Stat 21 Test 2

Due: Dec 8, 2020 by noon ET

This test is due on to be submitted on Gradescope on **December 8** by **12:00pm ET**. Please use the `#test_questions` channel on Slack to post any clairfication questions. Do not ask questions like "Is [this] the right answer?"

You must submit your solutions as a single **PDF** document uploaded to **Gradescope**. You may use R markdown to write up your solutions alone or you may use R markdown and hand-written solutions. You must show all of your work, including code input and output. Please make sure each problem is **clearly labeled** and that any handwritten components (such as pictures or equations) are easily readable in the PDF document. You may want to use a service like CamScanner (https://www.camscanner.com/) to help you upload handwritten pages and Small PDF (https://smallpdf.com/merge-pdf) to merge multiple PDFs into a single document.

You are permitted to reference all class material and use the internet (though I am not sure it will be very helpful). You are not permitted however, to get assistance from any person online or otherwise.

- 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 written statements and relevant plots.
- Each problem is worth 50 points for a total of 100 points possible.
- In order to knit this document, make sure you have installed the following packages in your version of RStudio: `ggplot2`, `tidyverse`, `gridExtra`, `knitr`

Future adjustments: make students select variables to use in the final mode (multicollinearity is a big issue here). Maybe have them try two versions and compare the two. Only Height or Drop shoul be used (not both) so eliminate one from the original data set for use next time.

Suppose we have randomly surveyed 91 roller coasters across the US. We are interested in investigating the relationship among a set of predictor variables with the quantitative response variable, the maximum speed of the coaster (mph). The quantitative predictor variables we are going to consider include

- the length of the track (in feet),
- the duration of the ride (in seconds),
- the highest climb the ride reaches (in feet), and
- the lowest drop the ride reaches (in feet).

The categorical predictor variables we are going to consider are

- the type of track (wooden or steel) and
- whether or not the ride has a loop/inversion (1 for yes, 0 for no).

Use the R code below to import this data set into RStudio. (This code makes sure that there are no missing data entries in the sample.)

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

```
## # A tibble: 91 x 7
```

```
##    Track Speed Height  Drop Length Duration Inversions
##    <chr> <dbl>  <dbl> <dbl>  <dbl>    <dbl>      <dbl>
##  1 Steel  35     51.8  49.8  1312.       95          0
##  2 Steel  41     94    66    2423        60          1
##  3 Steel  42     70    64    3100       120          0
##  4 Steel  45     56    47     635        66          1
##  5 Steel  49.7  108.   95.2  2625.      130          1
##  6 Steel  49.7  108.   95.2  2625.      130          1
##  7 Steel  49.7  108.   95.2  2625.      130          1
##  8 Steel  49.7  108.   95.2  2625.      130          1
##  9 Steel  49.7  108.   95.2  2625.      130          1
## 10 Steel  49.7  108.   95.2  2625.      130          1
## # ... with 81 more rows
```

Note: For this data set, we are given that the data is a random sample so we can assume that it is representative of all roller coasters in the US. You can also assume that the data is independent for this example.

Problem 1 (5 points)

Before fitting a model, the first step is to process your data. Preform any necessary processing steps here and briefly justify each step. ~~We are eventually going to try to determine which predictor variables have the largest effect on coaster speed.~~

```
coasters2 <- coasters %>% mutate(Inversions_cat = factor(Inversions), ## Step 1
                                 Track_cat = factor(Track))           ## Step 2
```

[Write your explanation of your steps here. For example...

Step 1: Then write your justification for each step down here.]

Problem 2 (5 points)

Create a scatter plot matrix for all of the quantitative variables and create box plots for each of the categorical variables (with coaster speed on the vertical axis). Does there appear to be any evidence of multicollinearity among the quantitative predictors?

```
## Write your code for the matrix of scatter plots here
```

```
## Write your code for the box plots for Track type here
```

```
## Write your code for the box plots for Inversions here
```

[Write your answer to Problem 2 here.]

Problem 3 (5 points)

Write the equation for the estimated MLR model that includes all of the predictor variables. Make sure you clearly define all of your variables including any indicator variables.

```
## Put your code to fit the model here
```

[Write your answer to Problem 3 here.

You may write the equation out with words like this: y_hat = beta0_hat + beta1_hat*x1

Or you can write a mathematical equation between dollar signs like this: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$. (Just make sure there is no white space space immediately after the first dollar sign or immediately before the second dollar sign!)]

Problem 4 (10 points)

Before we can use the model from Problem 3, we need to check some assumptions by investigating the residual plot. Plot the standardized residuals against the fitted values and create a normal probability plot for the standardized residuals. What can you conclude about your estimated MLR model for this data set based on these two plots?

```
## Create your residual scatter plot here
```

```
## Create your normal probability plot here
```

[Write your answer to Problem 4 here.]

Problem 5 (20 points)

(a) In the summary output for this regression model there are two R-squared values provided. Which R-squared value should we use and why? What does this value represent (in the context of these data)? (10 points)

(b) What is $\hat{\sigma}$ based on this model? What does this number represent in the context of these data? (5 points)

(c) What is the average difference in the speed of a steel roller coaster versus a wooden roller coaster (given all other input variables are the same)? (5 points)

```
## Put your code for Problem 5 here
```

[Your answer to Problem 5 goes here.]

Problem 6 (10 points)

Use the `filter` function to create a new data set that contains only the rows of data corresponding to roller coasters that do not have any inversions (i.e. loops).

(a) Write out the estimated regression equation. (4 points)

(b) Create a residual plot. (4 points)

(c) Create a normal probability plot for the standardized residuals.(2 points)

```
## Code for Problem 6 part a
```

```
## Code for Problem 6 part b
```

```
## Code for Problem 6 part c
```

[Write your answer for Problem 6 here.]

Problem 7 (10 points)

Use the `filter` function to create a new data set that contains only the rows of data corresponding to roller coasters that do have inversions.

(a) Write out the estimated regression equation. (4 points)

(b) Create a residual plot. (4 points)

(c) Create a normal probability plot for the standardized residuals. (2 points)

```
## Code for Problem 7 part a
```

```
## Code for Problem 7 part b
```

```
## Code for Problem 7 part c
```

[Write your answer for Problem 7 here.]

Problem 8 (Table 1 - 5 points, Table 2 - 10 points)

Based on the previous problems, fill out the following tables (by replacing the XX's with the correct numbers). The first table compares the coefficient of determination of the three different models from Problems 3, 6, and 7. The second table compares the confidence intervals and prediction intervals for each of the three models. For the second table, use the input values of:

- Track = Steel
- Height = 111 ft
- Drop = 95 ft
- Length = 2555 ft
- Duration = 75 s

Model	R squared	$\hat{\sigma}$
MLR - Problem 3	XX	XX
MLR - Problem 6	XX	XX
MLR - Problem 7	XX	XX

Model	CI for mean response	PI for new response
MLR - Problem 3	[XX, XX] (Inversions)	[XX, XX] (Inversions)
	[XX, XX] (No Inversions)	[XX, XX] (No Inversions)
MLR - Problem 6	[XX, XX]	[XX, XX]
MLR - Problem 7	[XX, XX]	[XX, XX]

```
## Code for Problem 8 here
```

Problem 9 (20 points)

You've now analyzed this data with two different methods (and three different models). The question I'm sure you've been waiting for is... which analysis should we use and why? Using the information from the previous problems, should we use `Inversions` as a predictor variable in a single MLR model for this data OR should we separate the data and analyze the speed of roller coasters with and without loops separately OR does it even matter? Provide a definite, but succinct answer. (No more than 5 sentences.)

[Write your answer here.]

Extra Credit

If the response rate for the course evaluation is higher than 85%, everyone will get 2 points added to their grade for this test ;)