DS 301 Homework 1
Due: Feb. 2, 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: Bias-variance decomposition

a. On a single plot, provide a sketch of typical curves for (squared) bias, variance, expected test MSE, training MSE, and the irreducible error 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. Define in plain language (so that a non-data scientist can understand) what the quantities expected test MSE, training MSE, bias, variance and irreducible error mean.

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

d. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

Problem 2: Multiple linear regression

For this problem, we will use the `College` 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(College)`. To find out more about the data set, we can type `?College`.

a. How many observations are in the data set? How many variables?

b. Pull up the help page on the College data set (`?College`). Copy/paste the one-line description of the data set here.

c. Extract the 278th row of the data set. What college does it correspond to? Copy/paste that row here.

d. Obtain the average graduation rate across all colleges. How does Iowa State University compare?

e. Fit a simple linear regression model with the predictor as student-to-faculty-ratio and the response as graduation rate. Summarize the output from the model: the least square estimators, their standard errors, and corresponding p-values. **Do not just copy/paste raw `R` output here.** Interpret your results; what does the $\hat{\beta}_1$ tell us about the relationship between these two variables?

f. Draw the scatterplot of Y (graduation rate) versus X (student-to-faculty-ratio) and add the least squares line to the scatterplot.

g. Obtain the fitted values $\hat{y}_i$ and the residuals e_i from this simple linear regression model. Print the first 5 fitted values and the corresponding residuals.

h. Obtain the predicted graduation rate when the student-to-faculty ratio is 10. Print that value here.

i. This data is from 1995. Suppose we would like to use insights from this dataset to make predictions on graduation rates in 2022. Unfortunately, we do not have any data currently available for 2022. However, we can still get a sense of the prediction accuracy of our model on **data it has never seen before**. Explain in plain language how you will do this. Now implement it in `R` and obtain a realistic estimate of the prediction error for your trained model.

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) :

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

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

b. Generate 100 observations Y_i using the true population regression line. You may 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. Note: you should fit a multiple linear regression model here.

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. Note: you should fit a multiple linear regression model here.

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