", cells(b(fmt(3)star) se(par)) stats (N r2) replace /*Output regression results*/

rvfplot, yline(0) /*RVF Plot*/
graph export rvfplot.png, replace /*Save Plot*/

estat hettest /*Heteroscedasticity Test*/

Robust Regression regress totexp female income suppins phylim actlim totchr, robust eststo robust1

regress totexp age female income suppins phylim actlim totchr, robust eststo robust2

esttab robust1 robust2 using "myresultsrobust.csv", cells(b(fmt(3)star) se(par)) stats (N r2) replace

log close /*Close log file*/

STATA Resources

STATA Online Resources

• STATA manuals are freely downloadable from the above site

http://www.stata-

press.com/manuals/documentation-set/

 Typing help [topic] in the command window is also useful, but the online manuals generally contain more detail/examples

STATA Online Resources

UCLA Institute for Digital Research and Education

• List of topics and STATA resources can be found here:

http://www.ats.ucla.edu/stat/stata/webbooks/r eg/default.htm

Other STATA Resources

- Jones, A.M., Rice, N., d'Uva, T.B., Balia, S. 2013. <u>Applied</u> <u>Health Economics - Second Edition</u>, Routledge Advanced Texts in Economics and Finance. Taylor & Francis
- Cameron, A.C., Trivedi, P.K. 2010. <u>Microeconometrics</u> <u>Using Stata – Revised Edition</u>, Stata Press books.
- Allison, P.D. 2009. <u>Fixed Effects Regression Models</u>, Quantitative Applications in the Social Sciences. SAGE Publications.
- Wooldridge, J. M. (2010). <u>Econometric analysis of cross</u> section and panel data. MIT press
 - Solutions manual (sold separately) contains STATA code and output

Useful sites to find and download Canadian data

 Ontario Data Documentation, Extraction Service and Infrastructure (ODESI) website:

http://search2.odesi.ca/

 Computing in the Humanities and Social Sciences (CHASS) at U of T

http://www.chass.utoronto.ca

Thanks for Listening

Good luck with STATA!