Homework 3: Simple Linear Regression

FIRSTNAME LASTNAME

September 05, 2023

Instructions:

The purpose of this homework assignment is to practice basic simple linear regression. Make sure you read each question carefully. In each question, I will give you a task to do, and I will tell you want I want you to output. You can write as much code as you want in each code chunk, but make sure you complete the task and only print the output I asked you to print. Don't sort the data unless you are told to sort the data. You should remove the "#" sign in each code chunk before writing your code. **You should knit your RMD file to a PDF after you answer every question.**

The **SocGrades** dataset from the **heplots** package in R contains data collected on 40 students in an introductory sociology course. Our focus is primarily going to be on the variables **pretest**, **midterm1**, **midterm2**, and **final**. For the 40 students, we know how many points students got on each of these assessments. You will need to install the **heplots** package to access the dataset and to get the **library(heplots)** function to work. Once you install and load the library, you can use `?SocGrades` to access the documentation about this dataset or you can just search for the documentation online. Do not leave "?SocGrades" in your RMD file if you do this. This is for temporary reading of documentation in RStudio.

After you are done, knit the RMarkdown file to PDF and submit the PDF to Gradescope under HW3.

Questions

Q1 (2 Points)

Load the dataset **SocGrades** from the **heplots** package into your global environment using the `data()` function. Then, use the `str()` function to show a preview of the **SocGrades** dataset.

```
#
```

Q2 (3 Points)

Create a data frame named **SocGrades2** that only contains the variables named **pretest**, **midterm1**, **midterm2**, and **final**. Then, use the `str()` function to show a preview of the **SocGrades2** dataset. For all future questions, you will use the **SocGrades2** dataset.

```
#
```

Q3 (4 Points)

Instructors sometimes give pretests to make sure students are prepared to take the course. Sometimes pretests can be used to inform students about their chance of succeeding in the class. I want you to create a scatterplot of the final exam score versus the pretest score. From this picture, do you believe the pretest is useful for predicting the student's success on the final exam. Put your response in complete sentences in the designated space below and explain why or why not.

```
#
```

Response in Complete Sentences: REPLACE EVERYTHING IN ALL CAPS HERE WITH YOUR ANSWER IN COMPLETE SENTENCES

Q4 (3 Points)

The final exam is just one of three exams. In this sociology course, the course grade for students is a weighted average of their **midterm1**, **midterm2**, and **final** scores. Because the final exam is cumulative, the instructor puts 40 percent of the student's grade on their **final** score. The remaining 60 percent is evenly split between **midterm1** and **midterm2** (30 percent weight on each). I want you to create a new variable in **SocGrades2** called **course** which represents the student's course grade (in decimal form) based off the weighting scheme mentioned above. In order to calculate this, you need to know that the first midterm had a maximum of 80 points, the second midterm had a maximum of 100 points, and the final exam had a maximum of 170 points.

The final course grade should be expressed in decimal form so 0.93 represents 93 percent in the course. Then, I want you to create a scatterplot of the new variable **course** versus **pretest**.

```
#
```

Q5 (4 Points)

We want to know if a linear model can be useful for understanding the relationship between the points on the first midterm and the points on the final exam. Since the final exam always comes after the first midterm, assume that **final** is your response variable. Fit a simple linear regression to the mentioned relationship, and use the `summary()` function to show the regression table.

Then, in complete sentences, I want you to interpret the estimate of the slope to a student in the course who may have very little background in statistics. You should explain the slope in the context of the data. Think about what would be important to tell the student so that it is helpful in understanding the impact their score on the first midterm has on their score on the final exam. Make sure you are precise in your language so you are not misleading someone to believe something that isn't guaranteed to be true.

```
#
```

Response in Complete Sentences: REPLACE EVERYTHING IN ALL CAPS HERE WITH YOUR ANSWER IN COMPLETE SENTENCES

Q6 (7 Points)

Now, I want you to fit a simple linear regression based on the relationship between the points on the second midterm and the points on the final exam. Again, assume that the points on the final exam is your response variable. In your output, I don't care how much code you write, but I only want to see the following things in your output:

1. Summary of model using `summary()` function (1 point)
2. Scatterplot showing the raw data with a "red" regression line plotted over the points. (2 points)
3. Boxplot of the residuals (1 point)
4. Scatterplot of the residuals versus the fitted values with a "blue" horizontal line representing perfect prediction. (2 points)
5. Numerical prediction of the number of points on the final exam for a student who scored 80 points on the second midterm. Your prediction should have a minimum of 3 decimal places. (1 point)

```
#
```