# Stat 21 Homework 8

Your name here  Collaborators: [list any collaborators here]

Due: Tuesday, May 4th, by noon EST

This assignment is due on to be submitted on Gradescope by **Tuesday, May 4th** by **12:00pm 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_hw08.pdf (e.g. and "sthornt1_stat21_hw08.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`

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, col_names = c("car", "mpg", "displacement", "weight", "transmission_type
                       cols(transmission_type = col_factor())) %>% na.omit
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 corresponds to a larger vehicle size.)

**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.]

---

An engineer studied the effect of four variables on a dimensionless factor used to describe pressure drops in a screen-plate bubble column. The variables collected in this experiment are the following

- $y$ is the response variable, a dimensionless factor for the pressure drop through a bubble cap
- $x_1$ is the superficial fluid velocity of the gas (cm/s)
- $x_2$ is the kinematic viscosity
- $x_3$ is the mesh opening (cm)
- $x_4$ is a dimensionless number relating the superficial fluid velocity of the gas to the superficial fluid velocity of the liquid.

Use the code below to import the data into R and use this data to answer Problems 4-6.

```
library('tidyverse')
pressure_drop <- read_table2(url("http://www.swarthmore.edu/NatSci/sthornt1/DataFiles/pressure_drop.txt
head(pressure_drop)
```

```
## # A tibble: 6 x 5
##      x1    x2    x3    x4     y
##   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  2.14    10  0.34 1      28.9
## 2  4.14    10  0.34 1      31
## 3  8.15    10  0.34 1      26.4
## 4  2.14    10  0.34 0.246  27.2
## 5  4.14    10  0.34 0.379  26.1
## 6  8.15    10  0.34 0.474  23.2
```

## Problem 4

Fit a multiple linear regression model relating this dimensionless number, $y$, to the four regressors and test for the significance of the regression model. Then, use individual t-tests to assess the contribution of each regressor in the model. Finally, state and interpret the adjusted R-squared value for this model.

**Solution Problem 4:**

[Write your solution here.]

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

## Problem 5

Fit a multiple linear regression model relating $y$ only to the regressors $x_2$ and $x_3$. Test for the significance of this simpler regression model, use individual t-tests to assess the contribution of each regressor in the model and state and interpret the adjusted R-squared value for this model.

**Solution Problem 5:**

[Write your solution here.]

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

## Problem 6

For each of the models in Problems 4 and 5, perform a thorough residual analysis. For both models address the following

a) Are the assumptions necessary for estimation reasonable?

b) Are the assumptions necessary for inference reasonable?

c) Find a 99% CI for the regression coefficient of the predictor $x_2$ in both models. Briefly explain any differences and state which interval you would report if you had to pick one.

**Solution Problem 6:**

[Write your solution here.]

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

---

A data set called `msleep` is automatically included in the `tidyverse` package. Use the code below to clean up this data set so we can fit a MLR model to the data relating the total time slept per night (in hrs), ($y$), to the quantitative predictor measuring the animal's brain weight (in kg), $x_1$, and to the conservation status of the animal which can be on of three levels: domesticated, least concern, or high concern. Use this data, `msleep3`, to answer Problems 7 and 8.

```
msleep2 <- msleep %>%
           na.omit %>%
           mutate(conservation_cat =
                    conservation %>% factor %>% fct_collapse("hc" = c("vu","en","nt")))
msleep3 <- msleep2 %>% mutate(conservation_cat2 = relevel(conservation_cat, ref="domesticated"))
head(msleep3)
```

```
## # A tibble: 6 x 13
##   name  genus vore  order conservation sleep_total sleep_rem sleep_cycle awake
##   <chr> <chr> <chr> <chr> <chr>              <dbl>     <dbl>       <dbl> <dbl>
## 1 Grea~ Blar~ omni  Sori~ lc                  14.9       2.3       0.133   9.1
## 2 Cow   Bos   herbi Arti~ domesticated         4         0.7       0.667  20
## 3 Dog   Canis carni Carn~ domesticated        10.1       2.9       0.333  13.9
## 4 Guin~ Cavis herbi Rode~ domesticated         9.4       0.8       0.217  14.6
## 5 Chin~ Chin~ herbi Rode~ domesticated        12.5       1.5       0.117  11.5
## 6 Less~ Cryp~ omni  Sori~ lc                   9.1       1.4       0.15   14.9
## # ... with 4 more variables: brainwt <dbl>, bodywt <dbl>,
## #   conservation_cat <fct>, conservation_cat2 <fct>
```

## Problem 7

Fit the following main effects MLR model to the data

$$Y \mid (x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon,$$

where $Y$ is the total time slept per night (in hrs), $x_1$ is the animal's brain weight (in kg) and $x_2$ and $x_3$ are indicator variables that identify the conservation status of the animals in the following way

$$x_2 = \begin{cases} 1, & \text{if animal is of least concerning status (besides domesticated)} \\ 0, & \text{otherwise} \end{cases},$$

$$x_3 = \begin{cases} 1, & \text{if animal is of highest concerning status} \\ 0, & \text{otherwise} \end{cases}.$$

Based on this model state the following

(a) the estimated regression equation;

(b) the average effect of brain weight on time slept;

(c) the average effect on time slept for mammals of high concern;

(d) the average effect on time slept for mammals of least concern.

**Solution Problem 7:**

[Write your solution here.]

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

## Problem 8

Fit the following interaction effects MLR model to the data

$$Y \mid (x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + +\beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \epsilon,$$

where all the variables are defined in Problem 7.

Based on this model state the following

(a) the estimated regression equation;

(b) the average effect of brain weight on time slept;

(c) the average difference in time slept between mammals in the highest concern category and the least concern category;

(d) the average effect on time slept for mammals of least concern status.

**Solution Problem 8:**

[Write your solution here.]

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

---

A solid-fuel rocket propellant loses weight after it is produced. Suppose we want to model the weight loss(in kg) as a function of the number of months since the rocket was produced. Read the following observed data into R using the code below and use this to answer Problems 9-10.

```
rocket_data <- tibble(months = c(0.25,0.50,0.75,1.00,1.25,1.50,1.75,2.00,2.25,2.50),
                      weight_loss = c(1.42,1.39,1.55,1.89,2.43,3.15,4.05,5.15,6.43,7.89))
```

The code below fits what is called a second-order polynomial regression model to this data.

```
mod9 <- lm(weight_loss ~ poly(months,2), rocket_data)
```

The regression model here is

$$Y \mid x = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon,$$

where $Y$ is the weight loss and $x$ is the months since production.

## Problem 9

First, we are going to perform a thorough residual analysis based on this second-order polynomial regression model. Create and clearly label the following plots based on this model

(a) a scatter plot of $x$ and $y$;

(b) a scatter plot of the studentized residuals and $x$;

(c) a scatter plot of the studentized residuals and $x^2$;

(d) a plot of the studentized residuals vs the fitted values;

(e) a Normal probability plot of the studentized residuals.

Then, state which of the regression model assumptions seem reasonable for these data.

**Solution Problem 9:**

[Write your solution here.]

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

## Problem 10

Test for the significance of the quadratic term and state whether or not we should delete the quadratic term from the model based on the results of this test and the information from Problem 9.

**Solution Problem 10:**

[Write your solution here.]

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

## Bonus Problem

For +5 bonuse HW points, 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.