Introduction to Stata

Alicia Doyle Lynch
Harvard-MIT Data Center (HMDC)

Documents for Today

- Find class materials at: http://libraries.mit.edu/guides/subjects/data/training/workshops.html
  - Several formats of data
  - Presentation slides
  - Handouts
  - Exercises
- Let's go over how to save these files together

Organization

- Please feel free to ask questions at any point if they are relevant to the current topic (or if you are lost!)
- There will be a Q&A after class for more specific, personalized questions
- Collaboration with your neighbors is encouraged
- If you are using a laptop, you will need to adjust paths accordingly

Organization

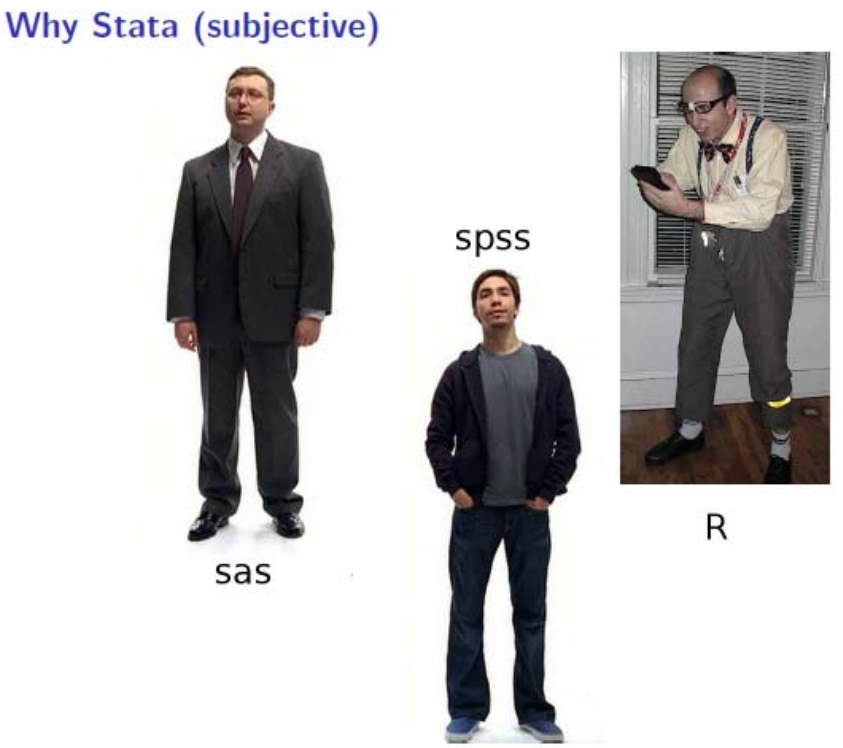
- Make comments in your Do-file rather than on hand-outs
  - Save on flash drive or email to yourself
- Stata commands will always appear in red
- "Var" simply refers to "variable" (e.g., var1, var2, var3)
- Pathnames should be replaced with the path specific to your computer and folders

Assumptions and Disclaimers

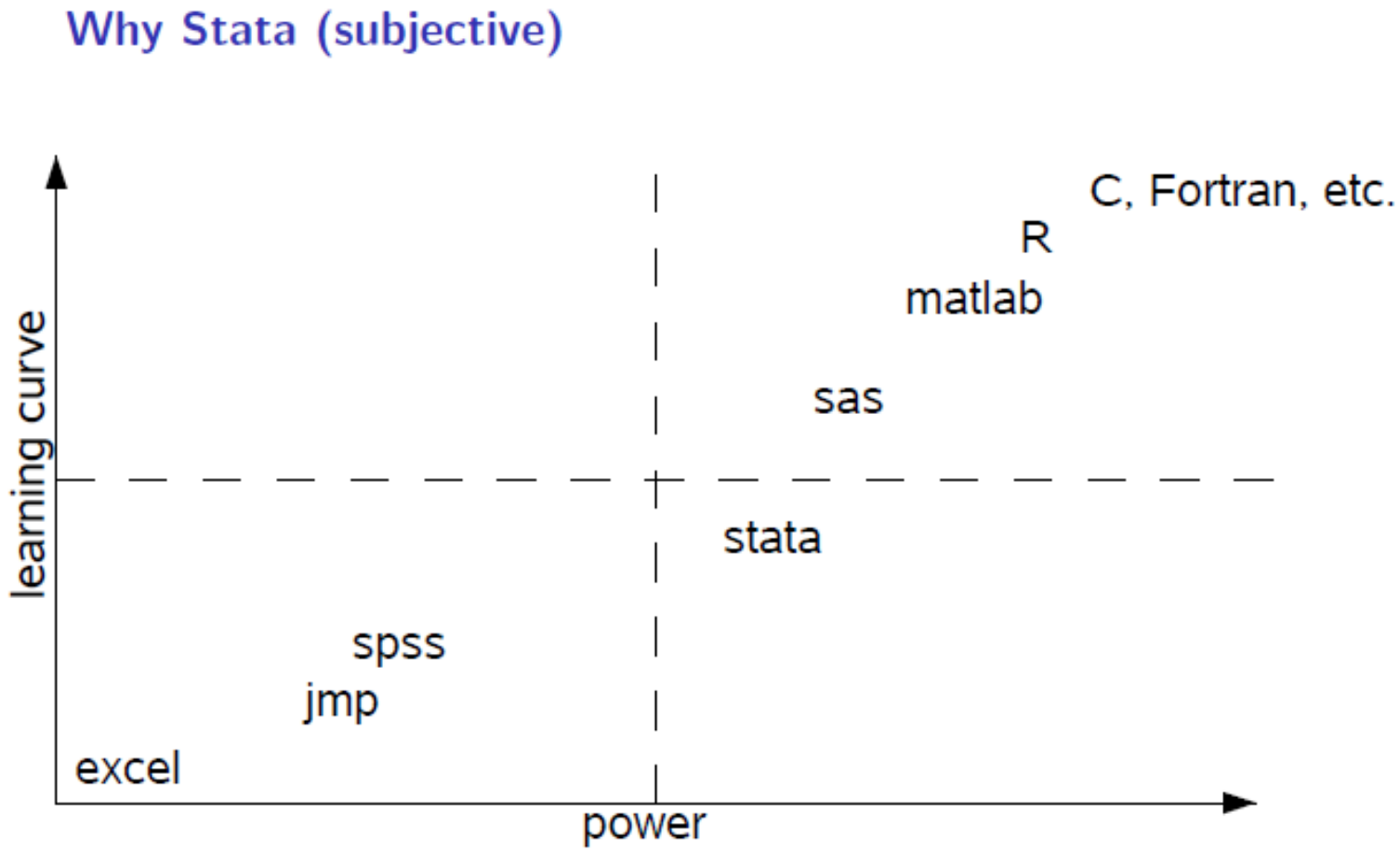
- This is an **INTRODUCTION** to Stata
- Assumes no/very little knowledge of Stata
- Not appropriate for people already well familiar with Stata
- If you are catching on before the rest of the class, experiment with command features described in help files

Why Stata?

- If you know Stata, it is likely you will not need any other software packages
- Used in a variety of disciplines
- Great guides available on web (as well as in Dewey Library)



Why Stata?


Why Stata (subjective)
learning curve
power
C, Fortran, etc.
R
matlab
sas
stata
spss
jmp
excel

Which Stata is right for you?

Package	Max. no. of variables	Max. no. of right-hand variables	Max. no. of observations	64-bit version available?	Fastest: designed for parallel processing?	Platforms
Stata/MP	32,767	10,998	unlimited*	Yes	Yes	Windows, Mac (64-bit Intel), or Unix
Stata/SE	32,767	10,998	unlimited*	Yes	No	Windows, Mac, or Unix
Stata/IC	2,047	798	unlimited*	Yes	No	Windows, Mac, or Unix
Small Stata	99	99	1,200	No	No	Windows (32-bit) or Mac (32-bit)

How do I get Stata?

- Your Department IT
- Athena terminals at MIT
- HMDC labs
- Buy it: educational or grad plan
- http://libraries.mit.edu/guides/subjects/data/software/index.html

Opening Stata

- In your Athena terminal (the large purple screen with blinking cursor) type

  add stata

  xstata

- Stata should come up on your screen
- Always open Stata FIRST and THEN open Do-Files (we'll talk about these in a minute), data files, etc.

Stata Interface

- Comprised of four windows:
  - Results
  - Command
  - Review
  - Variables
- Review and Variable windows can be closed (user preference)
- Command window can be shortened (recommended)

Do-Files

- A fifth window, called a “Do-file” is also important
- Open Do-File via icon or with dropown menu
- You can type all the same commands into the Do-File that you would type into the command window
- BUT…the Do-file allows you to SAVE your commands
- Your Do-file should contain ALL* commands you executed
  - *at least all the “correct” commands!

Command Window vs. Do-File

- I recommend never using the command window
- Saving commands in Do-File allows you to keep a written record of everything you have done to your data
  - Allows easy replication
  - Allows you to go back and re-run commands, analyses and make modifications

Let's get started

- Open up a new Do-File
- Before we do anything, we need to tell Stata how much memory to use

  set mem 500m, perm

  – "Perm" makes this permanent (everytime you open Stata, it will allow 500m of memory)

Opening Files in Stata

- Look at bottom left hand corner of Stata screen
  - This is the directory Stata is currently reading from
- We can also see this by typing pwd in our Do-File editor
- Use dir to see what is in the directory
  - If your datafile is not there, Stata will not open it!

Opening Files in Stata

- When I open Stata, it tells me it's using the directory:
  - afs/athena.mit.edu/a/d/adlynch
- But, my files are located in:
  - afs/athena.mit.edu/a/d/adlynch/IntroStata
- I'm going to tell Stata where it should look for my files:

  cd "~/IntroStata"

A Note About Path Names

- If your path has no spaces in the name (that means all directories, folders, file names, etc. can have no spaces), you can write the path as is
- If there are spaces, you need to put your pathname in quotes

Data File Commands

- Once we use the cd command to set the pathname, we no longer have to worry about the path
- Retrieving your data file:

  use datasetname.dta
- Saving your data file:

  save datasetname.dta
  - This command should be followed by ", replace" if you're writing over an existing file

Where's my data?

- Data editor (browse)
- Data editor (edit)
  - Never use!
- Always keep any changes to your data in your Do-file
- Avoid temptation of making manual changes by viewing data via the browser rather than editor

Creating a Log File

- A log file is a recording of your Stata session
- Basically, it's saving a copy of your results window
- To create a log file:

  log using logname

  – If you're writing over an existing log, this command should be followed by ", replace"

How to Start Every Do-File

- 1. Set memory
- 2. Call up dataset
- 3. Begin log file

```
/*DESCRIPTION OF FILE*/
set mem 500m
use datasetname.dta
log using logname
```

Stata Help

- Easiest way to get help in Stata – just type "help" followed by topic

  help regress

- "Search" also works – but is less specific
- Generally, if you google "Stata [topic]," you'll get some helpful hits
- UCLA website: http://www.ats.ucla.edu/stat/Stata/

General Stata Command Syntax

- Most Stata commands follow the same underlying principles
- Command variable(s), options

  sum var1 var2, detail

  – CAUTION – in some cases, if you type a command and don't specify a variable, Stata will perform the command on all variables in your dataset
- You can find command-specific syntax in the help files

General Stata Command Syntax

- Always label your Do-file immediately
- Use comments throughout
  - Stata needs to be told what is a comment and what is a command:
    - *comment
    - /*comment comment comment comment comment comment comment comment comment*/

What if my data is not a Stata file?

- Delimited, ASCII (text file)

  insheet using gss.csv, clear

  outsheet using gss_new.csv, replace comma

- Stata will open SAS transport files

  fdause gss.xpt

What if my data is from another statistical software program?

- SPSS/PASW will allow you to save your data as a Stata file
  - Go to: file > save as > Stata (use most recent version available)
  - Then you can just go into Stata and open it
- StatTransfer

What if my data is in excel?

- You can copy and paste your excel file directly into Stata's data editor
- You need to make sure that all of your columns have labels
- After you paste, you will see a prompt asking, "Is the first row data or variable names?"
  - Select "treat first row as variable names"
- Or, if you save as .xml use syntax:

```
xmluse gss.xml, doctype(excel) firstrow
```

Exercise 1: Importing Data

Descriptive Statistics

- Review your data carefully
  - describe
  - sum
  - codebook
  - list
  - tab
- Remember, if you run these commands without specifying variables, Stata will produce output for every variable

Once you have successfully imported

- View data visually with a histogram

  hist varname

- Interested in normality of your data? You can as Stata to draw the normal curve over your histogram

  hist varname, normal

Variable and Value Labels

- You never know why and when your data may be reviewed
- ALWAYS label every variable no matter how insignificant it may seem
- Stata uses two sets of label commands
  - 1. variable labels
  - 2. value labels

Variable and Value Labels

- Label variable hh_inc “household income”

  la var inc “household income”
- Want to change the name of your variable?

  rename oldvarname newvarname

Variable and Value Labels

- Value labels are labels you put on the values that variables take on (e.g., “yes,” “no,” “1,” “2,” “3”)
- Value labels are a two step process:
  - 1. “define” a value label
  - 2. Assign defined label to variable(s)

Variable and Value Labels

- Let's define a value label for yes/no responses

  la define example 1 "Yes" 0 "No"
- Stata knows what our label means, but now we need to assign it to variable(s)

  la val var1 var2 var3 example
- Label define useful when you have multiple variables with the same value structure
  - Less useful when you have only one variable with the corresponding value structure
- If you have many variables, you can search labels using: lookfor

  lookfor income

Exercise 2: Variable Labels and Value Labels

Data Manipulation Commands

- After ensuring variables were correctly imported you may wish to create new variables or modify existing variables
- NEVER delete or write over an original variable

Useful Data Manipulation Commands

- == equal to (status quo)
- = used in assigning values
- != not equal to
- > greater than
- >= greater than or equal to
- & and
- | or

Data Manipulation Commands

- Creating a new variable? Start off with "gen" command

  gen newvar = .

  – This creates a new, blank variable space

- Next, start adding your qualifications

  replace newvar=1 if var1==2

  replace newvar=2 if var1==2 & var2==2

  replace newvar=3 if var1==2 | var2==2

Data Manipulation Commands

- Recoding variables

  recode varname (1=2) (2=3)

- Deleting variables

  drop varname

- Keeping a subset of variables

  keep var1-varn

The “By” Command

- Sometimes, you’d like to generate output based on different categories of a single variable
  - For example, say you want to look at happiness based on whether an individual is male or female
- The “by” command does just this

```
bysort sex: tab happy
hist happy, by(sex)
```

Exercises 3 & 4

Other Services Available

- MIT's membership in HMDC provided by schools and departments at MIT
- Institute for Quantitative Social Science
  - www.iq.harvard.edu
- Research Computing
  - www.iq.harvard.edu/research_computing
- Computer labs
  - www.iq.harvard.edu/facilities
- Training
  - www.iq.harvard.edu/training
- Data repository
  - http://libraries.mit.edu/get/hmdc

Thank you!

All of these courses will be offered during MIT's IAP and again at Harvard during the Spring 2011 semester.

- Introduction to Stata
- Data Management in Stata
- Regression in Stata
- Graphics in Stata
- Introduction to R
- Introduction to SAS

Sign up for MIT workshops at:
http://libraries.mit.edu/guides/subjects/data/training/workshops.html

Sign up for Harvard workshops by emailing:
dataclass@help.hmdc.harvard.edu