DS 303 Homework 1
Due: Sept. 8 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: Bias-variance decomposition

a. Provide a sketch of typical (squared) bias, variance, expected test MSE, training MSE, and the irreducible error curves on a single plot, as we go from less flexible statistical learning methods towards more flexible methods. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be 5 curves. Make sure to label each one.

b. Explain why each of the five curves has the shape displayed in part (a).

Problem 2: Multiple linear regression

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

```
install.packages("ISLR2") #you only need to do this one time.
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(Boston)`. To find out more about the data set, we can type `?Boston`.

We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

a. How many rows (n) are in the data set? How many variables are in the data set? What does the variable `lstat` represent?

b. Fit a simple linear regression model with `crim` as the response and `lstat` as the predictor. Describe your results. What are the estimated coefficients from this model? Report them here.
   Note: a simple linear regression is just a regression model with a single predictor.

c. Repeat this process for each predictor in the dataset. That means for each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

d. Fit a multiple regression model to predict the response using all of the predictors. You can do this from a single line of code:

```
lm(crim~.,data=Boston)
```

Summarize your results. For which predictors can we reject the null hypothesis: $H_0 : \beta_j = 0$?

e. How do your results from (c) compare to your results from (d)? Create a table (or a plot) comparing the simple linear regression coefficients from (c) to the multiple regression coefficients from (d). Describe what you observe. How does this provide evidence that using many simple linear regression models is not sufficient compared to a multiple linear regression model?

f. First `set.seed(1)` to ensure we all get the same values. Then, split the `Boston` data set into a training set and test set. On the training set, fit a multiple linear regression model to predict the response using all of the predictors. Report the training MSE and test MSE you obtain from this model.

g. On the training set you created in part (f), fit a multiple linear regression model to predict the response using only the predictors `zn, indux, nox, dis, rad, ptratio, medv`. Report the training MSE and test MSE you obtain from this model. How do they compare to your results in part (f)? Are these results surprising or what you expected?

Problem 3: 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.

Let's pretend we know that the true underlying population regression line is as follows (this is almost never the case in real life) :

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

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

b. Generate 100 observations Y_i under this normal error 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 $\beta_2 = 5$.