Intro to Stata for Political Scientists

Andrew S. Rosenberg
Junior PRISM Fellow

Department of Political Science

THE OHIO STATE UNIVERSITY

Workshop Description

- This is an **Introduction** to Stata
- I will assume little/no prior knowledge
- Probably not appropriate for those with previous experience/knowledge
- *BUT*, we will be going through parts I-IV of the self-teaching package
- *AND*, I have the answers. . . so that may be worth something to you

Learning Objectives

1. Get a brief intro to statistical computing
2. Familiarize yourself with Stata and the Stata interface
3. Learn about the many resources that can help you
4. Go through all the commands that you need for the first couple assignments
5. Have fun :D

Disclaimer

- Statistical computing is frustrating at first
- You **will** make a lot of 'dumb' mistakes
- None of this is a reflection on your intelligence or self-worth
- Also. . . I don't use Stata because `R` is free and open-source...shhhhh
- BUT, it's super important to be competent in multiple software packages (in case you are ever a PRISM fellow, for instance)

Resources

- Accordingly, I relied on a lot of help to make these slides and prepare for this talk
- Shout out to Ista Zahn from the Institute of Quantitative Social Science at Harvard for making his materials available

Some great resources:

- Type 'help' followed by topic or command, e.g., help regress
- `http://www.ats.ucla.edu/stat/stata/`
- `http://www.statalist.org`
- `www.youtube.com`
- `http://stackoverflow.com/questions/tagged/stata`

But, your greatest ally is. . .

`http://bfy.tw/1fvM`

So why are we doing this?

"The advent of the computer has certainly revolutionised the practice of statistics." (Speed, 1985)

"It is clear that both the theory and practice of statistics are being revolutionised by the computer and that, as a result, radical changes are taking place in the teaching of statistics" (Lunn. 1985)

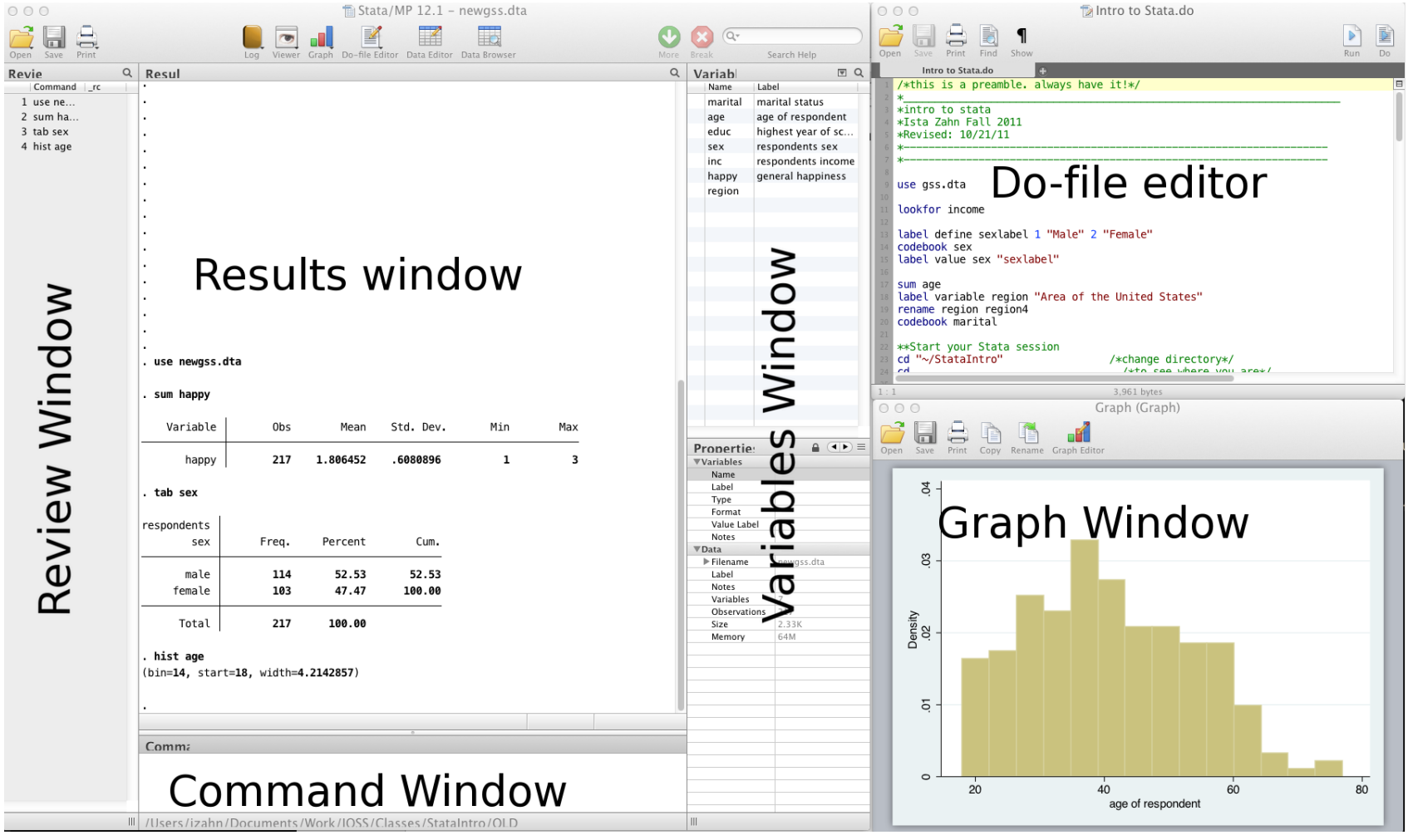
So, why are we doing this?

1. Computers can help us to do what we did before, but *way* more efficiently
2. Computers can do things nobody thought would be possible (Bayesian stats, etc.)

Why Stata?

1. Used in a variety of disciplines
2. Can accomodate user functions
3. There are a lot of resources online
4. There are some decent student discounts

The Stata Interface



Do Files: Your Future Self's Best Friend

- Think of it as your lab notebook
  1. What did I do?
  2. Why did I do it?
  3. How did I do it?
- Anything you type in the console, you can put in a do file
- **BUT**, do files allow you to *save* your commands
- It should contain EVERYTHING you run
- This is why you should never use the console or GUI to make changes

Quick Philosophy of Science Aside (Sorry...)

- We want reproducible research because it is the *only* thing a scholar can guarantee
- Unless it is purely descriptive, nothing can be proven with a **single** study...except maybe in math (with logical proofs, etc.)
- We want to be accountable, transparent and clear about how we do our research
- Those of you who have done replications understand that this is important and frustrating

General Stata Command Syntax

- Most Stata commands follow the same underlying principles
  - Command varlist, options, e.g., sum var1 var2, detail
- **BE AWARE**: 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

Commenting and Formatting Syntax

- Begin a do file by commenting (with an *) and describing what it does
- Make sure you comment throughout!!!
- Single Line and Block Comments

```
// comment
describe var
/*
comment block comment block comment block comment
block comment block comment block
*/
```

- Use /// to break varlists over multiple lines:

```
// break commands over multible lines
describe var1 var2 var2 ///
var4 var5 var6
```

Beginning a Session

1. Launch Stata
2. Open up a new do file from the menu
3. Type and Run some code!

```
// change directory
cd "~/Desktop"
// start a log file to record your stata session
log using myStataLog, replace
// Pause / resume logging with "log on" / "log off"
// Summarize some fake data
summarize mtcars
```

The Beginning of EVERY do File

1. Describe what the file does
2. Change directory to your working directory
3. Begin log file
4. Read in data
5. Save data under new name (if making changes to dataset)

Getting Data into Stata

- Open and save datasets with use and save commands
- Can read in data using insheet (data from spreadsheet) or infile (unformatted text data), etc.

```
// open the gss.dta data set
sysuse auto.dta
// saving your data file:
save newauto.dta, replace
/* the "replace" option tells stata it's OK to
write over an existing file */
```

Dealing with data

- Data editor (browse)
- Data editor (edit)
- **Don't use the data editor!** (why?)
- *Always* keep any changes to your data in your Do-file

Important Commands

1. describe: Gives labels for all variables and other stuff
2. sum: Gives summary and descriptive stats (mean, sd, min/max, etc.)
3. list: Prints observations (your Y's)
4. tab: Crosstabulates variables
5. browse: View the data like a spreadsheet (excel-style)

Commands you will need this week

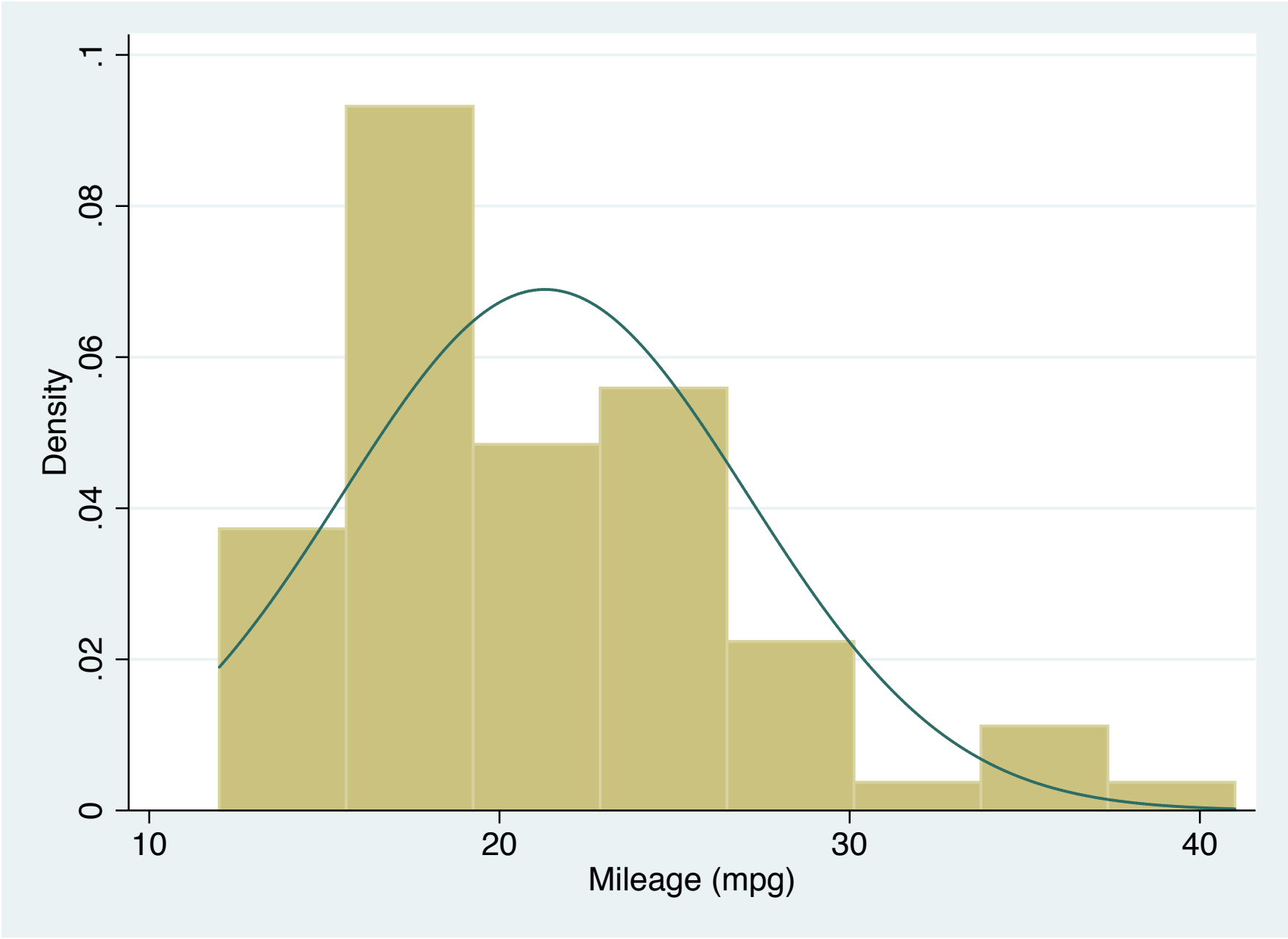
1. clear: removes specified data and variables from memory
2. using: `infile a b c using "~:My Data:myfile.raw"`
3. gen: generate a new variable
4. replace: replace contents of existing variable
5. drop: eliminate variables
6. if: used in above commands to do stuff conditionally
   - `summarize mpg if foreign==0`
7. aw: used to specify analytic weights
   - `tab mpg [aw= price]`
8. quietly: perform command but suppress output
   - `quietly regress mpg weight foreign headroom`

Basic Plotting

- Draw a histogram using hist

```
hist mpg
/* Interested in normality of your data? You can tell
Stata to draw the normal curve over your histogram*/
hist mpg, normal
```

Histogram Example

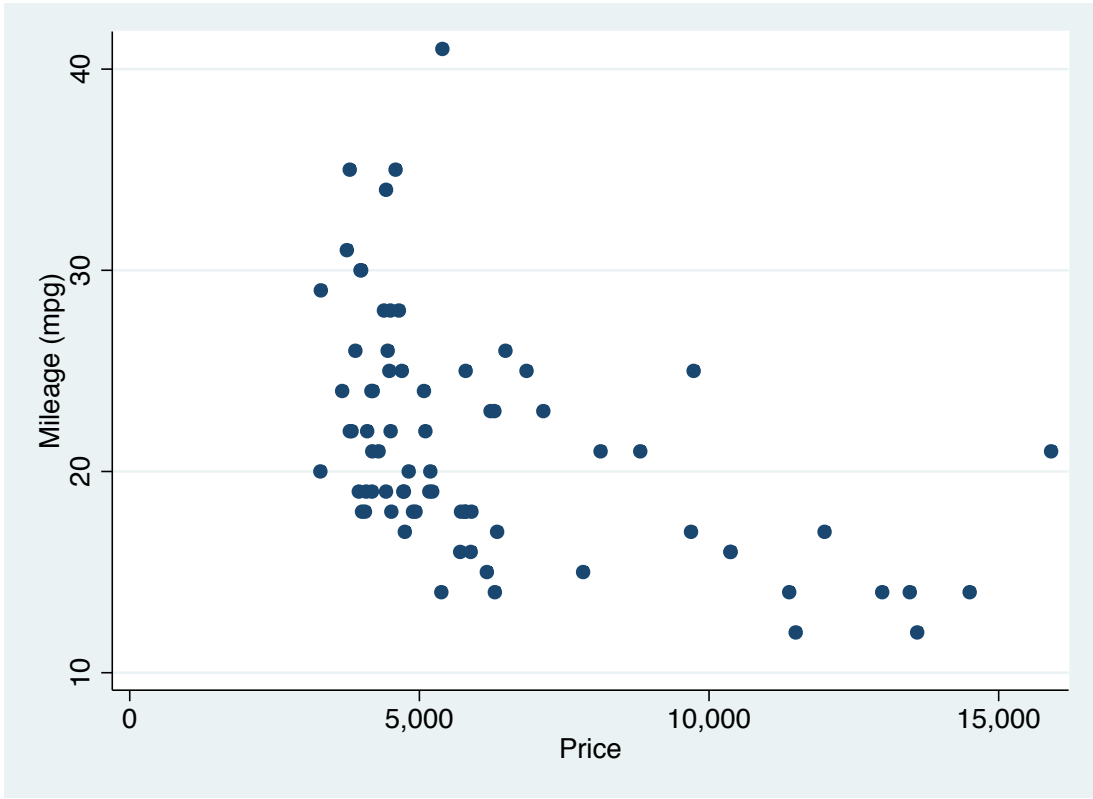


Basic Plotting

- Make a scatterplot

```
/* scatterplots */
twoway (scatter mpg price)
```

Scatterplot Example

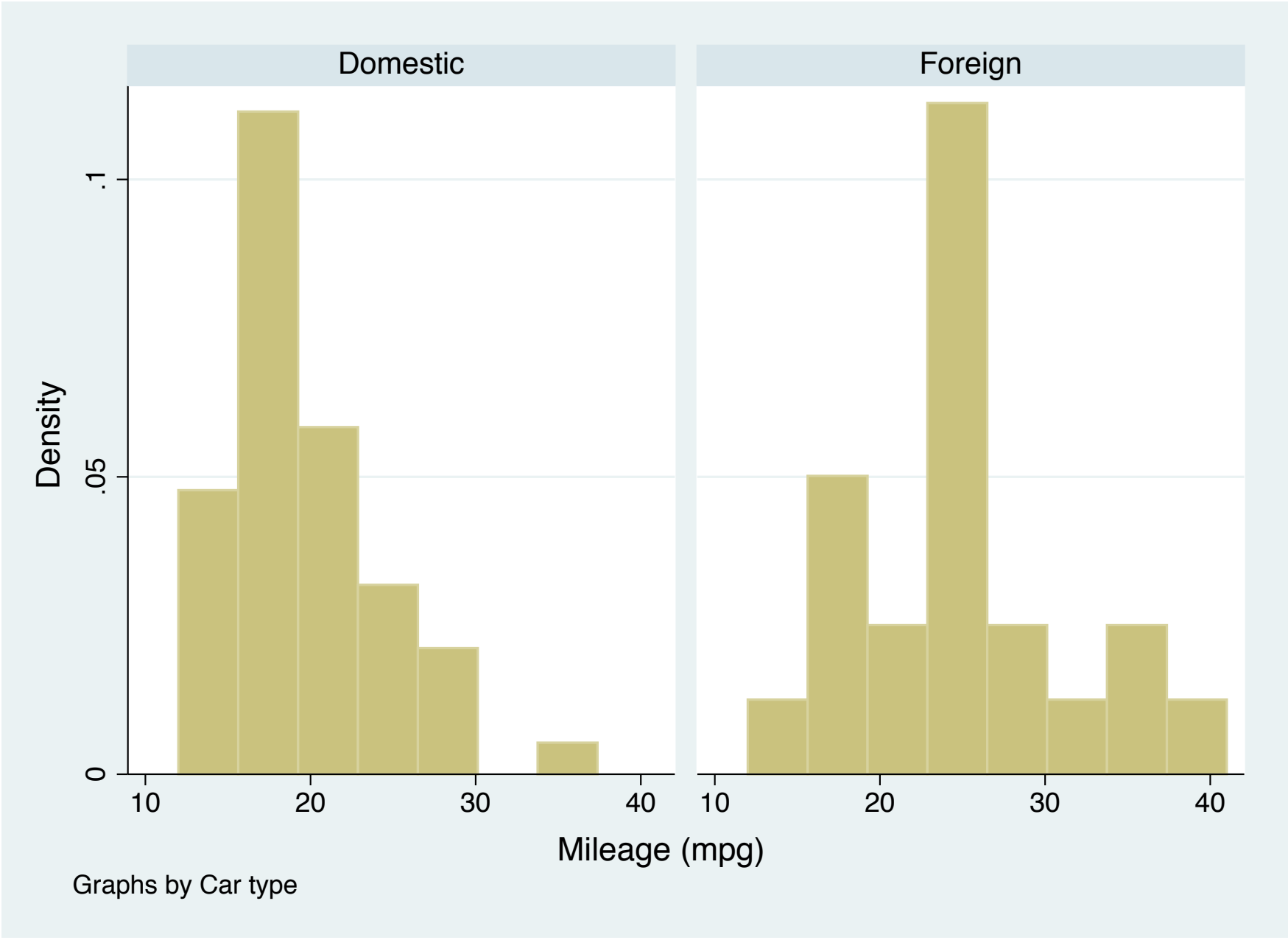


The All-Imporant “by" Command

- Often, one wants to do an operation based on the value of another variable
- The ‘by’ command does this

```
/* tabulate mpg separately for foreign and domestic cars */
bysort foreign: tab mpg
/* not all commands can be used with the by prefix.
some, (like hist) have a "by" option instead */
hist mpg, by(foreign)
```

See how cool 'by' is???



Working with Subsets of Data

- It's important to be able to look at specific rows of your data
- For example, "what is the average income for women (Sex = 1)?"
- There are a bunch of operators that help with this
  1. == (Equal to)
  2. != (not equal to)
  3. <, etc. (less than, etc.)
  4. <= (Less than or equal to)
  5. & (AND)
  6. | (OR)
- Example: `mean mpg if price < 8000`

Generating and Replacing Variables

- To create a new variable, use 'gen'

```
/* create a new variable named mean_price"
equal to price minus the mean of price */
mean price
gen mean_price = price - 6165.87
```

- Often, it's easier to generate a blank variable and fill it in based on values of existing variables

```
/* generate a column of missings */
gen dummy_price = .
/* Next, start adding your qualifications */
replace dummy_price=0 if price<6165 & foreign == 1
replace dummy_price=1 if price>6165 & foreign == 1
```

Renaming Variables

- Renaming variables is easy
- rename *old name* **new name**

```
/* rename a confusing name */
rename inch90 income_house_90
```

Recoding or Dropping Variables

- Recode 'foreign'

```
/* recode foreign into domestic */
recode foreign (1=0) (0=1), gen(domestic)
```

- Drop a variable

```
drop make // delete make of car variable
keep price-foreign // keep the rest
```

OK!

- Now you should be able to get started on your homework!
- I have described all of the commands needed to do the exercises
- However, I skipped some of the specific stuff that is discussed in the packet
- For example, I didn't go over adding weights explicitly
- But, I think all of the assumed knowledge from the packet is covered
- I'll be here to answer your questions

If you need further help

- PRISM office hours (Tuesdays and Thursdays and by appointment)
- My email: rosenberg.1108@osu.edu

Thanks for your attention!

Table of Contents

- Resources
  - Your Greatest Ally
- Statistical Computing
  - Inspirational Quotes
  - The Why
- Intro to Stata
  - Why Stata?
  - What does it look like?
  - Do Files
  - Getting Started
  - Reading in Data
- Important Details
  - Important Commands
  - Commands for HW
  - Plotting
  - The “by" Command
  - Subsets
- Playing with Variables
  - Gen or Replace
  - Renaming
  - Recode or Drop