DS 303 HOMEWORK 2
DUE: SEPT. 05, 2023 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: Properties of least square estimators via simulations

Simulations are a very powerful tool data scientists use to deepen our understanding of model behaviors and theory.

Suppose we know that the true underlying population regression model is :

$$Y_i = 2 + 3 \times X_{i1} + 5 \times \log(X_{i2}) + \epsilon_i \quad (i = 1, \ldots, n), \quad \epsilon_i \sim \mathcal{N}(0, 1^2).$$

a. What are the true values for β_0, β_1, and β_2?

b. Generate 100 Y_i observations from the true population regression model. You can use the following code to generate X_1 and X_2:

```
X1 = seq(0,10,length.out =100) #generates 100 equally spaced values from 0 to 10.
X2 = runif(100) #generates 100 uniform values.
```

c. Draw a scatterplot of X_1 and Y and a scatterplot of X_2 and Y. Describe what you observe.

d. Design a simple simulation to show that $\hat{\beta}_1$ is an unbiased estimator of β_1.

e. Plot a histogram of the sampling distribution of the $\hat{\beta}_1$'s you generated. Add a vertical line to the plot showing $\beta_1 = 3$.

f. Design a simple simulation to show that $\hat{\beta}_2$ is an unbiased estimator of β_2.

g. Plot a histogram of the sampling distribution of the $\hat{\beta}_2$'s you generated. Add a vertical line to the plot showing $\beta_2 = 5$.

h. Propose an unbiased estimator for $\text{Var}(\epsilon_i)$. Prove (using statistics and math) that your proposed estimator is unbiased. No code/simulations should be used here.

i. Use your answer from (h) to directly compute an unbiased estimator for $\text{Var}(\epsilon_i)$ for our simulated data. Report that numeric value here.

Problem 2: Review of regression concepts

Evaluate whether 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:

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

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

c. 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.

d. 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.

e. The expected test MSE can be smaller than the irreducible error.

f. The training MSE can be smaller than the irreducible error.

Answer the following questions.

g. Consider the patient dataset we went over in lecture. Suppose your colleague adds a new predictor to the dataset called 'unhappiness'. For each patient, this is the average of their disease severity score and anxiety score. Higher scores indicate more unhappiness. Your colleague now proposes fitting a linear regression model that looks like

```
lm(satisf ~ age + severe + anxiety + unhappy, data = patient).
```

Is this problematic? Explain why. Your explanation should include a clear definition on what it means for a matrix to be full rank.

h. How does the RSS (defined as $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$) behave each time we add a predictor to the model? Does it increase, decrease, stay the same, or not enough information? Explain in plain language the rationale behind its behavior.

Problem 3: Expected test MSE

For a real dataset, we cannot obtain the expected test MSE. It requires knowledge of the true model, irreducible error, and access to an infinite number of training sets. In simulations, we can get close to obtaining the expected test MSE and this is exactly what we'll do. Suppose we know that the true population regression line is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon.$$

a. Suppose $\beta_0 = \beta_1 = \beta_2 = 1$, and $\epsilon \sim N(0, 1)$. Generate $n = 100$ observations for Y_i under this model. You can use the following code to generate X_1:

```
X1 = seq(0,5,length.out =100)
```

Produce a plot of Y and X_1 and print that here.

b. Ideally, to compute the expected test MES we would have an infinite number of training sets. That's not computationally feasible, so instead let's just simulate 1000 training sets (each with $n = 100$). That means you'll need to simulate ($n = 100$) Y values 1000 times. There is no need to generate new X_1's (think about why). For each of these 1000 training sets, train 5 models of increasing complexity ($M_1 - M_5$). M_1 will be the linear regression model, M_2 includes a 2nd order-term, M_3 includes a 3rd order term, and so on until M_5. For each model, store the predicted value of Y when $X_1 = 1$. Report the first 5 predicted values for each model here.

c. Create a test set of 1000 observations: (x_0, y_0). For each test observation, let $x_0 = 1$. Generate y_0 using the true regression line with $x_0 = 1$. Report the first 5 values in your test set.

d. Use the results from above to obtain the *expected test MSE* for each of the five models when $x_0 = 1$. Report the five expected test MSEs here. Which model has the smallest expected test MSE?

e. Produce a plot with expected test MSE on the y-axis and model complexity (1-5) on the x-axis. Present that plot here.

f. Explain the behavior of your results in the context of the bias-variance tradeoff.