(FGLS)
Chapter 6

6.2.1
6.2.2
6.2.3
6.3

(LDV)
Chapter 10

10.1.2
10.1.3
10.2
10.3
10.4
10.5.1

(Dummy)
Chapter 7

7.1.2
7.2
7.4.1

(IV)
Chapter 8

8.2
8.5
8.6
8.7.4
8.8
8.10
8.A
8.B.

AB

B.1.1
B.1.2
B.2
B.3
B.4
B.5

CREATE RECIPE / ALGORITHM of Data Sc.

STATA CHEATSHEETS

Start ' you have data sets → merging
↳ dropping vars
↳ generating vars
↳ new refined data set
↳ labelling

use .dta, clear (1)

describe (2) summarize (3)

-N (largest observation)
-n (current)
→ use

generate, replace,
(new var) (modifies)
Transform.

generate x = y*100
x = y/z
Sum x
replace x = y/z *100

gsort, sort (4)
↳ descending!
given! ascending 2
gsort y -z.
(descending)

(sort)
list
(5) l/5 first5
-5/l last5
gsort y +z
(asc.)
Sort & list x y z in l/5

gsort -z
Sort z list } gsort -z
list x y z in 1/5

if (6) } gen z = y if x > 5000
Sort /p
list /p q x z, sep(0)

missing val (7) - '.' (say failed condls)
Sum z if x > 5000.

(8) generate dummies
gen z = 0
replace z = 1 if x <= 5000
gen y = 0
replace y = 1 if x > 5000
list /p x z y

CAREFUL
gen y = 1 if x > 5000
missing created not 0.

(9) missing val as largest num in list.

gen y = (x <= 5000) if x < .

gen z = (x > 5000) if x < 0

TODAY

listen to ma'am recordings

Summary Statistics

Stata Bauer IHDS practice play

Variables

---

if for subset of data analysis

sum hh q r if x == 1

varlist and if } qualifiers

by x, sort : sum p, q r } equivalent

sum p q r if x == 1 } (10)

by varlist: - for discrete categories

x = { NE, SE }

by x, sort : sum p q r
↳ automatic

then for each x value

sum p q r if x == "NE"

(2) bg( ) - (11) (not by varlist:)

tabstat p, by(x) statistics (N mean sd min max)

by( ) eg by(region) table with sum for each region

bysort: - repeats command for each value of region.

EXAMPLE

x and y to one dummy z = 1 for x=1 and 2 for y=1

→ gen z = x + 2 * y

by z, sort: sum a b c

label (12)

label data "abc"

label variable x "population"

desc x

(13) drop and keep

drop varlist if (condn, range)

keep varlist.

(14) rename x y
(old) (new)

MB01
MB02

temp? fix MB nucleidity

(15) save xyz.dta, replace

(16) outsheet
outreg (csv)

DATA TRASFORMS

Start from the first do files.

list all do files.

- How data is transformed.

keep if x == 1

keep to q t s

bysort varlist : egen z = max(x)

* Advanced [bysort egen]

Say create a var = 1 for just one obs in each var

- bysort z : gen y1 = (_n == 1)

egen y2 = tag(z)

after new var (say using egen) - label and drop duplicates.

(13) generate dummy var z

gen z = 0

replace z = 1 if x == 1 | y == 13

(14) merge m:1 using 2data

matrix

drop _merge

(15) collapse (mean) [weight] by (varlist)

merge 1:1 using gini (varids)

We can use egen for say max, min and use to get avg.

So that create ordering.

egen a = min(inc)
egen b = max(inc)
gen r = 1 if
inc <= b
avg

- N (for count) (16)

gen people_count = _N

(17) bys varlist : gen peoplecount = _N

bys varlist new var : gen percent = (_N / peoplecount * 100

(IHDS rank 1 (low) 2 (below avg) 3 (abv avg

[3 indices Now] - index
govt index
pvt index

STATISTICAL ANALYSIS using STATA

Cheatsheet:

→ use dataset

→ _N (highest)
- _n (current)

→ generate x = y / z

→ summarize x (continuous)

→ tabulate x (categorical)

→ replace x = x / 100

→ sort x -y

→ list x y z, sep by(x)

Q1) List 5 US states w/ lowest pop'n out of all i, l(i)

→ sort pop
→ list State region pop in 1/5

Q2) List 5 States w/ highest.

→ gsort -pop #descending
→ list State region pop in 1/5.

→ gen x2 = x if y > 100

* '.' - missing

Q3) Create a dummy off of a continuous.

→ gen S = 0

→ replace S = 1 if pop <= 5000

→ gen L = 0

→ replace L = 1 if pop > 5000

Q) why not gen S = 0
gen S = 1 if pop <= 5000
gen L = 0

A) for pop > 5000 S would be set missing and not 0.

* (.) missing treated as ∞

∴ all values < . = True!

ALTERNATE

gen S = (pop <= 5000) if pop <.

gen L = (pop > 5000) if pop <.

→ replace new = 5 if old == 2.

→ recode old (2 = 5), gen(new)

STATA CHEATSHEETS

I SHORTCUTS

1) F2 - Describe data
2) ctrl + 8 - open data editor
3) ctrl D - highlighted execution of do

II SET UP

- pwd - dir
- cd PATH - dir *.dta
- log using FILE.txt, replace
- SSC install PACKAGE

III IMPORT DATA

- use FILE.dta, clear
- import excel FILE.xlsx, /*
  */ sheet("Sheet1") cellrange(A2:H11)
  firstrow

- import delimited FILE.csv, /*
*/ rowrange(2:11) colrange(1:8) varnames(1)

IV SYNTAX - Command.

- [by var:] command [var2] [=exp] [if exp]
[in range] [weight] [using FILE] [,options]

foreg

" bysort RO4: Summarize price if foreign
==0 & price <= 9000, detail "

Analysis

- bysort or just by - Application of
cmd across each
unique comb of
vars in var1(vector)

[var2] - Vector to apply command (f'n) to

[=exp] - Same output as new variable

V LOGICAL OPERATORS

- +, −, *, /, ^
- & ! | == =!= < >=

VI DATA TYPES

- missingdata (no data)
  true/false (boyte)
  String
  int long float double

- gen var2 = string(var1) → say var1 = 1 var2 = "1"
- tostring var1, gen(var2)
  decode var1, gen(var2)

VI DATA EXPLORE

- describe
- Count "count if var > 500"
- Codebook var gives overview
- Summarize or Sum var
- inspect var (histogram)
- histogram var1, var2

missing val treated as largest number (positive) thus to exclude as to whether value < ".."

levelsof var1 : unique vals of var1

- browse
- clist (compactify)

display (di) var1[4]

- list var1 var2 if var2 > 10,000
  & !missing(var2)