Stat 21 Homework 5

Your name here Collaborators: [list any collaborators here]

Due: Monday, April 5th, by 8am EST

This assignment is due on to be submitted on Gradescope by **Monday, April 5th** 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_hw05.pdf (e.g. and "sthornt1_stat21_hw05.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

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

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

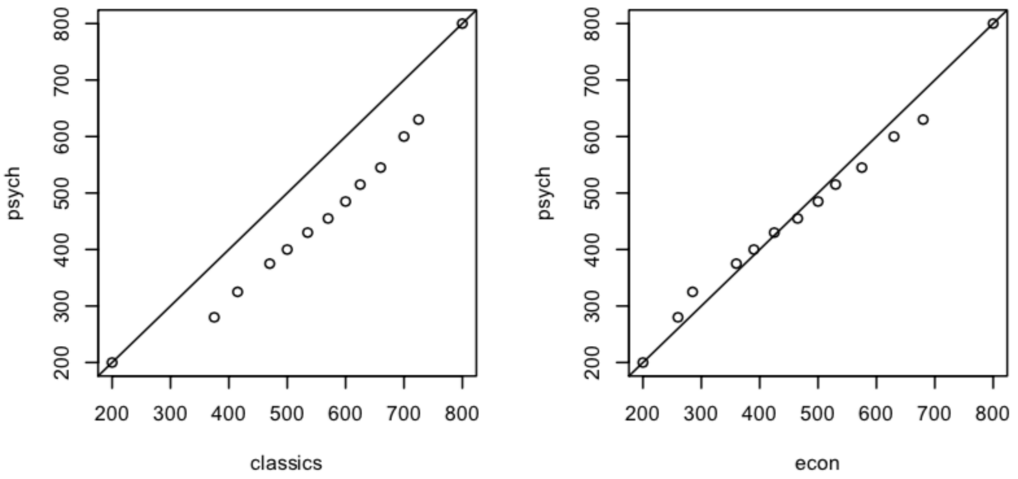
where height is measured in inches and weight is measured in lbs.

(a) Is the following statement a correct interpretation of the regression line: “If someone gains 10 pounds, he will get taller by $(0.047)(10) = 0.47$ inches”? If not, provide a better explanation for the meaning of the slope.

(b) Suppose the variance of our model errors is $\sigma^2 = 2$. What percentage of all 200-pound men are taller than 74 inches? (Assume the regression model assumptions are met.)

Problem 2

Below are two Quantile-quantile plots of GRE General Test Verbal scores for students intending graduate study in psychology, classics, and economics. Here we are comparing the psychology test scores to the classics and to economics. We are interested in how the population psychology students compares to the population of classics students and to the population of economics student. How do you interpret the patterns (deviations from the diagonal lines) in these plots? Explain in 2-3 sentences.



Problem 3

Suppose we are interested in studying the effectiveness of the recycling/composting programs at Swarthmore. I.e. we are investigating the waste that is disposed in the trash/recycle/compost bins across campus. Use your imagination to come up with three different research questions related to this topic in the case where

1. We have two numerical variables of interest;
2. We have one numerical variable of interest and one categorical variable of interest;
3. We have two categorical variables of interest.

For each setting clearly state the variables of interest, the roles of the variables (predictor/response), and the statistical research question you wish to answer. Each research question you come up with should be answerable by one of: a simple linear regression, an ANOVA model, or a chi-squared test.

Problem 4

Let's consider the data set called *msleep* which is contained in the R package `ggplot2`.

```
library(ggplot2)
head(msleep)
```

```
## # A tibble: 6 x 11
##   name  genus vore  order conservation sleep_total sleep_rem sleep_cycle awake
##   <chr> <chr> <chr> <chr> <chr>              <dbl>     <dbl>       <dbl> <dbl>
## 1 Chee~ Acin~ carni Carn~ lc                  12.1      NA        NA      11.9
## 2 Owl ~ Aotus omni  Prim~ <NA>                17         1.8      NA       7
## 3 Moun~ Aplo~ herbi Rode~ nt                  14.4       2.4      NA       9.6
## 4 Grea~ Blar~ omni  Sori~ lc                  14.9       2.3       0.133   9.1
## 5 Cow   Bos   herbi Arti~ domesticated         4         0.7       0.667  20
## 6 Thre~ Brad~ herbi Pilo~ <NA>                14.4       2.2       0.767   9.6
## # ... with 2 more variables: brainwt <dbl>, bodywt <dbl>
```

This data set looks at the amount of time spent sleeping for different mammals and records other factors such as brain and body weight of these animals. Suppose we are interested in the total amount of sleep an animal gets (variable name `sleep_total`) as predicted by the total body weight of the animal (variable name `bodywt`).

a) There is a qualitative variable named *order* in this data set. Fit two separate linear regression models for the animals of *order* "Carnivora" and of *order* "Primates". Report the estimated regression equations and print the summary of the two linear models.

b) What is the estimate of the variance of the random error for each of these two models?

c) Do you think we should combine the data and just fit a single linear regression model to both *orders* Carnivora and Primates? Justify your answer with 1-2 sentences and (possibly) a supporting plot.

The following data set was collected in 2018-2019 and recorded different attributes of skyscrapers in NYC. In Problems 5-10 we are going to investigate how the height (in meters) (Y) of a skyscraper depends on the number of stories (i.e. floors) it has (x).

```
skyscrapers <- read_csv(
                url("http://www.swarthmore.edu/NatSci/sthornt1/DataFiles/skyscraper_data.csv"))
head(skyscrapers)
```

```
## # A tibble: 6 x 8
##      ID Building_name    height_meters height_ft floors  year material  purpose
##   <dbl> <chr>                    <dbl>     <dbl>  <dbl> <dbl> <chr>     <chr>
## 1     1 30 Hudson Yards           387.      1268     73  2019 concrete~ office
## 2     2 3 World Trade C~          329.      1079     69  2018 composite office
## 3     3 35 Hudson Yards           308.      1010     71  2019 concrete  residen~
## 4     4 220 Central Par~          290.       952     70  2019 concrete  residen~
## 5     5 15 Hudson Yards           279.       914     70  2019 concrete  residen~
## 6     6 The Centrale              245.       803     64  2019 concrete  residen~
```

Problem 5

Make a scatter plot of the observed predictor and response variables and report

(a) The estimated regression equation.

(b) The value of the standard deviation of height.

(c) The value of R-squared.

Problem 6

(a) Make a residual plot for the regression model you fit in Problem 5. Are there any apparent violations of the regression model assumptions? Explain your answer in four sentences or less.

(b) Make a Normal probability plot of the standardized residuals to determine if the residuals look like they come from a Normal distribution. Interpret the plot in three sentences or less.

Problem 7

(a) Calculate a 95% confidence interval for the model parameter β_1, the slope of the variable `floors`. How would you explain the meaning of this confidence interval, in the specific context of this data set, to an architect who has never taken a statistics class?

(b) Test the hypothesis that $\beta_1 = 0$ at an $\alpha = 0.05$ significance level. State your null and alternative hypotheses and report the test statistic and p-value. Interpret, in the context of the problem, the results of this test in two sentences or less.

Problem 8

(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 report to them a prediction interval or a confidence interval? Justify your answer in three sentences or less.

(b) Create a scatter plot of the observed data and overlay on this plot the estimated regression line and the confidence and prediction bands.

Problem 9

Note that there is a categorical variable called "purpose" included in the data set. Suppose we are interested in determining if there is a significant difference in the average height of a building depending on its purpose. Using `height` (in meters) as the response and purpose as the explanatory variable, fit an ANOVA model to this data after excluding the data point for the only hospital (see code below). What does the result of the ANOVA F-test indicate?

```
skyscrapers_edit <- skyscrapers %>% filter(!(purpose=="hospital"))
```

Problem 10

We know that the significance level (α) represents the probability of a false positive (i.e. a type I error) in a hypothesis test. Related to this concept is the probability of correctly detecting a positive. In statistics, this probability is called the **power** of a test and is often denoted as $1 - \beta$ where β is the probability of a type II error. (Hence the power of a test is the probability of NOT making a type II error.) What if we wanted to collect more skyscraper data to increase the power of our test in Problem 9? Using this example as a guide: https://stats.idre.ucla.edu/r/dae/one-way-anova-power-analysis/ and assuming we can collect enough data so that all categories for the variable `purpose` have the same number of observations (i.e. we have a balanced ANOVA design), how many more observations of hospitals, hotels, offices, and residential buildings specifically do we need to achieve 85% power?

Bonus Problem

For +2 additional possible homework points, answer the following questions regarding Problems 7-10 from HW 4.

(a) What is the predictor variable and what is the response variable for this statistical model?

(b) Without standardizing the response variable, what are the p-value for the ANOVA F-test and the MSE based on this ANOVA model?

(c) Without standardizing the response variable, display the Normal probability plot for states without any voter ID laws.

Now create a new data object called `voter_turnout2` that contains a third column called `turnout_rate_std` that corresponds to the standardized version of `turnout_rate`.

(d) Using `turnout_rate_std` as the response variable, what are the p-value for the ANOVA F-test and the MSE based on this ANOVA model?

(e) Using `turnout_rate_std` as the response variable, display the Normal probability plot for states without any voter ID laws.

(f) Compare your answers from parts (b) and (c) to parts (d) and (e). Briefly explain any differences or similarities you see.

HW 4 setting:

The increase in the passage of voter identification laws ("Voter ID") has been labelled as a push to reduce fraudulent or illegal voting in the United States. Most of these laws have come into effect by requiring voters to present a photo ID either at time of registration or at the time of voting. The National Conference of State Legislatures (NCSL) identified four different levels of voter identification requirements: Strict Photo ID Law, Non-Strict Photo ID, Strict Non-Photo ID Law, and NonStrict Non-Photo ID. Non-strict identification laws are defined as "at least some voters without acceptable identification have an option to cast a ballot that will be counted without further action on the part of the voter". Strict identification laws are defined as "voters without acceptable identification must vote on a provisional ballot and also take additional steps after Election Day for it to be counted."

In 2004, not a single state enforced a photo identification requirement at the polls, but since then the numbers have increased drastically. Although designed to reduce fraudulent voting, several studies have found that there are a very small number of fraudulent votes to begin with, so the laws have very little impact. Rather, these laws have acted to raise the costs of voting and reduce convenience, especially for certain groups of voters who are unlikely to have a photo ID.

The following (real) data set contains information on what level of voter identification law was in place for each state in the 2016 US presidential election (no low, non-strict non-photo ID law, strict non-photo ID law, non-strict photo ID law, strict photo ID law) and what was the percentage of voter turnout for each state that year. (No law states do not require a government issued ID in order to submit a ballot.)

```
voter_turnout <- tibble(law_type = c(rep("No law",18),
                                 rep("Non-strict non-photo", 14),
                                 rep("Strict non-photo",2),
                                 rep("Non-strict photo",8),
                                 rep("Strict photo",8)),
                   turnout_rate = c(52.8,56.5,61.6,68.4,70.7,66.4,67.2,74.1,62.8,57.4,64.4,
                                    54.5,56.8,64.5,66.2,63.6,63.7,59.5,50.2,64.7,56.8,
                                    52.3,60.8,71.4,61.8,62.2,58.6,64.2,63.7,70,61,54.9,56.7,
                                    62.9,51.4,58.6,59.1,64.7,60,59.2,42.3,64.5,
                                    69.5,66.1,51.1,55.2,57.7,56.4,59.1,58.8))
```