Stat 21 Test 3

Due: May 16, 2020 by 12:00pm EST

This test is due on to be submitted on Gradescope by **May 16** at **12:00pm EST**. Please use the `#test3_questions` channel on Slack to post any clairfication questions. Do not ask questions like "Is [this] the right answer?" or "How can I get R to produce [this] plot?"

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. You are not permitted however, to get assistance from any other 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.
- In order to knit this document, make sure you have installed the following packages in your version of RStudio: `ggplot2`, `tidyverse`, `gridExtra`, `knitr`.
- If you are running into issues with the Swarthmore RStudio server, you may submit your soulutions to me as an RMarkdown document in Slack via DM. Name your file using the convention: `[SwatID]_stat21_test3.rmd`. You may use this online R compiler to double check the the R code chunks of your document: https://rdrr.io/snippets/.
- If at any point you want to use the 'select' function to select specific columns from a data object, use 'dplyr::select' rather than 'select'.

You may write mathematical equations 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 1

a) How does drinking alcohol affect average life expectancy?

Solution: [Write your solution here.]

b) Suppose someone consumes about 2.5 alcoholic drinks per day and smokes regularly. Calculate (provide the formula for) the average life expectancy for this individual.

Solution: [Write your solution here.]

```
## Put any R code you use for your calculations here
```

c) What is the average difference in lifespan between smokers and non-smokers? Show your work.

Solution: [Write your solution here.]

```
## Put any R code you use for your calculations here
```

d) Supposing we fix the amount of alcohol consumption; is there a statistically significant relationship between life expectancy and smoking status? Justify your answer.

Solution: [Write your solution here.]

e) How much of the variation in lifespan can be explained by smoking status and alcohol consumption?

Solution: [Write your solution here.]

f) Which of the questions above are questions of statistical inference?

Solution: [Write your solution here.]

Problem 2

a) Write the down the main effects equations (in numbers) for predicting the average tooth growth for Guinea pics who:

```
1) Have supplement type absorbic acid and dosage of 1.0 mg
2) Have supplement type absorbic acid and dosage of 2.0 mg
3) Have supplement type orange juice and dosage of 1.0 mg
4) Have supplement type orange juice and dosage of 2.0 mg
```

Solution:

1. [Write the equation for a.1 here.]
2. [Write the equation for a.2 here.]
3. [Write the equation for a.3 here.]
4. [Write the equation for a.4 here.]

b) Write the down the interaction effects equations (in numbers) for predicting the average tooth growth for Guinea pics who:

```
1) Have supplement type absorbic acid and dosage of 1.0 mg
2) Have supplement type absorbic acid and dosage of 2.0 mg
3) Have supplement type orange juice and dosage of 1.0 mg
4) Have supplement type orange juice and dosage of 2.0 mg
```

Solution:

1. [Write the equation for b.1 here.]
2. [Write the equation for b.2 here.]
3. [Write the equation for b.3 here.]
4. [Write the equation for b.4 here.]

c) Based on the R output above and the residual plots on the next page, which model do you think is a better choice, the one with interactions or the one without interactions? Justify your answer.

Solution: [Write your solution here.]

d) For whichever model you chose in part (c), explain the relationship between supplement type and tooth growth in language that can be understood by an average high school student.

Solution: [Write your solution here.]

Problem 3

```r
biomass <- read_table2(url(
  "http://www.swarthmore.edu/NatSci/sthornt1/DataFiles/biomass_data.txt"), col_names = TRUE)
```

a) What is the average (absolute) difference in biomass (in gm^{-2}) for moss found in Smith Island versus moss found in the Snows Marsh? Show your work.

Solution: [Write your solution here.]

```r
## Put any R code you use for your calculations here
```

b) What is the average (absolute) difference in biomass (in gm^{-2}) if the potassium (variable K) in the soil increases by 500 ppm? Show your work.

Solution: [Write your solution here.]

```r
## Put any R code you use for your calculations here
```

c) The code below considers the same predictor, fit by the same exact data only now the quantitative predictor variables have been standardized. (The residual plot and Normal probability plots for this model are also shown on the next page. Read the titles carefully!) Based on this version of the model, what is the average (absolute) difference in biomass (in gm^{-2}) if the potassium (variable $\boldsymbol{K}$) in the soil increases by 500 ppm? Show your work.

Solution: [Write your solution here.]

```
## Put any R code you use for your calculations here
```

(d) Why might we want to consider the standardized model in part (c) rather than the original model?

Solution: [Write your solution here.]

Problem 4

a) Based on this model of the transformed response, what is the average difference in biomass when the potassium in the soil increases by 500 ppm? Show your work.

Solution: [Write your solution here.]

```
## Put any R code you use for your calculations here
```

c) Are the answers to part (a) of Problem 4 and part (b) of Problem 3 directly comparable? Why or why not?

Solution: [Write your solution here.]

Problem 5

```
coasters <- read_table2(url(
  "http://www.swarthmore.edu/NatSci/sthornt1/DataFiles/roller_coasters.txt"),
  col_types = "f?????f") %>%
  filter(!is.na(Duration) & !is.na(Inversions))
```

(a) Create a scatter plot matrix showing only the quantitative variables (predictors and response) 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?

Solution: [Write your solution here.]

```
## Put your R code here
```

(b) Fit the following four regression models to this data and write out the corresponding estimated regression equations as well as the adjusted coefficient of determination for each model.

Model 1: $E[Y|\text{length, height}] = \beta_0 + \beta_1\text{length} + \beta_2\text{height} + \epsilon$

Model 2: $E[Y|\text{length, height, track}] = \beta_0 + \beta_1\text{length} + \beta_2\text{height} + \beta_3 w_{track} + \epsilon$, where $w_{track} = \begin{cases} 1, & \text{if steel} \\ 0, & \text{otherwise} \end{cases}$

Model 3: $E[Y|\text{duration, drop}] = \beta_0 + \beta_1\text{duration} + \beta_2\text{drop} + \epsilon$

Model 4: $E[Y|\text{duration, drop, loop}] = \beta_0 + \beta_1\text{duration} + \beta_2\text{drop} + \beta_3 w_{loop} + \epsilon$, where $w_{loop} = \begin{cases} 1, & \text{if has inversion} \\ 0, & \text{otherwise} \end{cases}$

Solution:

1. [Write the equation for model 1 here.]
2. [Write the equation for model 2 here.]
3. [Write the equation for model 3 here.]
4. [Write the equation for model 4 here.]

```
## Put your R code here
```

(c) Perform a thorough analysis of the studentized residuals for each of the four models you fit in part (b). Make sure every plot is clearly labeled and include at least a one sentence statement explaining the relevance of each plot. (Hint: Make sure you load the library("MASS").)

Solution:

```
## Put your R code for model 1 here
```

[Interpret the plots for model 1 here.]

```
## Put your R code for model 2 here
```

[Interpret the plots for model 2 here.]

```
## Put your R code for model 3 here
```

[Interpret the plots for model 3 here.]

```
## Put your R code for model 4 here
```

[Interpret the plots for model 4 here.]

(d) Suppose your friend's favorite ride is the Wicked Twister and they want to know what is the estimated maximum speed this coaster reaches. This a steel roller coaster that reaches a height of 215 ft, has a drop of 206 ft, is 675 ft long, lasts for 40 seconds, and has no inversions. Based on your answers to parts (a)-(c), which of the regression models are appropriate to use to answer your friend's question about the Wicked Twister and why?

Solution: [Put your answer here.]

(e) Wikipedia lists the speed of the Wicked Twister as 72 mph. Suppose we want to determine if there is a statistically significant difference between the speed reported by Wikipedia and the estimated maximum speed based on one of our models. Explain how we could do this and state which models from part (b) (if any) are appropriate to use to answer this question. Explain your answer.

Solution: [Put your answer here.]