Stat 21 Homework 6

Your name here Collaborators: [list any collaborators here]

Due: Monday, April 12th, by 8am EST

This assignment is due on to be submitted on Gradescope by **Monday, April 12th** 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_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 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`

Problem 1

Recall the skyscraper data set used in Homework 5 ("skyscraper_data.txt"). This data investigates how the height (in meters) of a skyscraper depends on the number of floors it has. (5 points)

a) Suppose a developer is working on a new building that has taken the 15 years to get the go-ahead. Suppose they are cheekily designing the building to have 15 floors, one for each year of struggle to get the building approved. If the architect needs to know how tall this building may be, would you use a prediction interval or a confidence interval? Justify your answer.

b) As we did in class, create a scatter plot of the observed data and overlay on this plot the estimated regression line and the confidence and prediction bands.

Solution Problem 1:

[Write your solution here.]

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

Problem 2

Suppose we have two random variables X and Y. What are the differences among the following assumptions regarding X and Y:

- X and Y are uncorrelated,
- X and Y are independent,
- X and Y have the same variance, and
- X and Y have the same distribution?

Solution Problem 2:

[Write your solution here.]

Problem 3

Sketch (by hand) residual plots (with predicted response values on the horizontal axis) that show each of the following:

- constant variance and linearity;
- non-constant variance and linearity;
- constant variance and non-linearity;
- non-constant variance and non-linearity.

You can draw these plots on paper and use CamScanner to take a photograph of your drawings. Once you knit this document to a PDF file, you can then convert your image files to PDF files and merge everything together using a website such as smallpdf.com.

Solution Problem 3:

[Write your solution here.]

Problem 4

The number of pounds of steam used per month at a plant is thought to be related to the average monthly ambient temperature. The past year's usages (per 1000 lbs) and temperatures follow.

```
steam_data <- tibble(
  month = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"),
  temperature = c(21,24,32,47,50,59,68,74,62,50,41,30),
  usage = c(185.79,214.47,288.03,424.84,454.68,539.03,621.55,675.06,562.03,452.93,369.95,273.98))
```

(a) Fit a simple linear regression model to the data.

(b) Test for significance of regression.

(c) Plant management believes that an increase in average ambient temperature of 1 degree will increase average monthly steam consumption by 10,000 lb. Do the data support this statement?

(d) Construct a 99% prediction interval on steam usage in a month with average ambient temperature of 58 degrees.

Solution Problem 4:

[Write your solution here.]

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

Problem 5

```
library('MPV')
NFL_data <- table.b1
names(NFL_data)
```

```
##  [1] "y"  "x1" "x2" "x3" "x4" "x5" "x6" "x7" "x8" "x9"
```

Use the code above to read Table B.1 from your textbook into your R library. (Make sure you install the MPV package first!) This data set concerns the performance of the 26 National Football League teams in 1976. It is suspected that the number of yards gained rushing by opponents (variable x8) has an effect on the number of games won by a team (variable y). Fit a simple linear regression model relating games won, y, to yards gained rushing by opponent, x8, and answer the following questions with this model.

(a) Display the analysis-of-variance table for this model and test for the significance of the regression model. State your null and alternative hypotheses and interpret the conclusion in the context of the problem.

(b) Find and interpret a 95% CI on the slope of the yards gained rushing.

(c) What percent of the total variability in the number of games won is explained by this model?

(d) Find a 95% CI on the mean number of games won if opponents' yards rushing is limited to 2000 yards.

Solution Problem 5:

[Write your solution here.]

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

The data shown below present the average number of surviving bacteria in a canned food product and the minutes of exposure to 300 degree Fahrenheit heat. Use this data to answer Problems 6-7.

```
bacteria_data <- tibble(bacteria_count = c(175, 108, 95, 82, 71, 50, 49, 31, 28, 17, 16, 11),
                        minutes_exposure = c(1,2,3,4,5,6,7,8,9,10,11,12))
```

Problem 6

Create a scatter plot of this data and then fit a SLR model with the number of bacteria as the response.

(a) Display the residual plot and calculate the coefficient of determination to comment on the adequacy of this model.

(b) Based on this model, what is the average effect on the bacterial growth per each additional minute of exposure?

Solution Problem 6:

[Write your solution here.]

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

Problem 7

Identify an appropriate transformed model for these data. (Hint: Transform the response variable by taking the logarithm or taking the inverse and using this transformed data as the new response variable. Create residual plots for each transformation you try to select which method you want to use.)

(a) Fit a SLR model to the transformed data and display the residual plot and calculate the coefficient of determination to comment on the adequacy of this model.

(b) What is the average effect on the bacterial growth per each additional minute of exposure?

Solution Problem 7:

[Write your solution here.]

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

For Problems 8-10, read the Wikipedia page for Simpson's Paradox: https://en.wikipedia.org/wiki/Simpson%27s_paradox. Then, import the "Stand your ground" data set using the code below. This data (from 2015) is related to the Stand Your Ground law in Florida. Each observational unit consists of a case where the Stand Your Ground law was a part of the defense strategy, the defendent's race (white or non-white), the victim's race (white or non-white), and a binary variable indicating whether or not the defendant was convicted. With this categorical data we are not going to fit a regression model but we are going to examine this data and look out for Simpson's paradox.

```
FL_stand_your_ground <- read_csv(
  url('http://www.swarthmore.edu/NatSci/sthornt1/DataFiles/stand_your_ground.csv'))
```

Problem 8

Create and print the following tables to summarize the data:

```
1. Defendant's race vs convicted for all observational units;

1. Defendant's race vs convicted for cases with minority victims only;

1. Defendant's race vs convicted for cases with white victims only;

1. The table created by adding Tables 2 and 3 together.
```

Solution Problem 8:

[Write your solution here.]

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

Problem 9

What are the overall conviction rates for minority and white defendants, respectively? What are the conviction rates for minority and white defendants among the cases with minority victims? What are the conviction rates for minority and white defendants among the cases with white victims?

Solution Problem 9:

[Write your solution here.]

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

Problem 10

Explain what is going on here in terms of Simpson's paradox and interpret what this means with respect to racial bias in the criminal justice system.

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 1 from HW 5.

Problem 1

In a survey of 988 men aged 18˘24, the regression equation for predicting height from weight was:

$$height = 62.4 + (0.047)(weight),$$

where height is measured in inches and weight is measured in lbs. Suppose the variance of our model errors is $\sigma^2 = 2$ and assume all regression model assumptions are met.

(a) What is the probability that a 180-pound man is between $65 - 70$ inches tall?

(b) What is the lower 75th quantile height for 180-pound men?