DS 303 Homework 2
Due: Sept. 13 2021 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: Multiple Testing Problem

Design and implement a simulation study to illustrate the multiple testing problem. Generate 1000 observations for 200 predictors $(X_1, X_2, \ldots, X_{200})$. Then generate 1000 Y observations such that Y has a relationship with only 5 of the 200 predictors. Explicitly:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \epsilon_i \quad (i = 1, \ldots, n), \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Decide on the values the parameters and report them (do not forget σ). Fit a multiple linear regression model on all 200 predictors and report the number of individual t-tests that are significant at $\alpha = 0.05$. Use this example to explain (in plain language, no statistics terminology), why we cannot depend on individual t-tests to tell us whether or not there is a relationship between at least one of the predictors and the response Y. Discuss the implications of the multiple testing problem on real applications outside the context of supervised learning. What tools are available to us to resolve this issue? Please make sure to submit your `R` code to receive full credit.

Problem 2: Review of regression concepts

Evaluate if the following statements are true or false and **justify your answer**.

a. When asked to state the true population regression model, a fellow student writes it as follows:

$$E(Y_i) = \beta_0 + \beta_1 x_i + \epsilon_i \quad (i = 1, \ldots, n).$$

b. The RSS (defined as $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$) must increase every time we add a predictor to the model.

c. For a given dataset, the training MSE will always be smaller than the test MSE.

d. The expected test MSE is defined as: $E(y_0 - \hat{f}(x_0))^2$. Here y_0 is from our training set and $\hat{f}()$ is the model we built from our training set. We evaluate $\hat{f}(x_0)$ on the x_0 values from our test set.

e. The bias-variance decomposition tells us that sometimes reducing the complexity of our model (for example, removing a predictor), can actually improve our expected test MSE.

f. When carrying out a hypothesis test, we (the user) set the type I error we're willing to accept.

Problem 3: Statistical Inference

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 (you'll need to find what the average competitor price is), median community income level, 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 (you'll need to find what the average competitor price is), median community income level, 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.

j Obtain the predicted carseat unit sales for all the same predictor values as in part (g), but set the price for car seats at each site to be 450. What is your prediction for Y? Does this make sense? Discuss how this reveals the limitations of our model.