Stat 21 Homework 6

Your name here Collaborators: [list any collaborators here]

Due: Nov 10th, by noon ET

This assignment is due on to be submitted on Gradescope on **November 10** by **12:00pm ET**. Please use the `homework-q-and-a` channel on Slack to post any related questions or error messages.

General instructions for all assignments:

You must submit your completed assignment as a single **PDF** document uploaded to **Gradescope**. For instructions on how to do this, please watch this 2 minute video: https://youtu.be/KMPoby5g_nE. You must use R markdown to write up your solutions. For any homework problems that involve coding in R, you must provide **both** the code and the requested output. You can find a R markdown homework template here: http://www.swarthmore.edu/NatSci/sthornt1/Stat021/Stat21.html. Please make sure each problem is **clearly labeled** and that any handwritten components (such as pictures or equations) are easily readable as pictures within the R markdown document. You may want to use a service like CamScanner (https://www.camscanner.com/) to help you upload handwritten pages.

You are allowed to work with your classmates on this homework assignment but you must disclose the names of anyone you collaborate with at the top of your solutions. Each homework assignment is worth 20 points. One problem will be chosen at random to be graded for correctness and the other problems will be graded for completion. At the end of the semester, your lowest homework grade will be dropped. No homework solutions will be provided.

- Use this file as the template for your submission. You can delete unnecessary text (e.g. these instructions) but make sure to keep the nicely formatted "Problem 1", "Problem 2", "a.", "b.", etc
- Upload your knitted HTML or PDF file to the Homework 1 submission section on Gradescope. Name this file as: [SwatID]_stat21_hw06.pdf (e.g. and "sthornt1_stat21_hw06.pdf"). You only need to upload one file, but please make sure that your graphs, code, and answers to each question appear in the appropriate place. If we cannot see your code/graphs/answers, we cant give you credit for your work!
- Your file should contain the code to answer each question in its own code block. Your code should produce plots/output that will be automatically embedded in the output file.
- Each answer must be supported by written statements (unless otherwise specified).
- Include the name of anyone you collaborated with at the top of the assignment.
- In order to knit this document, make sure you have installed the following packages in your version of RStudio: `ggplot2`, `tidyverse`, `gridExtra`, `gcookbook`, `knitr`

Use the code below to load the data set called `mileage` into R Studio. This data describles the gasoline mileage performance for 32 automobiles. The response variable is `mpg` (miles per gallon) and the predictor variables being considered are `displacement`, `weight`, and `transmission_type`. Use this data to answer problems 1-3, and 5.

```
mileage <- read_table2(url("http://www.swarthmore.edu/NatSci/sthornt1/DataFiles/mileage.txt"), skip=2, c
head(mileage)
```

```
## # A tibble: 6 x 5
##   car          mpg displacement weight transmission_type
```

```
##   <chr>      <dbl>        <dbl>  <dbl> <fct>
## 1 Apollo      18.9          350   3910 A
## 2 Omega       17            350   2860 A
## 3 Nova        20            250   3510 A
## 4 Monarch     18.2          351   3890 A
## 5 Duster      20.1          225   3365 M
## 6 JensonConv  11.2          440   4215 A
```

Problem 1

a) Build a linear regression model relating gasoline mileage (as the response) to engine displacement and the type of transmission. (Note that transmission type is a binary categorical variable.) Does the type of transmission significantly affect the mileage performance? Justify your answer.

b) Modify the model developed in part (a) to include an interaction between engine displacement and the type of transmission (A=automatic, M=manual). What is the average effect on gasoline mileage when the engine is automatic? What is the average effect on gasoline mileage when the engine is manual? (It may help to know that engine displacement has to do with the size of the vehicle, so, loosely speaking, larger displacement means a larger vehicle.)

Solution:

[Write your solution here.]

```
## Use this space for any R code for this problem
```

Problem 2

a) Build a linear regression model relating gasoline mileage (as the response) to vehicle weight and the type of transmission. Does the type of transmission significantly affect the mileage performance? Justify your answer.

b) Modify the model developed in part (a) to include an interaction between vehicle weight and the type of transmission. What is the average effect on gasoline mileage when the transmission is automatic? What is the average effect on gasoline mileage when the transmission is manual?

Solution:

[Write your solution here.]

```
## Use this space for any R code for this problem
```

Problem 3

Based off of the results in problems 1 and 2, if you were to build your own MLR to predict the vehicle gasoline mileage, what terms would you include in your regression model and why? (You do not have to fit this model, just base your answer off of different characteristics of the models fit in problems 1 and 2.)

Solution:

[Write your solution here.]

Problem 4

Read the article, "Scientists rise up against statistical significance" at https://www.nature.com/articles/d41586-019-00857-9. Answer each part of this question with 2-4 sentences.

(a) The article claims, "...researchers have been warned that a statistically non-significant result does not 'prove' the null hypothesis (the hypothesis that there is no difference between groups or no effect of

a treatment on some measured outcome)." Explain why failing to reject the null hypothesis does not prove that there is no effect. What does failing to reject the null hypothesis really mean instead?

(b) In the graphic "Beware false conclusions", results are shown from two studies: one that found "significant" results, and another that found "non-significant" results. The article claims that it is "ludicrous" to say that the second study found "no association." Briefly explain why this is the case.

(c) Regarding the same two studies in part (b), the article claims that it is "absurd" to say that the two studies are in conflict, even though one was "significant" and the other was "not significant". Briefly explain why this is the case.

(d) In the section titled "Quit categorizing", the article claims that, "Statistically significant estimates are biased... Consequently, any discussion that focuses on estimates chosen for their significance will be biased." Briefly explain why this is the case.

Solution:

[Write your solution here.]

Problem 5

Now that you've read the article for problem 4, revisit your answer to problem 3 and explain if you would change your final model suggestion or not and why. (2-4 sentences)

Solution:

[Write your solution here.]

Use the code below to load the data set called `airplanes` into R Studio. This data set was collected from a national publication advertising the sale of used aircraft in the early 1990s. The variables included in this dataset include the `year` the aircraft was sold, `TT` (the total flight time in hours), `SMOH` (the number of hours since the major overhaul), `DME` (whether it has distance measuring equipment), `LORAN` (whether it has long-range navigation based on satellite communication), `HP` (engine horsepower), `paint` (whether it has a new or recent paint job), and `price` of sale (in thousands of dollars).

```
airplanes <- read_table2(url("http://www.swarthmore.edu/NatSci/sthornt1/DataFiles/airplanes.txt"), col_
airplanes <- airplanes %>% select(-c(IFR))
head(airplanes)
```

```
## # A tibble: 6 x 9
##   price  year    TT  SMOH DME   LORAN ModeC    HP paint
##   <dbl> <dbl> <dbl> <dbl> <fct> <fct> <fct> <dbl> <fct>
## 1  33.5  1981  3200  1600 1     0     1       161 0
## 2  24.5  1974  3680  1680 0     1     1       161 0
## 3  24.9  1977  1340   300 0     0     0       161 0
## 4  21.9  1974  2090   400 0     0     0       161 1
## 5  27.9  1976  3270   940 0     1     1       161 0
## 6  27.5  1982  6320   320 0     0     1       161 0
```

Problem 6

(a) Make a scatterplot matrix using only the quantitative variables. Comment on any notable features or patterns (or lack of thereof). What is notable about the variable `HP`? What do you think you should do with this variable? (2-3 sentences)

(b) Which data point appears to be an outlier? Can you guess why it might be an outlier? For the remainder of this assignment, we will delete this point from the data for subsequent analyses. Delete

this outlier observational unit from the data using the following code and make a new scatterplot matrix of the quantitative variables. (2-3 sentences)

```
airplanes <- airplanes %>% filter(year>=1970)
```

Solution:

[Write your solution here.]

```
## Use this space for any R code for this problem
```

Problem 7

Fit a model using `price` as the response variable and using all of the predictor variables **except** `SMOH`.

(a) What is the value of R-squared?

(b) What is the estimate of the error variance?

(c) How do you interpret this estimate in the context of the problem?

(d) Which variables appear significant, and which do not?

Solution:

[Write your solution here.]

```
## Use this space for any R code for this problem
```

Problem 8

It seems plausible that `TT` and `year` might be collinear since older planes may have been flown more than newer ones. Based on the scatterplot matrix and the estimated correlation between these variables, does this appear to be the case? (2 sentences or less)

Solution:

[Write your solution here.]

Problem 9

Taking into account all the steps you've preformed in Problems 6-8 decide which predictor variables to include in a final regression model for this data. Write your estimated regression model and discuss whether or not the regression assumptions appear to be satisfied. Don't forget to include a residual plot and an Normal probability plot for the residuals. (4 sentences or less)

Solution:

[Write your solution here.]

```
## Use this space for any R code for this problem
```

Problem 10

Evaluate the fit of your model in Problem 9 by including an interpretation of the estimated regression coefficients (and their roles in the model) and an interpretation of the overall fit of the model. Justify your statements with statistical reasoning. (6 sentences or less)

Solution:

[Write your solution here.]

```r
## Use this space for any R code for this problem
```