Intro to Stata

Michele Claibourn, StatLab

Research Data Services
University of Virginia Library
`data.library.virginia.edu`

September 16, 2014

Getting to Know Stata

Stata

- A full-featured statistical programming language
- For Windows, Mac OS X, Unix/Linux
- Official website: `http://www.stata.com/`

Strengths

- Data manipulation
- Breadth and depth of statistics
- Graphics
- Extensibility (Stata, SSC archive)

Available

- Through UVa Hive
- At a discount through the Stata GradPlan (`http://www.stata.com/order/new/edu/gradplans/`)

Navigating Stata

When you open Stata, five windows are immediately visible

- Command: Where you type in the commands to make Stata go
- Results: Where your results are displayed
- Review: A running list of all commands youve used in the order in which youve used them
- Variables: A list of all variables in your dataset
- Properties: Review variable names, labels, value labels, notes, formats, storage types

Interacting with Stata: traditionally a command-line driven language

- Pull-down menus
- Type commands in the command window
- Type a list of commands in a do-file and execute the do-file

---

Wait...

Why should I type commands?

1. Scientific research requires reproducible research findings
2. A record of commands makes the research reproducible (for your future self as well)
3. Facilitates performing alternate analyses (in the end, it's more efficient)

Is there anything else I should know?

- Stata is case sensitive; commands are always in lower case
- Most commands can be abbreviated
- You can cycle through previous commands with PageUp and PageDown

Do-files, log-files

Log file

A record of results; logs capture all the text printed in the results window

Do file

A user program of commands; do files reproduce any manipulations and analyses of the data

Commenting the do file

- * Stata will ignore a line that begins with an asterisk
- // Stata will ignore anything on a line that follows two consecutive slashes
- /// Stata will ignore anything on a line that follows three consecutive slashes, and will add the next line to the end of the current line
- /* */ Stata will ignore anything between the opening and closing pseudo-parens

Example Data

The Chronicle of Higher Education College Completion (`collegecompletion.chronicle.com/`).[1] Includes data for 3,800 degree-granting institutions. Key variables:

- chronname: Institution name
- level: Level of institution (4-year, 2-year)
- control: Control of institution (Public, Private not-for-profit, Private for-profit)
- grad_100_value: Percentage of first-time, full-time, degree-seeking undergraduates who complete a degree or certificate program within 100 percent of expected time (bachelor's-seeking group at 4-year institutions)
- grad_150_value: Percentage of first-time, full-time, degree-seeking undergraduates who complete a degree or certificate program within 150 percent of expected time (bachelor's-seeking group at 4-year institutions)
- student_count: Total number of undergraduates in 2010
- med_sat_value: Median estimated SAT value for incoming students
- aid_value: The average amount of student aid going to undergraduate recipients
- endow_value: End-of-year endowment value per full-time equivalent student
- pell_value: Percentage of undergraduates receiving a Pell Grant

---

[1] Supported by the Bill & Melinda Gates Foundation.

To the do-file!

- The working directory: `cd`
- Reading data: `use` and `import delimited`
- Examining data: `describe`, `browse/edit`, and `list`
- Subsetting with `if`:

Operators	
equal	==
and	&
or	|
not	! or ∼

- Manipulating data: `rename`, `recode`, `generate`, and `label`
- Exploring data: `summarize`, `tabulate`
- Saving data: `save`
- Analyzing data: `tab, oneway, regress`
- Graphing data: `histogram, boxplot` and `graph twoway`

Getting Help

- The Stata manuals rock – and they're available as PDFs, hyperlinked to the on-line help (type `help`)
- Stata listserve searchable archive (`http://www.stata.com/statalist/`)
- UCLA IDRE Resources (`http://www.ats.ucla.edu/stat/stata/`)
- Mitchell, M. 2010. *Data Management Using Stata: A Practical Handbook.* Stata Press.
- Long, J.Scott. 2009. *The Workflow of Data Analysis Using Stata.* Stata Press.
- Acock, Alan C. 2012. *A Gentle Introduction to Stata*, Revised 3rd ed. Stata Press.

Some Useful Commands

General

help : online help on a specific command
ssc : access routines from the SSC Archive
log : log output to an external file
tsset : define the time indicator for timeseries or panel data
compress : economize on space used by variables
cd : change the working directory
clear : clear memory
quietly : do not show the results of a command

Data manipulation

generate : create a new variable
replace : modify an existing variable
rename : rename variable
sort : change the sort order of the dataset
recode : recode categorical variable
drop : drop certain variables and/or observations
keep : keep only certain variables and/or observations
encode : generate numeric variable from categorical variable
destring : convert string variables to numeric
describe : describe a data set or current contents of memory

Some Useful Commands, cont.

More data manipulation

use : load a Stata data set
insheet : load a text file in tab- or comma-delimited format
save : write the contents of memory to a Stata data set
outsheet : write a text file in tab- or comma-delimited format
append : combine datasets by stacking
merge : merge datasets (one-to-one or match merge)
contract : make a dataset of frequencies
collapse : make a dataset of summary statistics

Statistical commands

tab : abbreviation for tabulate: 1- and 2-way tables
table : tables of summary statistics
summarize : descriptive statistics
correlate : correlation matrices
ttest : perform 1-, 2-sample and paired t-tests
anova : 1-, 2-, n-way analysis of variance
regress : least squares regression
logit, logistic : logit model, logistic regression
probit : binomial probit model
predict : generate fitted values, residuals, etc.
test : test linear hypotheses on parameters

Some Useful Commands, cont. further

More statistical commands

ivregress : instrumental variables regression
prais : regression with AR(1) errors
ologit, oprobit : ordered logit and probit models
mlogit : multinomial logit model
poisson : Poisson regression
heckman : selection model
arima : BoxJenkins models, regressions with ARMA errors
xtreg, (fe, re) : fixed or random effects estimator
xtlogit : panel-data logit models
xtmixed : linear mixed (multi-level) models

Graphical commands

histogram x: histogram of the x variable
twoway scatter y x: a Y vs X scatterplot
twoway line y x: a Y vs X line plot
tsline Y time: a Y vs time time-series plot
twoway area y x: a Y vs X area plot
twoway rline y x: a Y vs X range plot (hi-lo) with lines
twoway lfit y x: a Y vs X least-squares fit line
twoway lfitci y x: a Y vs X least-squares fit line with confidence intervals
twoway lowess y x: a Y vs X lowess (locally weighted smoothed) line