Week 10 – Modeling Data

Joshua Rosenberg and Alex Lishinski

March 15, 2021

Welcome!

Welcome to *week 10*!

Record the meeting

Breakout rooms!

Starting with whoever has the most tidy current work space (home, office) . . .

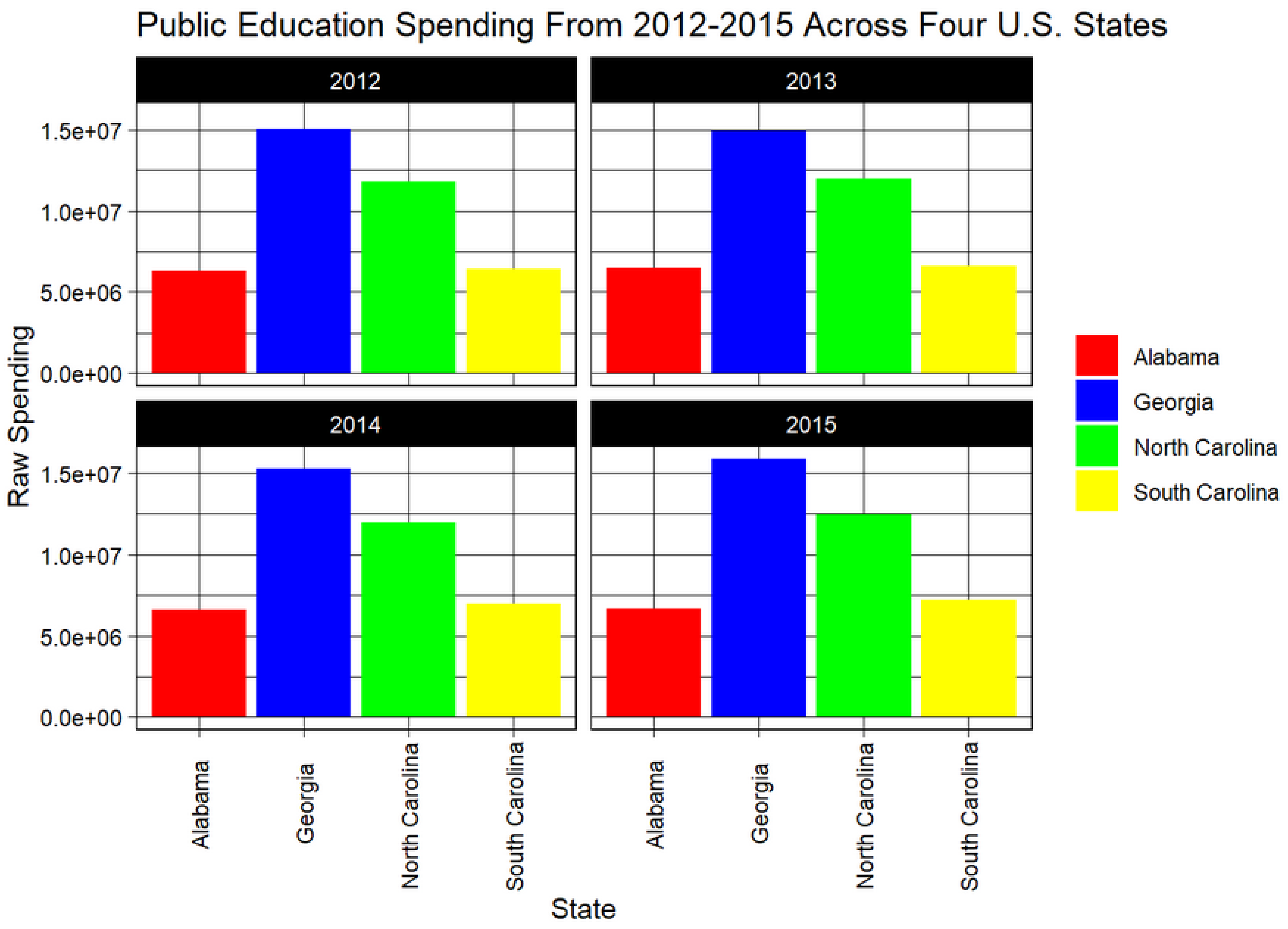
- What was most challenging about the last homework (in which you carried out an initial version of a complete analysis)?
- What was the most rewarding about the last homework?

Record the meeting

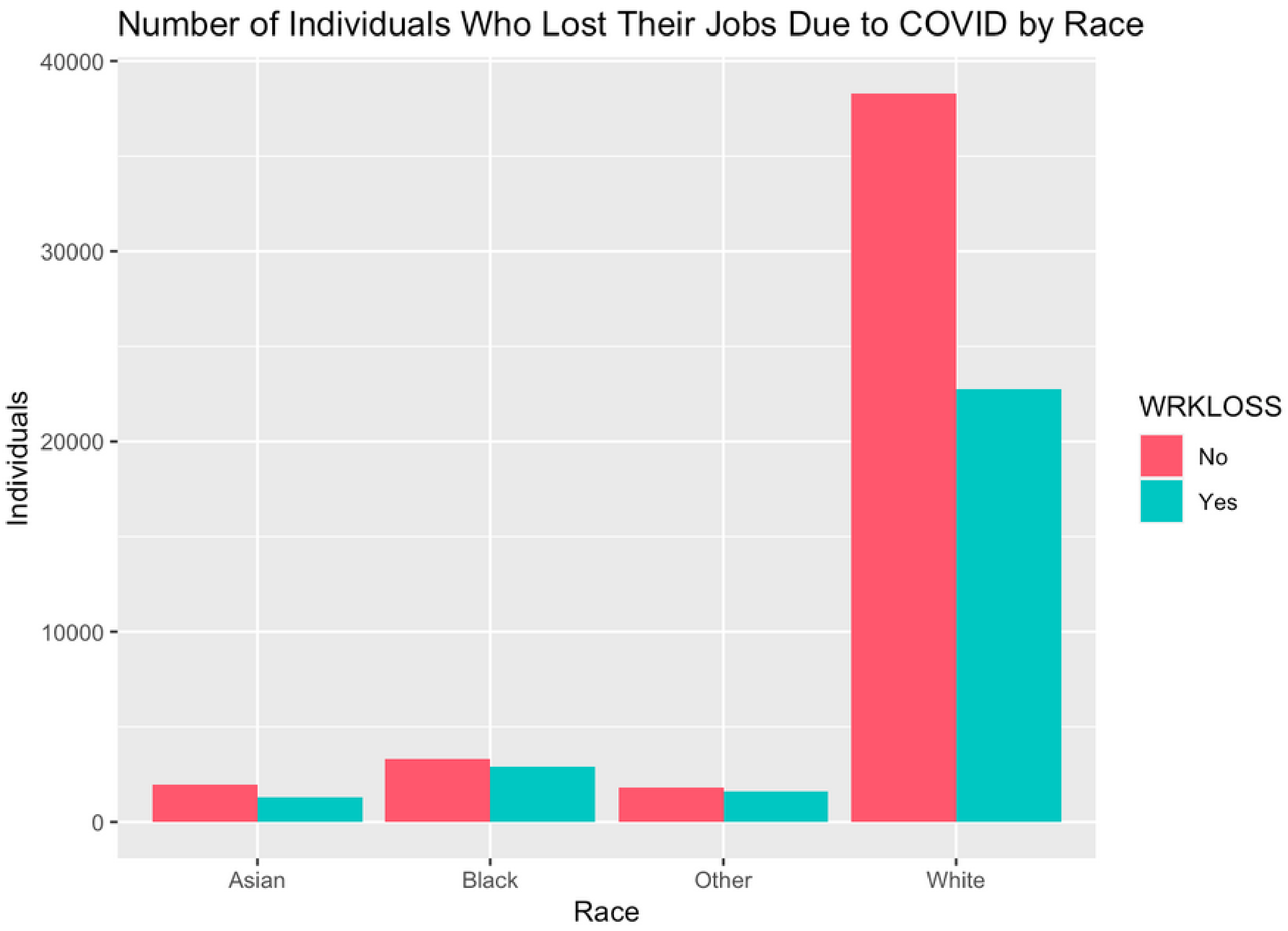
Review of last week's class

- From Soup to Nuts: Carrying out a complete analysis
- Exam 2

Homework highlights



Homework highlights



Exam recap

- Overall everyone did quite well!
- Many questions 100% correct
- A couple of items to highlight

Exam recap

A couple of items to highlight

Which of the following is one use case for the kable package?

- To format tables included in RMarkdown documents and reports, (Correct answer) 56 %
- To easily create plots of your data, (Incorrect answer) 6 %
- To create a correlation matrix, (Incorrect answer) 6 %
- To automatically calculate summary statistics for a data frame, (Incorrect answer) 31 %

Exam recap

A couple of items to highlight

For this question, we use "left data frame" to refer to the first data frame passed to the join function, and "right data frame" to refer to the second data frame passed to the join function.

For example, below, df1 is the left data frame, and df2 is the right data frame.

left_join(df1, df2)

See here (and the slides) for a description of different joins: https://datascienceineducation.com/c07.html#joining-the-data

Which function would you use if you wanted to join the left and right data frame based on a key (a variable present in both data frames), joining those rows in the right data frame to those that match a key in the left data frame, and keeping all of the rows in the left data frame?

- anti_join() 0%
- left_join() (Correct answer) 63%
- right_join() 13%
- semi_join() 25%

Exam recap

A couple of items to highlight

Which of the following have we not considered to be a data visualization layer?

- mapping between data and geometric object 0%
- theme 0%
- data 38%
- geometric object 13%
- title (correct answer) 50%

Mid–semester feedback

Thanks everyone who took the time to give feedback!

What should we keep doing?

- Homeworks
- In class demos

Mid–semester feedback

Thanks everyone who took the time to give feedback!

What should we do more of?

- Opportunities to try things in class
- Better explanation in HW

Mid–semester feedback

Thanks everyone who took the time to give feedback!

What should we do less?

- Base groups (not that much time)
- Showing too much stuff in demos (in too short time)
- Live coding Base R vs homework (Rstudio)

Mid–semester feedback

Thanks everyone who took the time to give feedback!

Other comments?

- More frequent quizzes/checks
- Ability to use different data sets

This week's topics

Overview

A. Final project presentations

B. A buffet of models

A. Final project presentations

- Thank you for adding your ideas to the final project brainstorm!
  - https://docs.google.com/presentation/u/3/d/1KWU5bhxZmV63vkNtQNxHo-2Yl_iYJmSW2oOZ_jtFeP0/edit#slide=id.p
- Today, each of you will briefly (1 min. or less) present on your final project idea
- We will each provide feedback via Jamboard
  - https://jamboard.google.com/d/1S6K3ED_jvCS5b-GP3wfk5S3o6g4Buj0wgq6fTlmUY2w/edit?usp=sharing

B. A buffet of models

There are a number of ways to understand variables about which you have data and the relationships between them.

One way is to create a **model**, a simplified *representation* of your data that can be informative to you (and others) about your data – and, maybe, what your data represents.

From this broad definition, models can take many different forms:

- A sample statistic (e.g., a *mean* of a variable)
- A relationship describing how two variables co–vary (e.g., a bivariate *correlation*)
- A linear regression model
- . . . (what models are common in your field?)

B. A buffet of models

One of the benefits of modeling your data within R is that many R packages share a common modeling syntax, or interface: the formula syntax.

This code represents the regression of `hp` upon `mpg`:

```
mpg ~ hp
```

This code often corresponds to the underlying mathematical/statistical equation:

$$\mathbf{mpg} = \alpha + \beta_1(\mathbf{hp}) + \epsilon$$

B. A buffet of models

Today, we'll focus on the linear regression model, but will also touch on the following:

- t–test
- ANOVA
- generalized linear model (i.e., Poisson or Logistic Regression)
- multi–level (or hierarchical linear) model

B. A buffet of models

There is a *lot* we can do with a linear regression model!

```
d <- read_csv("https://raw.githubusercontent.com/data-edu/dataedu/master/data-raw/wt01_online-scie

d
```

```
## # A tibble: 603 x 30
##    student_id course_id   total_points_possi… total_points_ear… percentage_earn…
##         <dbl> <chr>                     <dbl>             <dbl>            <dbl>
##  1      43146 FrScA-S216…                3280              2220            0.677
##  2      44638 OcnA-S116-…                3531              2672            0.757
##  3      47448 FrScA-S216…                2870              1897            0.661
##  4      47979 OcnA-S216-…                4562              3090            0.677
##  5      48797 PhysA-S116…                2207              1910            0.865
##  6      51943 FrScA-S216…                4208              3596            0.855
##  7      52326 AnPhA-S216…                4325              2255            0.521
##  8      52446 PhysA-S116…                2086              1719            0.824
##  9      53447 FrScA-S116…                4655              3149            0.676
## 10      53475 FrScA-S116…                1710              1402            0.820
## # … with 593 more rows, and 25 more variables: subject <chr>, semester <chr>,
## #   section <chr>, Gradebook_Item <chr>, Grade_Category <lgl>,
## #   FinalGradeCEMS <dbl>, Points_Possible <dbl>, Points_Earned <dbl>,
## #   Gender <chr>, q1 <dbl>, q2 <dbl>, q3 <dbl>, q4 <dbl>, q5 <dbl>, q6 <dbl>,
## #   q7 <dbl>, q8 <dbl>, q9 <dbl>, q10 <dbl>, TimeSpent <dbl>,
## #   TimeSpent_hours <dbl>, TimeSpent_std <dbl>, int <dbl>, pc <dbl>, uv <dbl>
```

B. A buffet of models

Estimating a model; seeing the result:

```
lm(FinalGradeCEMS ~ TimeSpent_hours, data = d)
```

```
##
## Call:
## lm(formula = FinalGradeCEMS ~ TimeSpent_hours, data = d)
##
## Coefficients:
##     (Intercept)  TimeSpent_hours
##         65.8085           0.3648
```

B. A buffet of models

Saving the output to an *object* and printing a summary of the results

```
m1 <- lm(FinalGradeCEMS ~ TimeSpent_hours, data = d)
summary(m1)
```

```
##
## Call:
## lm(formula = FinalGradeCEMS ~ TimeSpent_hours, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.136  -7.805   4.723  14.471  30.317
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     65.80851    1.49120   44.13   <2e-16 ***
## TimeSpent_hours  0.36484    0.03889    9.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.71 on 571 degrees of freedom
##   (30 observations deleted due to missingness)
## Multiple R-squared:  0.1335,    Adjusted R-squared:  0.132
## F-statistic: 87.99 on 1 and 571 DF,  p-value: < 2.2e-16
```

B. A buffet of models

Making the model more complex – a multiple regression

```
m2 <- lm(FinalGradeCEMS ~ TimeSpent_hours + int + Gender, data = d)
summary(m2)
```

```
## 
## Call:
## lm(formula = FinalGradeCEMS ~ TimeSpent_hours + int + Gender, 
##     data = d)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -66.593  -7.382   4.761  14.534  30.618 
## 
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     69.61325    7.06075   9.859   <2e-16 ***
## TimeSpent_hours  0.36962    0.04198   8.804   <2e-16 ***
## int             -0.99359    1.58756  -0.626    0.532    
## GenderM         -0.54962    2.06489  -0.266    0.790    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 21.03 on 499 degrees of freedom
##   (100 observations deleted due to missingness)
## Multiple R-squared:  0.1375,	Adjusted R-squared:  0.1323 
## F-statistic: 26.51 on 3 and 499 DF,  p-value: 6.362e-16
```

B. A buffet of models

Adding an interaction

```
m3 <- lm(FinalGradeCEMS ~ TimeSpent_hours + int*Gender, data = d)
summary(m3)
```

```
##
## Call:
## lm(formula = FinalGradeCEMS ~ TimeSpent_hours + int * Gender,
##     data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.812  -7.636   4.664  14.415  33.093
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      80.93390    8.70113   9.302   <2e-16 ***
## TimeSpent_hours   0.36890    0.04182   8.820   <2e-16 ***
## int              -3.65595    1.98802  -1.839   0.0665 .
## GenderM         -30.73798   13.81410  -2.225   0.0265 *
## int:GenderM       7.21687    3.26560   2.210   0.0276 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.95 on 498 degrees of freedom
##   (100 observations deleted due to missingness)
## Multiple R-squared:  0.1458,    Adjusted R-squared:  0.139
## F-statistic: 21.26 on 4 and 498 DF,  p-value: 3.358e-16
```

B. A buffet of models

t–test

```
m_t_test <- t.test(FinalGradeCEMS ~ Gender, data = d)
m_t_test
```

```
## 
##     Welch Two Sample t-test
## 
## data:  FinalGradeCEMS by Gender
## t = -0.30379, df = 327.71, p-value = 0.7615
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.579370  3.354211
## sample estimates:
## mean in group F mean in group M
##        77.01877        77.63135
```

B. A buffet of models

ANOVA

```
m_anova <- aov(FinalGradeCEMS ~ subject, data = d)
m_anova
```

```
## Call:
##    aov(formula = FinalGradeCEMS ~ subject, data = d)
##
## Terms:
##                   subject Residuals
## Sum of Squares   13484.46 269057.23
## Deg. of Freedom         4       568
##
## Residual standard error: 21.76447
## Estimated effects may be unbalanced
## 30 observations deleted due to missingness
```

B. A buffet of models

Multi–level model

```
library(lme4)
m5 <- lmer(FinalGradeCEMS ~ TimeSpent_hours + int*Gender + (1|course_id), data = d)
summary(m5)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: FinalGradeCEMS ~ TimeSpent_hours + int * Gender + (1 | course_id)
##    Data: d
## 
## REML criterion at convergence: 4433.8
## 
## Scaled residuals: 
##     Min      1Q  Median      3Q     Max 
## -3.4970 -0.4169  0.2413  0.6507  2.3171 
## 
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  course_id (Intercept)  46.47    6.817  
##  Residual              384.21   19.601  
## Number of obs: 503, groups:  course_id, 26
## 
## Fixed effects:
##                  Estimate Std. Error t value
## (Intercept)      74.22969    8.45385   8.781
## TimeSpent_hours   0.43078    0.04128  10.435
## int              -2.84129    1.89455  -1.500
## GenderM         -26.55507   13.10001  -2.027
## int:GenderM       6.39449    3.09236   2.068
## 
## Correlation of Fixed Effects:
##             (Intr) TmSpn_ int    GendrM
## TimSpnt_hrs -0.239                     
## int         -0.963  0.091              
## GenderM     -0.595  0.021  0.611       
## int:GenderM  0.583 -0.027 -0.609 -0.989
```

Live coding

Let's head over to the following file for a demonstration: `week-10-demo.R`

Logistics

This week

- Homework 10: Available tomorrow by noon tomorrow; **Due by Thursday, 4/1**
- Reading:
    - Walkthrough 3: Using School-Level Aggregate Data to Illuminate Educational Inequities: https://datascienceineducation.com/c09.html
    - https://r4ds.had.co.nz/model-intro.html
    - https://r4ds.had.co.nz/model-basics.html

Final Project

- Final project
  - Flesh out final project idea based upon feedback (this forthcoming week)
  - Then receive feedback from us (the following week)

Random

- Do you have an interest in a class on social network analysis (more of a general theoretical and methodological approach that can be brought to bear on the analysis of face-to-face and digital networks) and the analysis of social media data? If so, please let us know.

- Are you interested in a graduate-level certificate in educational data science?
  - https://docs.google.com/document/d/e/2PACX-1vRhJTuCQfpEx9uZI57pucjyr_quIR9Vv5ZZdxvu4GSrdD5IkIQyUTsWX5NyuHiPiwOMPtkL

Wrapping up

In your base group's Slack channel:

- What is one thing you learned today?
- What is something you want to learn more about?
- Share your feelings in GIF form!