Stat 21 Homework 7

Your name here Collaborators: [list any collaborators here]

Due: Tuesday, April 27th, by 8am EST

This assignment is due on to be submitted on Gradescope by **Tuesday, April 27th** by **8:00am EST**. Please use the `homework-q-and-a` and `r-q-and-a` channels on Slack to post any related questions.

Note: You will automatically **lose 5 points** if you do not select the pages associated with the solutions for each of the homework problems when uploading to Gradescope. If you need assistance figuring out how to do this, please see the video below and message me if you still have questions about how to do this!

General instructions for all assignments:

You must submit your completed assignment as a single **PDF** document 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 on Moodle in the Homework section.

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. One problem will be chosen at random to be graded for correctness and the other problems will be graded for completion. No homework solutions will be provided. You may check your answers with others during office hours or anytime outside of class.

- 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 PDF file to the Homework 1 submission section on Gradescope. Name this file as: [SwatID]_stat21_hw07.pdf (e.g. and "sthornt1_stat21_hw07.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 when you upload your file. 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 (.pdf) file.
- Each answer must be supported by a written statement (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`

The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the ACT test score (X_1) and from a binary categorical indicator of whether or not the student had chosen a major field of concentration at the time their application was submitted. The categorical predictor variable is coded so that

$$X_2 = \begin{cases} 0, & \text{if the major field was undecided} \\ 1, & \text{if the student had indicated a major field of concentration at the time of application} \end{cases}.$$

The results of the study are presented in the data below. Use this data to answer Problems 1-4.

```
library("tidyverse")
admissions <- read.csv(url(
              "http://www.swarthmore.edu/NatSci/sthornt1/DataFiles/admissions_data.csv"))
head(admissions)
```

```
##     GPA ACT major
## 1 3.897  21     0
## 2 3.885  14     1
## 3 3.778  28     0
## 4 2.540  22     1
## 5 3.028  21     0
## 6 3.865  31     1
```

Problem 1

Explain the interpretation of each of the regression model coefficients within the context of this data:

$$Y_i \mid x_{1,i}, x_{2,i} = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i,$$

where $x_{2,i} = \begin{cases} 0, & \text{if student } i \text{ is undecided} \\ 1, & \text{if student } i \text{ indicated a major field of concentration at the time of application} \end{cases}$, $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$, and all the errors are independent.

Solution Problem 1:

[Write your solution here.]

```
## Uncomment this line and put any r-code you used for your solution here
```

Problem 2

Fit the regression model from Problem 1 and state the estimated regression equation.

Solution Problem 2:

[Write your solution here.]

Problem 3

Test whether the x_2 variable can be dropped from the regression model using an $\alpha = 0.01$ significance level. State the hypotheses and the conclusion.

Solution Problem 3:

[Write your solution here.]

Problem 4

Create and interpret a plot of the standardized residuals versus the fitted values and a Normal probability plot of the standardized residuals. What can we conclude about the regression model assumptions?

Solution Problem 4:

[Write your solution here.]

```
## Uncomment this line and put any r-code you used for your solution here
```

A hospital administrator wished to study the relation between patient satisfaction (Y), and patient's age (X_1, in years), severity of illness (X_2, an index), and anxiety level (X_3, an index). The administrator randomly selected 46 patients and collected the data presented below, where larger values of Y, X_2, and X_3 are, respectively, associated with more satisfaction, increased severity of illness, and more anxiety. The results of the study are presented in the data below. Use this data to answer Problems 5-10.

```
satisfaction <- read.csv(url(
                "http://www.swarthmore.edu/NatSci/sthornt1/DataFiles/patient_satisfaction.csv"))
head(satisfaction)
```

```
##   satisfaction age severity anxiety
## 1           48  50       51     2.3
## 2           57  36       46     2.3
## 3           66  40       48     2.2
## 4           70  41       44     1.8
## 5           89  28       43     1.8
## 6           36  49       54     2.9
```

Problem 5

Create a histogram for each of the predictor variables. Comment on any noteworthy features of these plots.

Solution Problem 5:

[Write your solution here.]

```
## Uncomment this line and put any r-code you used for your solution here
```

Problem 6

Create a matrix scatter plot for all of the variables (predictor and response). Where do you see indication of a linear relationship?

Solution Problem 6:

[Write your solution here.]

```
## Uncomment this line and put any r-code you used for your solution here
```

Problem 7

Fit the regression model

$$Y_i \mid x_{1,i}, x_{2,i}, x_{3,i} = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \epsilon_i$$

where $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$, and all the errors are independent, to this data and state the estimated regression equation. How is $\hat{\beta}_2$ interpreted here?

Solution Problem 7:

[Write your solution here.]

```
## Uncomment this line and put any r-code you used for your solution here
```

Problem 8

Create a plot of the standardized model residuals by the fitted values and create a Normal probability plot of the standardized residuals. Interpret these plots and summarize your findings.

Solution Problem 8:

[Write your solution here.]

```
## Uncomment this line and put any r-code you used for your solution here
```

Problem 9

For the sake of this exercise, suppose the assumption of independent Normally distributed errors is applicable to this model. Test for the significance of the regression model at the $\alpha = 0.10$ level. State the hypotheses and the conclusion of the test. What does this imply about the model coefficients?

Solution Problem 9:

[Write your solution here.]

```
## Uncomment this line and put any r-code you used for your solution here
```

Problem 10

For the sake of this exercise, suppose the assumption of independent Normally distributed errors is applicable to this model. Obtain a 90% interval estimate of the mean patient satisfaction when $x_1 = 35$, $x_2 = 45$, and $x_3 = 2.2$ and interpret this interval in the context of the problem.

Solution Problem 10:

[Write your solution here.]

```
## Uncomment this line and put any r-code you used for your solution here
```

Bonus Problem

For +2 additional possible homework points, answer the following questions based on on Problem 2 from HW 6.

Suppose we have two random variables X and Y. Use the `rnorm()` function in R to create examples of the following instances:

(a) x_{obs} and y_{obs} are both samples of size 20 from Normal distributions but the samples are uncorrelated. Show the samples are uncorrelated by using the `cor()` function.

(b) x_{obs} and y_{obs} are independent samples from the same Normal distribution. There is no method to verify this so make sure your code clearly demonstrates that each sample is independent of the other.

(c) x_{obs} and y_{obs} both come from a Normal distribution with the same variance but with different means. Show that the samples have the same variance by using the `var()` function.

(d) x_{obs} and y_{obs} are both draws from the same Normal distribution. Verify this by creating a QQ-plot comparing the quantiles of x_{obs} to the quantiles of y_{obs}. (The pseudo-code for this plot is included below but it assumes that you have stored both x_{obs} and y_{obs} in a tibble called `data_part_d` with column names "xobs" and "yobs".)

```
## Code for part (a)


## Code for part (b)


## Code for part (c)


## Code for part (d)
#nq <- 31
#p <- (1 : nq) / nq - 0.5 / nq
#ggplot(data_part_d) +
#  geom_point(aes(x = quantile(xobs, p), y = quantile(yobs, p))) +
#  labs(title="QQ-plot for two samples", y='Y quantiles', x='X quantiles')
```