D-Lab Stata Cheatsheet

Chris Kennedy - University of California, Berkeley – January 2016

In Progress

1. Introduction to Stata

use "your_file.dta", clear – open a dataset.

help use – learn more about a command

clear – unload the current data from memory.

Ctrl-r – keyboard shortcut to quickly go back to a previous command.

describe - list the variables, total observations, and variable types & labels.

count – report the total number of observations in the dataset.

count if var > 10 - count how many observations meet certain criteria

summarize – show the mean, median, max, min of one or more variables.

tab my_var – show a breakdown of values for one variable.

tab my_var, **missing** – show a breakdown of values for one variable, and don't hide missing values.

tab1 my_var1 my_var2 – shows separate breakdowns for multiple variables.

tab my_var1 my_var2 – show a breakdown of values across two variables.

tab my_var1 my_var2, **chi2** – show a breakdown of values across two variables, and run a chi-square test of independence.

list – display all observations in the dataset.

list var1 var2 – display certain variables for all observations in the dataset.

gen new_var = 5 – create a new variable and set its value.

replace new_var = 4 – update the value of an existing variable.

replace new_var = 3 **if** other_var == 2 – update the value of an existing variable, for observations that meet certain criteria.

rename old_varname new_varname - change the name of a variable.

save "my_data.dta", replace – save the current data to a file, overwriting any existing file.

histogram myvar – plot a histogram of a variable.

scatter var1 var2 – scatterplot of two variables.

pwd – show the current working directory.

cd “other_directory/” – change the working directory to another directory.

set more off – disable the pause feature when showing multiple pages of output.

log using "my_log.log", replace – start a log file, and overwrite the file if it already exists.

log close – stop logging (put at the very end of your .do file).

export delimited using "my_data_export.csv", replace - create a csv text export of the current dataset.

import delimited - load a csv data file.

export excel using "my_data_export.xls", firstrow(variables) replace- create an Excel export of the current dataset.

import excel using "my_data_export.xls", firstrow clear - load an Excel data file.

label variable myvar "this is my var" - create a text description of a variable.

clonevar newvar = oldvar - create a copy of a variable, including any labels.

label define my_label 1 "Option 1" 2 "Option 2" - create a set of values and their corresponding string descriptions

label list my_label - show the values and string descriptions of a value label.

label values myvar my_label - assign an existing set of value labels (from "label define") to a variable in the dataset.

label data "this is my dataset, 2016-01-05" - provide a description of the dataset.

order var1, after(var2) - change the order of a variable in the dataset.

drop var1 var2 - remove specific variables

drop if var1 > 10 - remove observations that meet a certain criteria

keep if var1 > 10 - remove observations that don't meet a certain criteria

display "Some output" – output a message.

compress - reduce the filesize of the dataset if possible.

2. Stata Data Analysis

findit mdesc – search for a user-written command that could be installed

ssc install mdesc – install a user-written command from the Stata software archive.

mdesc - review any missing data for each variable in the dataset

gen missing_indicator = **missing**(myvar) - create an indicator/dummy for missing data in another variable.

gen missing_indicator = myvar == . - another way to do the same thing.

recode age (18/29 = 1) (30/50 = 2) (else = 3), gen(age_recoded) - recode a variable based on its values.

sort var1 var2 - re-order the dataset based on the value of one or more variables, in ascending order

gsort +var1 -var2 - re-order the dataset based on the value of one or more variables, in ascending or descending order

set seed - set the random number generator starting point

set sortseed - when sorting on a variable, ensure that ties are broken in the same random order.

gen my_order = **_n** - save an observation's order in the dataset (1, 2, 3, ..., n).

gen rand = **runiform()** - create a random number for each observation in the dataset.

reg y_var x1 x2 – fit an OLS regression.

reg y_var x1 x2, **robust** – fit an OLS and use robust standard errors.

reg y_var x1 x2, r **cluster**(village_id) – OLS with robust clustered SEs.

predict y_hat - predict y_hat after a regression

quietly - hide any output from a command

return list - show the custom values that a previous command created

ereturn list - show the custom regression-related values that a previous command created

regression vectors - _b[varname] to access beta coefficients, _se[varname] for standard errors.

logit y_var x1 x2 - fit a logistic regression

logit y_var x1 x2, **nolog** - fit a logistic regression and hide the optimization log

corr x1 x2 x3 - correlation table

ttest outcome, by(group_var) - t-test

by - operate on subsets of a sorted dataset.

bysort - operate on subsets of a dataset and sort automatically

egen - generate a new variable with advanced functions

duplicates report - check for duplicate values in a dataset

duplicates tag - record the number of duplicates for each observation

duplicates drop - remove records that are duplicated

twoway (scatter) (lfit) - create a scatterplot chart and add a linear regression line.

append - append one dataset to the currently loaded dataset.

merge - combine a dataset with the currently loaded dataset.

graph matrix var1 var2 var3 -- bivariate scatterplot matrix to visusalize correlations.

graph export

estout

outreg2

3. Stata Programming

xi - create indicator/dummy variables for a categorical variable.

missing values

factor variables

interaction terms

capture log close – close an open log file if it exists, and if not ignore the error message.

preserve - make a temporary backup of the current dataset.

restore - restore the temporary backup of the dataset.

local myvar = 1 - create a programming variable (not in the dataset) and set it to 1.

global myvar = 1 - same thing, but allow other do files to also see the variable.

foreach var in var1 var2 var { } - run certain commands seperately for each variable in a list

forvalues var in 1/10 { } - run certain commands seperately for each value in a given range

confirm numeric - check if a variable is numeric or a string.

reshape - change a dataset from wide to long format or vice versa.

round()

floor()

ceil()

set obs 100 - create a blank dataset with 100 observations

assert age >= 50 - give an error if a certain condition is not met, for debugging purposes.

display as error "Ran into an error" - output a message with color-coding

graph combine - combine two charts into a single image [chart with hist example]

4. Advanced Stata Programming

datasignature set - record a unique numeric summary of the current dataset (cryptographic hash)

datasignature confirm - check if anything in the dataset has been modified

matrix my_betas = e(b) - save the matrix result from a command, e.g. a regression.

matrix list my_betas - display a saved matrix

ds, has(type numeric) - describe variables in a dataset that are a certain type

local var_list: **list** r(varlist) – exclude_list - remove variables from a list

levelsof my_var, local(my_local) - determine how many unique values a variable has and save in a local macro

timer on 37 - start a timer and call it #37

timer off 37 - stop timer #37

timer list 37 - display how long has elapsed for timer #37

timer clear 37 - reset the timer

Show timer output in an easier to read way:

```
timer off 37 timer list 37 dis as text "Hours: " as result round(r(t37) /
3600, 0.01) dis as text "Minutes: " as result round(r(t37) / 60, 0.01) dis as
text "Seconds: " as result round(r(t37))
```

local : word count - count how many words are in a string

tempvar - create temporary variables in a dataset with unique names, useful for ado commands

tempname - create temporary local macros with unique names, useful for ado commands

Stata Resources

1. Stata book recommendations (Stata Press)
    - A Gentle Introduction to Stata (http://www.stata-press.com/books/gentle-introduction-to-stata/) - great companion for the trainings

    - The Workflow of Data Analysis using Stata (http://www.stata-press.com/books/workflow-data-analysis-stata/) - helpful tips on organizing do files, working on teams, improving productivity, and reducing errors
    - An Introduction to Stata Programming (http://www.stata-press.com/books/introduction-stata-programming/) - advanced book on Stata programming
2. External Stata tutorials
    - Stanis Koleniov's Stata tutorials (http://web.missouri.edu/~kolenikovs/stata/Duke/)
    - UCLA Statistical Consulting Group (http://www.ats.ucla.edu/stat/stata/)
    - statalist.org (http://www.statalist.org/)

Stata at Berkeley

- Citrix virtual workstation (http://ist.berkeley.edu/is/platforms/citrix)
- D-Lab computers (http://dlab.berkeley.edu/space)
- Library Data Lab computers (http://www.lib.berkeley.edu/libraries/data-lab)
- StataCorp educational discount (http://www.stata.com/order/new/edu/gradplans/student-pricing/)