Intro to Data Management In Stata

Adam Okulicz-Kozaryn
Harvard-MIT Data Center (HMDC)

dataquest@help.hmdc.harvard.edu

Overview

Preliminaries/Basics

Import/Export

Labels

Variables Manipulation

10 Minutes Break

Observations

Combine Data

Best Practices

Organization

- ◇ Please do interrupt and ask questions if questions are relevant to the currrent topic or if you are lost
- ◇ For further questions there will be a Q & A after the class
- ◇ Collaboration with your neighbours is encouraged
- ◇ Slides/Exercises assume you use lab computer; If you have laptop adjust (e.g. paths) accordingly
- ◇ If you are ahead of time:
  - ▶ help others
  - ▶ experiment with commands
  - ▶ read help files

Organization cont'd

- Make comments in the code file (we will download it), not on your handouts – you are going to use code/commands, not the handouts
- Save commented code file on flash drive or email to yourself

Outline

Preliminaries/Basics

Import/Export

Labels

Variables Manipulation

10 Minutes Break

Observations

Combine Data

Best Practices

Assumptions and Disclaimers

◇ Prerequisite: our Introduction to Stata or basic knowledge of Stata
Assume: Everybody used Stata before
Assume: Everybody knows how to use do-files

◇ This is **introduction** to data management, and covers only most popular features – it does not cover all Stata data management capabilities.

General Tips

◇ Use GUI/Command Window for playing around with data only

◇ Save in Do-File everything permanent

◇ Use comments

◇ Use TAB for auto-completion of variables

◇ Press Page-UP to get previous command in Command Window

Exercise 0: Files for Today

- ◇ Find class materials `http://stathelp.iq.harvard.edu/stata_data_mgmt`
- ◇ Right-click, Save Link As, and put on C:\ drive,
  go to C:\ and unzip it: right-click, select win-zip and extract here
- ◇ There are several formats of the same data, presentation slides, handouts, exercises, and all code we will use today in the do-file
- ◇ Data we use is a subset of General Social Survey:
  `http://www.norc.org/GSS+Website/`
- ◇ e.g. income, education, gender

Outline

Preliminaries/Basics

Import/Export

Labels

Variables Manipulation

10 Minutes Break

Observations

Combine Data

Best Practices

Delimited, ASCII (text file) [Covered in Stata Intro]

- **.csv, .tab, ...** Open with text editor first to see how it looks
- Variables delimited by comma, tab, etc.
- Stata will usually figure delimiter out

Fixed Format, ASCII (text file)

- **.txt, .dat, ...** They will either tell you or open it in text editor and figure yourself
- You need a dictionary. Dictionary specifies column numbers for variables
- There are several ways to do it...
- We will use the simplest approach
- do-file

Import/Export Tips

- Use the following commands often to make sure that Stata did what you thought it did: You will be surprised how often Stata misbehaves:
- d
- sum
- edit
- exercise 1

Outline

Preliminaries/Basics

Import/Export

Labels

Variables Manipulation

10 Minutes Break

Observations

Combine Data

Best Practices

Variable Names, Labels, and Value Labels

- Variable name is... a variable name, e.g. educ
- Variable label describes variable, e.g. “Highest degree completed”
- Value label describes values that a variable takes on (output of tab and tab,nola), e.g.

  “primary school” 1

  “high school” 2

  “college or university”3
- do-file

Labels Tips

◇ Give variables short names

◇ Labels prevent confusion later and for others

◇ They automatically appear on graphs, regressions, etc.

◇ Use lookfor, especially if you have many variables

◇ exercise 2

Outline

Preliminaries/Basics

Import/Export

Labels

Variables Manipulation

10 Minutes Break

Observations

Combine Data

Best Practices

Operators

- == equal to (status quo)
- = used for assigning values
- ! = not equal to
- > greater than
- >= greater than or equal to
- & and
- | or
- replace happy=1 if(educ>10 | inc>=10) & (unemp!=1 & div!=1)

Basics [Covered in Stata Intro]

- Most standard variables manipulation (e.g. generating, transforming, and recoding variables) can be done with:
- gen and replace
- or:
- recode
- do-file

Egen

- `egen` means “extended generate”
- Powerful, difficult, and confusing
- For details: `help egen`; Examples:
- `egen max_inc=rowmax(hh_inc r_inc)`
- `egen avg_inc=mean(inc)`
- `gen dev_inc=inc-avg_inc` $(x - \bar{x})$

By, Sort, Egen

- by: will run a command by some group
- You always need to sort the group first
- So always use by sort: or in short: bys:
- bys marital: egen avgm_inc=mean(inc)
- As usual, don't forget to check if Stata did what you think it did
- do-file

Tostring/Destring is About Storage Type

◇ After running `d` in "storage type" column **str** denotes a string(word), everything else is a number

◇ Run `edit` and note colors: red is string, black is number, blue is number with label

◇ Number can be stored as a string

◇ String cannot be stored as a number

◇ From number to string

```
tostring marital, gen(m_s)
```

◇ From string to number

```
destring m_s, gen(m_n)
```

◇ do-file

Encode/Decode is About Values

- Convert string into numeric

  `encode region, gen(reg_s)`

- `decode` will replace values with labels

- **Encode/Decode is about values**
- **Tostring/Destring is about storage type**
- do-file

Missing Values

- Stata understands missing as a very big number
- For instance, if income is coded from 1 to 26 and we generate high income, this is **wrong:**

```
gen hi_inc=0
replace hi_inc=1 if inc>15
```

(1 for >15 and “.”)

- It should be:

```
gen hi_inc=.
replace hi_inc=1 if inc>15 & hi_inc<26
replace hi_inc=0 if inc>0 & hi_inc<16
```

- do-file

Tips

- Use tab, mi to see if there are any missings
- Be careful about strings
- Remember that number can be stored as string
- You cannot do algebraic manipulations on string
- Use operators – you can do anything with your data using them
- Manipulation of variables is difficult. Remember to double check what you did : tab <var> , mi and tab <var>, nola mi and/or codebook <vars>, tab(100)
- exercise 3

Outline

Preliminaries/Basics

Import/Export

Labels

Variables Manipulation

10 Minutes Break

Observations

Combine Data

Best Practices

Outline

Preliminaries/Basics

Import/Export

Labels

Variables Manipulation

10 Minutes Break

Observations

Combine Data

Best Practices

Keep/Drop

- Keep first 10 obs

  ```
  keep in 1/10
  ```

- Keep obs on condition

  ```
  keep if marital==1
  ```

- Instead of `keep` you may use `drop`

  ```
  drop if marital>1
  ```

- `keep` and `drop` also work for variables:

  ```
  drop marital
  ```

- do-file

Sort, Order

- Sort on marital's values

  ```
  sort marital
  ```

- Sort on marital's and income's values

  ```
  sort marital inc
  ```

- Make marital 1st var

  ```
  order marital
  ```

- Put vars in alphabetic order

  ```
  aorder
  ```

- do-file

_n _N

- To make operations based on row order it is useful to use _n and _N
- gen id=_n
- gen total=_N
- edit
- gen previous_id=id[_n-1]
- do-file

Collapse

- We already learned `bys:` and `egen:`

```
bys marital: gen count_marital_group=_N
bys marital: egen count_id=count(id)
```

- A similar, but more radical, is `collapse`

```
collapse inc educ, by(region)
```
(mean is default)

```
collapse (count) id, by(marital)
```

- do-file

Tips

- Both collapse and bys: egen can be used to calculate group statistics
- collapse produces new dataset with N equal number of groups
- bys: egen adds a new variable with group statistic that is constant within a group
- _n+/−<number> is useful with panel/time series data
- exercise 4

Outline

Preliminaries/Basics

Import/Export

Labels

Variables Manipulation

10 Minutes Break

Observations

Combine Data

Best Practices

Merge

- Combines variables (Same Obs)
- Let's generate some data first
- use gss.dta, clear
- gen id=_n
- keep id region
- save gss1.dta, replace (**using**)
- use gss.dta, clear
- gen id=_n
- keep id inc (**master**)
- merge id using gss1.dta, sort (combine with (**using**)
- do-file

Merge Contn'd

- After merging **always** do:
- tab _merge
- variable _merge takes on 3 values:
- **3** Obs in both datasets
- **1** Obs in master only
- **2** Obs in using only
- do-file

Append

- Combines Observations (Same Var)
- Let's generate some data first
- `use gss.dta, clear`
- `keep in 1/50`
- `save gss1.dta, replace` (**using**)
- `use gss.dta, clear`
- `keep in 51/100` (**master**)
- `append using gss1.dta` (combine with (**using**)
- do-file

Xpose, Reshape

◇ xpose interchanges Vars and Obs

◇ reshape converts wide-to-long/long-to-wide

◇ help reshape

◇ reshape long var, i(id) j(year)

◇ var is a common variable that repeats, i.e. prefix,

◇ id is always unique (eg. made by gen id=_n)

◇ year is a new variable that takes unique part from variable that repeats, i.e. suffix

◇ do-file

Reshape Example

```
use gss.dta, clear
ren inc inc1
gen inc2=2*inc1
gen id=_n
reshape long inc, i(id) j(period)
edit
```

Tips

- ◇ Can Also Merge One-To-Many

  `http://users.ox.ac.uk/~sjoh2052/datamanipulation.htm`

- ◇ After reshape and merge always make sure that you got what you expect
- ◇ reshape may be confusing – use help file !
- ◇ Let's do Exercise 5

Outline

Preliminaries/Basics

Import/Export

Labels

Variables Manipulation

10 Minutes Break

Observations

Combine Data

Best Practices

Do-files

- ◇ Have a do-file that produces final results from raw data
- ◇ Always keep raw data intact
- ◇ Then manipulate it and save again, even several times
- ◇ At the end of your project you may end up with several datasets at different levels of advancement
- ◇ Then you may begin your stata session at any level
- ◇ Still your full do-file has to produce very final results from very raw data

File organization

- Always have raw data and codebook–you will go back and forth between them
- Have one directory for the whole project–keep everything in one place
- If project is big have subdirectories
- Keep one version of your project on one drive
- Back-up at least once per week

Corectness

- ◇ Double check after every maniuplation (at least at the beginning)
- ◇ Double check the whole do-file once finished
- ◇ Use as much descriptive statistics as possible
  (1) To get more familiar with data
  (2) To avoid mistakes, e.g. age of -9
- ◇ Then you may begin your stata session at any level
- ◇ See my example on the website

Thank You !

- Please fill evaluations AND give us some comments/feedback – we do care for these classes and want to make them better
- Come to other classes we offer and tell your friends about our classes `http://www.iq.harvard.edu/statistical_software_2009_2010`

A Word From Our Sponsor !

- Institute for Quantitative Social Science http://iq.harvard.edu
- Data Collection, Management

  http://www.iq.harvard.edu/data_collection_management_analysis
- Research Computing Environment

  http://www.iq.harvard.edu/research_computing
- Computer Labs (Software, Books) http://www.iq.harvard.edu/facilities
- Training http://www.iq.harvard.edu/training