Third Lab

Tom Hanna

University of Houston

tlhanna@uh.edu

February 8, 2021

Overview

Format

- Answer specific stats questions from lab
- Brief stats *overview* for this week's lab
- Additional questions
- Go over the "lab survey" from last lab
- Technical *overview* of this week
- Technical questions
- I will email you a copy of the survey questions in advance to reference as you run the lab

Root Mean Squared Error

Why do we use the Root Mean Squared Error instead of just a general error term?

- Root MSE is to MSE as Standard Deviation is to Variance - more easily interpreted
- We use Mean Squared Errors for the same reason we use squares in computing variance
- OLS Assumption: The error term has a population mean of zero
- So, just summing the positive and negative errors gives us a zero sum

Principles this week

- Model fit and model specification
- Error variance
- Standard errors
- Underspecification: Omitted variable bias
- Overspecification: Multicollinearity and irrelevant variables

R^2 or the Coefficient of Deterination

How much does the model explain?

$$R^2 = 1 - \frac{SSR}{SST}$$

Important things

- Variance-Covariance Matrix
- Variance Inflation Factor

- **Question:** "In the line ""CAdta¡-read.dta(here(""data"",""caschool.dta""))"", what does the ¡- do? Two sentences should suffice. Hint: It does something to the left side and uses the right side to do it. "
- **Answer: ASSIGNS** It creates a new **dataframe** object called CAdata and **assigns** the value from the right side to the left side. The code to the right opens a file and the data from that file is assigned to the dataframe on the left.

Question 3

- **Question:** In the line "CAdta¡-read.dta(here("data","caschool.dta"))", what kind of object is being created? That is, what will "CAdta" be once it is created? This is a one (maybe two) word answer.
- **Answer: dataframe** - *This is important R terminology*

Question 4

- **Question:** In the code, VCM ¡- cov(newdata[,c(2,4:6)]), what does **"newdata[,c(2,4:6)]"** indicate or do? Hint: what part of the dataframe "newdata" is being used here?
- **Answer:**It indicates that only the values from column 2, and columns 4 through 6, will be used in the dataframe "newdata" will be used.

Question 5

- **Question:** In the code, newdata ¡- CAdta[myvars], what is being assigned to the new dataframe "newdata"? Put differently, newdata is a _____ of CAdata. Hint: If you don't know the R terminology, think about sets.
- **Answer:subset** and **subsetting** - This is R terminology that means almost exactly what you'd think it would mean if you just think about set terminology.

Question 6

"Conventional wisdom holds that how much school districts spend is an important determinant of students' test scores." An alternate theory might be that how money is spent is more important than the amount. There are at least two variables in this data set that involve spending categories at least indirectly. Both include a total number and a per student or student ratio number. If we wanted to compare the per student commitment to these two important alternative spending categories, what variable names would we use? Any reasonable answer is acceptable. I was thinking of the

two spending categories of technology, represented by COMP_STU and teachers represented by STR.

Question 7

- **Question:** More specifically than just "California schools," when and where is the education data from? (You may cut and paste the answer.)
- **Answer:**"The data used here are from all 420 K-6 and K-8 districts in California with data available for 1998 and 1999."

Question 8

- **Question:**In finding the simple regression slope, the variance of the explanatory variable is in the _____ and the covariance of the explanatory and dependent variables is in the _____?
- **Answer:**denominator, numerator

Question 9

- **Question:**In finding the simple regression line, the intercept is the mean of the _____ variable minus the beta-1 coefficient times the mean of the _____ variable?
- **Answer:** outcome/dependent/Y - first blank; explanatory/independent/X - second blank

EC1

- **Question:**In the table following, "Using the means, variances, and covariances, you can calculate the simple regression line for the student-to-teacher ratio's effect on test scores. This entails filling in this table:" the explanatory variable in the table doesn't match the sentence. What is the variable actually used in the table and the lab script? (Hint: this refers to the version with Spring 2020 and Burak's name at the top.)
- **Answer:** expn_stu Expenditures per student
- The point of this was to give you a point for some quick proofreading because it's this sort of swapping variable names that can lead to results that make no sense.

EC2

Question In the following code, if we were interested in the variables "teachers" and "comp_stu", but we were not interested in "el_pct" how would you change the code?

```
# Dropping Variablesmyvars <- names(CAdta) %in% c("district",
 "str", "testscr", "avginc", "el_pct", "expn_stu")

 newdata <- CAdta[myvars]
```

Answer:

```
# Dropping Variablesmyvars <- names(CAdta) %in% c("district",
 "str", "testscr", "avginc", "teachers", "comp_stu", "expn_stu")

 newdata <- CAdta[myvars]
```

EC3

- **Question:**This answer is in the updated worksheet or the lab video. When you start working with a new dataset compiled by a person or organization, especially a large one, what is the first step you should take to familiarize yourself with the information in the set? Look at the _____.
- **Answer:** The Codebook!

Technical Errors

- Please don't copy from directly from Word or a webpage into the surveys.
- It looks like this:

```
"I suppose we could type <span style=""color:
 rgb(0,0,0);font-family: &quot;Open Sans&quot;
 , sans-serif;font-size: 13.0px;font-style:
 normal;font-weight: normal;letter-spacing:
 normal;text-align: left;text-indent: 0.0px;
```

- Copy to a plain text file first.
- Notepad in Windows
- TextWrangler in Mac
- https://asu.secure.force.com/kb/articles/FAQ/How-to-Paste-Text-into-Blackboard

Technical Errors

- Working directory - you don't need this anymore
- Caveat: You have to run *install.packages("here")* just one time
- Caveat 2: You have to use the updated lab scripts that include *library("here")* near the top
  - If you don't, you will get "function not found" error when you use the "here" function
- Caveat 3: You have to use R Studio Projects
  - If you don't, you will get a file not found or directory not found error

Nothing too challenging

I will run through

- Installing *here*
- Opening a Project
- Performing a Git pull

Stargazer

- Copy and paste formatted results to LaTeX
- Lets you produce tables in AJPS, APSR, IO, AER or other formats format
- By default: publish regression coefficients with standard errors, and the most commonly reported statistics
- You can edit the tables in LaTeX as needed
- You will produce a table, put it in LaTeX and email it to me
- The example took me about 60 seconds including two edits in TexStudio

Example from Lab line 14

Table: Correlation Matrix

	avginc	calw_pct	meal_pct	el_pct	comp_stu	expn_stu	str
avginc	1	-0.513	-0.684	-0.307	0.195	0.314	-0.232
calw_pct	-0.513	1	0.739	0.320	-0.152	0.068	0.018
meal_pct	-0.684	0.739	1	0.653	-0.204	-0.061	0.135
el_pct	-0.307	0.320	0.653	1	-0.251	-0.071	0.188
comp_stu	0.195	-0.152	-0.204	-0.251	1	0.287	-0.307
expn_stu	0.314	0.068	-0.061	-0.071	0.287	1	-0.620
str	-0.232	0.018	0.135	0.188	-0.307	-0.620	1

The End