Manipulating and visualizing data in R

Frank Edwards

1/25/2019

School of Criminal Justice, Rutgers - Newark

Review HW1

- Challenges?

Understanding the R environment

Let’s check installations

- R
- RStudio
- Git

Let’s clone the course repository

In RStudio Terminal, navigate to desired folder using command cd

```
https://github.com/f-edwards/intermediate_stats.git
```

The R and RStudio environment

http://stat545.com/block002_hello-r-workspace-wd-project.html

Using RMarkdown

http://stat545.com/block007_first-use-rmarkdown.html

Data input/output

http://stat545.com/block026_file-out-in.html

Basics of data frames in R

http://stat545.com/block006_care-feeding-data.html

Introducing dplyr

http://stat545.com/block009_dplyr-intro.html

Model objects in R

- Work through basics of lm()
- Interpret output
- Extract objects from model object

Working with (generalized) linear models

Basic syntax of lm

To produce:

$$y = \beta_0 + \beta_1 X + \varepsilon$$

We use:

lm(y ~ x)

Start a new R Project

1. Save it as lecture2.RProj in your intermediate_stats folder
2. move the data from ./data to ./lecture2/data (drag and drop or cp from terminal)

If using terminal (try it!) input

- mkdir data
- cd ..
- cp ./data/lecture_2_demo.csv ./lecture2/data/lecture_2_demo.csv

Read in data

```
cj_budgets<-read_csv("./data/lecture_2_demo.csv")
```

Note that we use a relative file path here (what's that!?)

Explore the data

- Evaluate integrity with head()
- Evaluate column classes with str()
- Evaluate summary statistics with summary()

Develop a theory for relationships between variables

Let's look the outcome \texttt{exp_police_pc}, which is the total expenditures on police per capita (divided by the total population) in a county across municipal and county governments (local PDs, sherrif's depts).

What could predict police budgets?

```
head(cj_budgets)
```

```
## # A tibble: 6 x 7
##   fips  exp_police_pc officers_pc rev_prop_tax_pc violent.cri
##   <chr>         <dbl>       <dbl>           <dbl> <lgl>
## 1 01001          169.        83.4            158. TRUE
## 2 01021          170.        32.2            242. TRUE
## 3 01033          193.        31.9            256. TRUE
## 4 01061          164.        29.9            260. TRUE
## 5 01081          187.        51.6            331. TRUE
## 6 01083          164.        34.4            170. FALSE
## # ... with 2 more variables: segregation.bw.high <lgl>,
```

Visualize the data, univariate

- hist()

Visualize the data, bivariate

- Look at crosstabs with table()
- Using base, plot()

Fit a single variable model

Interpret the model

- What does each coefficient mean

Visualize the fit

We can use coef() to extract coefficient estimates.

How could we use coef() to plot a fitted line?

Fit a regression with multiple variables

- Interpret the coefficients
- Visualize the fit using coef() and plot()

Fit a model with an interaction

- Interpret the coefficients
- Visualize the fit using coef() and plot()

Introducing prediction and counterfactuals

- Create fake data
- Use coef() to generate expected values
- Use predict() to generate expected values

Discuss HW2

- HW2 asks you to provide a brief analysis and writeup on the cj_budgets data
- Use RMarkdown to explore the data and write up your findings
- Construct a multivariate regression model and interpret your results
- Predict values for a theoretically interesting counterfactual
- Provide both criminological and statistical interpretations of your findings