Introduction to Stata

Hsueh-Sheng Wu

CFDR Workshop Series

Fall 2010

BGSU
Center for Family and Demographic Research

Outline of Presentation

- What is Stata?
- Stata user interface
- How to use Stata to manage data
- An example of Stata command file: Data Management
- Reminder of Using Stata
- Strengths and Limitations of Stata
- Where to Find Help
- Conclusions

What Is Stata?

Stata is one of statistical software packages, like SAS, SPSS, MINITAB , or BMDP.

Similar tasks across these software:

- Data Management
- Data Analysis
- Ability to use graphs to present analysis results

Differences among these software:

- User interface
- Data format
- Efficiency in managing/analyzing data and presenting results
- Syntax rules
- Some statistical analysis may be available in one package, but not the other

Stata User Interface

- Four task windows
  - Command window: You type in the command here and press Enter to submit the command
  - Results window shows the results after commands were executed
  - Review window shows the list of executed command
  - Variables window shows the list of variables in memory

Selected Functions of 8 Drop-down Menus

File
- Open and save data, graphic, do, or log files
- Import and export data files
- Print files
- Exit Stata

Edit
- Copy and paste text, graphic, and tables
- Set preferences of Stata

Data
- Examine and change the data

Graphics
- Provide graphic presentations of the variables

Statistics
- Provide various statistical tests

User
- User-supplied Stata commands (download from Internet)

Window
- Navigate through different windows

Help
- Find solutions for Stata

Functions of 11 Buttons

- Open
- Save
- Print
- Log
- New viewer
- Bring the graph window to the front
- New do-file editor
- Data Editor
- Data Browser
- Go
- Break

Data Menus

- Describe data
- Data Editor
- Data browser
- Create and change variables
- Sort
- Combine data set
- Labels
- Notes
- Variable utilities
- Matrices
- Other utilities

Graphs Menu

- Easy graphs
- Two way graphs
- Overlaid two way graphs
- Bar charts
- Dot charts
- Pie charts
- Histogram
- Box plot
- Scatter plot matrix
- Distributional graphs
- Smoothing & densities
- Regression diagnostic plots
- Time series graphs
- Cross-sectional time-series line plot
- Survival analysis graphs
- ROC analysis
- Quality Control
- More statistical graphs
- Table of graphs
- Manage graphs
- Change Scheme/Size

Statistics Menus

- Summaries, Tables, & tests
- Linear models & related
- Binary outcomes
- Ordinal outcomes
- Categorical outcomes
- Count outcomes
- Endogenous covariates
- Selection models
- Generalized Linear models
- Non-parametric analysis
- Time-series
- Multivariate time series
- Longitudinal/panel data
- Survival analysis
- Observational/epidemiologic analysis
- Survey data analysis
- Multivariate analysis
- Resampling
- Post estimation
- Other

Two Ways of Using Stata

- Interactive mode
  - Using menus and buttons
- Text mode
  - Write command lines in the command window
  - Write command lines in a command file (i.e., do file) and execute the do file

Two Ways of Using Stata (Continued)

- A free program editor software (PFE), which you can download at http://www.lancs.ac.uk/staff/steveb/cpaap/pfe/
- Advantages of PFE
  - Small and portable
  - Save the command files in an ASCII format
  - Allow you to look at the long line of data
  - You can keep the log file open, while ask Stata to rewrite the log file.
  - Allow you to search and replace special characters in the command and log files

How to Manage Data?

What are data?

Assuming you collect information on gender and age from three respondents, including Paul, Jim, and Sandy. The following table summarizes the information:

name	gender	age
Paul	male	25
Jim	male	35
Sandy	female	25

- The data set is a table consisting columns and rows
- Each column represent a variable (i.e., Name, Gender, or Age)
- Each row, except the first row, represents the information collected from each respondent (i.e., Paul, Jim, or Sandy)

What Does Data Management Mean?

- Read data into Stata
- Take a look at the data file
- Change the order of observations or variables
- Add labels
- Modify variables
- Create new variables
- Merge data
- Create a subset of data
- Save as a Stata data file

Use Stata to Read in Data

- if the data file is a Stata file,
  - use the file menu
  - use the open button
  - use the command line “use file name, clear”
- if the data are an excel, SAS, or SPSS file,
  - use Stat/transfer software to translate the file into a Stata data file.
- If you need to input the data yourself,
  - type data into an excel file, and use Stat/Transfer to transfer it into a Stata file
  - use data editor option and then rename the variables

Take a Look at the Data

- Find the attribute of data
  - count
  - describe
- Take a look at the values of a variables
  - list
  - tab1
- Look at the variable for some observations
  - list name in 1
  - list name in 1/3
  - list name if age ==25
  - list name if age ==25 & name ==“Paul”

Take a Look at Data (continued)

- Important operators
  - Examples
    - > Greater than
    - < Less than
    - = = Is or is equal to (equality symbol)
    - > = Greater than or equal to
    - < = Less than or equal to
    - & And
    - | Or
    - ~ Not
    - + addition
    - - subtraction
    - * multiplication
    - / division
    - ^ power
  - Symbols only, no mnemonic equivalents
  - Symbols need to be in correct order
    - >= NOT =>

Change Orders of Observations or Variables in the Data

- Change the order of observations
  - sort name
  - sort age
- Change the order of variables
  - order age gender name
  - move age name

Modify Variables

- Change the name of a variable
  - rename age age2

- Change the value of a variable

  Example 1: change the age of 25 to age of 35
  - recode age (25 =35)
  - replace age = 35 if age == 25

  Example 2: change the age of 25 to age of 35 for Paul only
  - recode age (25 =35) if name == “Paul”
  - replace age = 35 if age == 25 & name == “Paul”

  Example 3: change the age of 25 to age of 35 and age of 35 to age of 40
  - recode age (25 =30 35 = 40)
  - replace age = 30 if age == 25
    replace age = 40 if age == 35

- Change between numeric variables and string variables
  - change numeric variables to string variables: tostring age, gen(n_age)
  - change string variables to numeric variables: destring n_age, gen(n2_age)

Add Labels

- Three types of labels: data labels, variable labels, and value labels
- Add data label

  Example: add label to the data

  label data “Stata workshop 2009”
- Add variable labels

  Example: add a variable label to the age variable

  label age “the age of respondent”
- Add value labels

  Example: add values labels for the age variable

  label define agelable 25 “mid 20s” 35 “mid 30s”

  label value age agelable

Create New Variables

- Make a copy of an existing variable
  - gen age3 = age2
- Create a dummy variable
  - gen dummy
  - replace dummy =1 if gender ==“male”
  - replace dummy =0 if gender ==“female”
  - label dummy “dummy variable for gender”
- Create new variables from existing variables
  - gen age4 = age2 + age3
  - gen age5 =age 2 - age3
  - gen age6 = age2 * age3
  - gen age7 = age2 / age3
- Create new variable from the function of other variables
  - egen m_age = mean(age)
  - egen age4_2 = rowtotal (age2 age3)

Merge Data

Data_A

name	gender	age
Paul	male	25
Jim	male	35
Sandy	female	25

Data_B

name	gender	age
Joy	female	40

Data_C

name	Education
Paul	high school
Jim	college
Sandy	graduate school

Merge Data (continued)

- Add observations
  - open data_A
  - append using data_B
- Result:

name	gender	age
Paul	male	25
Jim	male	35
Sandy	female	25
Joy	female	40

Merge Data (Continued)

- Add variables
  - open data_A
  - merge name using data_C
  - Note: both data_A and data_C need to be sorted by name before merging them together
- Result:

name	gender	age	Education
Paul	male	25	high school
Jim	male	35	college
Sandy	female	25	graduate school

Create a Subset of Data

- Keep certain variables
  - keep name
- Delete certain variables
  - drop name
- Keep certain respondents
  - keep if name ==“Paul”
- Delete respondents
  - drop if name == “Paul”

Save a Stata file

- Use the file menu, select “save” or “save as”
- Use the “save” button
- Use the command line
  - save -file name-

An Example of Stata command file

```
/* Assign 30 megabyte of memory to Stata */
set mem 30m

/* set the maximal number of variable to 10,000 */
set maxvar 10000

/* Suppress the pause function in the result window */
set more 1

/* Open the log file and allow this log file to be overwritten */
log using “c:\temp\stata1.log”, replace

/* Clear the data set in the memory and then read data into Stata */
use “c:\temp\data_A.dta”, clear

/* Take a look at the data file */
describe
```

An Example of Stata Command file (Cont.)

```
/* Change the order of observations or variables */
    order age gender name
/* Add labels */
    label  age “the age of respondent”
/* Modify variables */
    recode age (25=35)
/* Create new variables */
    gen age2 = age
/* Append data */
    append using data _b
/* Merge data */
    sort name
    merge 1:1 using data _c
/* Create a subset of data */
    keep name age
/* Save a Stata file */
    save c:\temp\data_D.dta, replace
/* Close the log file */
    log close
```

Reminder of Using Stata

- limitations on the using Stata
  - 2,147,483,647 observations
  - 32,767 variables
  - 80 letters in the labels for data sets and variables
  - 32 letters for the name for a variable or value label
  - 244 letters in the value of a string variable
- Uppercase letters are treated differently from lowercase letters
- Beware of the logical flow in using Stata
  - You need to read a data before you can manage it.
  - You need to create a new variable before you can manage it
  - You need to sort the data sets first before you can merge them together

Strengths and Weaknesses of Stata

- Strengths
  - Stata is cross-platform compatible
  - Stat/Transfer software help you transfer data between Stata and other statistical software
  - You can easily learn how to use Stata even if you do not know the syntax
  - Stata is easily extensible
- Weaknesses
  - Some special statistical analyses were not available in Stata, e.g. structural equation modeling or item response analysis
  - Stata still takes a lot of time to use Stata to create graphs.
  - Stata may have problem analyzing large data sets.

Where to Find Help

- help and search
  - *help* tabulate
  - *search* tabulate
  - *search* rc 198
- *Useful website*
  - Stata website (www.stata.com)
  - UCLA (http://www.ats.ucla.edu/stat/stata/)
  - University of North Carolina (http://www.cpc.unc.edu/services/computer/presentations/statatutorial)
- User group (http://www.stata.com/statalist/)
- CFDR programming support
  - Hsueh-Sheng Wu @ 372-3119 or wuh@bgsu.edu

Conclusions

- Stata is a power software and very easy to use.
- Use the interactive mode to learn about Stata, and the use text mode for doing research
- Your ability as a researcher is the main determinant of the quality of your research.