DS 301 Homework 3
Due: Feb. 16, 2022 on Canvas by 11:59 pm (CT)

Instructions: Homework is to be submitted on Canvas by the deadline stated above. Please clearly print your name and student ID number on your HW.

Show your work (including calculations) to receive full credit. Please work hard to make your submission as readable as you possibly can - **this means no raw `R` output or code** (unless it is asked for specifically or needed for clarity).

Code should be submitted with your homework as a separate file (for example, a .R file, text file, Word file, or .Rmd are all acceptable). You should mark sections of the code that correspond to different homework problems using comments (e.g. `##### Problem 1 #####`).

Problem 1: Patient Data

You're working with a public health scientist as a consultant for their data analysis. They are interested in understanding the relationship between patient satisfaction (score from 0 to 100) with other predictors (age, severity of disease, and overall anxiety). The data they have collected can be found in the `patient.txt` dataset posted on Canvas. Read the data into `R` using the `read.table()` function.

You may use the following code:

```
patient = read.table("/.../patient.txt",header=FALSE)

# Instead of /.../, specify the pathway to the folder that you've saved the data set in.
# For example, if I saved the data in the Downloads folder my pathway would be:
# patient = read.table("/Users/lchu/Downloads/patient.txt",header=FALSE)

names(patient) = c("satisf","age","severe","anxiety")

head(patient) # check to make sure the data was read in correctly.
```

a. Fit a multiple linear regression model with patient satisfaction as the response (Y) and age (x_1), severity of disease (x_2), and anxiety (x_3) as predictors. Call this `model1`. What are the least-square estimators and their standard errors? Summarize your output in a table here.

b. Report the residual sum of squares for `model1`.

c. Report the residual sum of squares for the model with no predictors. You can fit this model using the following syntax: `null = lm(y ~ 1, data = patient)`, where y is patient satisfaction. Explain why the RSS of the null model must be larger than the RSS of `model1`.

d. Is `model1` a significant improvement over the model with no predictors? In other words, does at least one of the predictors have a relationship with Y? Carry out the appropriate

hypothesis test at $\alpha = 0.05$. You may assume the random errors (ϵ_i) are normally distributed. Write out the null/alternative hypothesis, test statistic, null distribution, decision rule, and conclusion.

e. Choose one regression coefficient and test whether it is zero or not at $\alpha = 0.05$. Write out the null/alternative hypothesis, test statistic, null distribution, decision rule, and conclusion.

f. Based on the model, obtain an estimate for the average satisfaction score for patients of age 77, disease severity 68, and an anxiety of 3. Quantify the uncertainty surrounding this estimate. Does the range of values make sense? **Explore the data** to justify your reasoning. Discuss what this tell us about the limitations of our model.

g. Explain the difference between what the function `predict()` outputs compared to what `model1$fitted.values` outputs in `R`.

h. Obtain an estimate for σ^2.

Problem 2: Consulting

Based on the results from Problem 1, the public health scientist has the following questions:

a. The scientist sees that you have set the significance level to be $\alpha = 0.05$. He wants to know what this $\alpha = 0.05$ means in the context of hypothesis testing. Explain.

b. The p-value for the predictor `anxiety` is 0.0647. It is not significant at $\alpha = 0.05$ so the scientist claims that the predictor is not meaningful and suggests fitting a model without this predictor. Do you agree or disagree with their claim? Justify your answer.

c. There are future plans to collect additional predictors to better understand factors affecting patient satisfaction. Suppose there will be a total of 12 predictors collected in the future. The scientist wants to determine whether or not at least one of these predictors is useful in predicting Y. He proposes fitting a model with all 12 predictors and then carrying out 12 individual t-test for each regression coefficient. If it at least one result is significant, he can conclude at least one of the predictors is useful in predicting Y. Explain in plain language why this might be a bad idea. What is the probability of seeing at least one significant result by chance? Use $\alpha = 0.1$.

Problem 3: Carseats Data

For this problem, we will use the `Carseats` data set which is part of the `ISLR2` package. To access the data set, load the `ISLR2` package into your `R` session:

```
library(ISLR2) #you will need to do this every time you open a new R session.
```

To get a snapshot of the data, run `head(Carseats)`. To find out more about the data set, we can type `?Carseats`.

We will now try to predict carseat unit sales (in thousands) using the other variables in this data set.

a. Fit a multiple linear regression model to predict carseat unit sales (in thousands) using all other variables as your predictors. What are the least-square estimates and their standard errors? Summarize your output in a table.

b. Assume that our random errors (ϵ_i) are normally distributed. Carry out the F-test at $\alpha = 0.05$. Write out the null/alternative hypothesis, test statistic, null distribution, p-value, and conclusion.

c. Choose one regression coefficient and test whether it is zero or not at $\alpha = 0.05$. Write out the null/alternative hypothesis, test statistic, null distribution, p-value, and conclusion.

d. Obtain an estimate for σ^2.

e. Interpret the R^2 from the fitted model.

f. Interpret the regression coefficients associated with Shelving Location.

g. Use the model to predict carseat unit sales when the price charged by competitor is average, community income levels is at its median, advertising is 15, population is 500, price for car seats at each site is 50, shelving location is good, average age of local population is 30, education level is 10, and the store is in an urban location within the US. What is your prediction for Y given these predictors? Construct an appropriate interval to quantify the uncertainty surrounding this prediction. Set $\alpha = 0.01$.

h. Use the model to predict carseat unit sales when the price charged by competitor is average, community income levels is at its median, advertising is 15, population is 500, price for car seats at each site is 50, shelving location is good, average age of local population is 30, education level is 10, and the store is in an urban location within the US. What is your estimate for $f(X)$ given these predictors? Construct an appropriate interval to quantify the uncertainty surrounding this estimation. Set $\alpha = 0.01$.

i Compare your results in (g) and (h). What do you observe? Explain why. Your explanation should include a discussion of reducible and irreducible error.