DS 3010 Homework 2

Due: **FEB. 2, 2026** 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** on your HW.

To receive full credit, **show all necessary work**, including calculations. Please make your submission as **clear and readable** as possible. **Do not** include raw R output or R code in your written submission unless **explicitly requested or necessary for clarity**.

Option 1: R Markdown (single file)

- You may submit **one R Markdown file** (`.Rmd`) and its rendered output (`.html` or `.pdf`).
- If you use R Markdown, you **do NOT need to submit a separate code file**.

Option 2: Separate files

- Submit a **written file** (Word or PDF of picture of hand-written answer) containing your answers and calculations.
- Submit a **separate file** containing your R code (`.R`, `.txt`, or `.Rmd`).

Clearly label sections of the code using comments, for example:

```
##### Problem 1 #####
```

Grading Note

Extra credit: Questions marked with **(practice)** are included to support your learning and provide more flexibility. These questions are graded for **completion** rather than correctness. A substantive, good-faith attempt on all practice questions will earn **10% completion credit**, added to the homework total.

Problem 1: Concept Review

(a) We fit a linear regression model $Y \sim \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$ to a dataset and obtain our least square estimates: $\hat{\beta}_0,\ \hat{\beta}_1,\ \ldots,\ \hat{\beta}_p$. Suppose we change the units of all the predictors X_j, $j = 1, \ldots, p$, to obtain a new set of predictors $Z_j = cX_j$, $j = 1, \ldots, p$, where c is a constant. For example, you can think of our original predictors X_j as measured in dollars and our new predictors Z_j as measured in cents. Then, we fit the model $Y \sim \hat{\alpha}_0 + \hat{\alpha}_1 Z_1 + \cdots + \hat{\alpha}_p Z_p$ to the transformed dataset.

What is the relationship between the least squares coefficients $\hat{\alpha} = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_p)$ and $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$? State it explicitly.

(b) Based on your answer to part (a), if we have predictors that are on different scales (i.e. the predictors age and income are different scales: age goes from 0–100 while income can go from 0 to infinity), do we need to standardize our predictors so that they are on the same scale? In other words, do we need to divide age and income by their standard deviation so that they both have a range from 0 – 1. Explain whether or not this is necessary for our linear regression model.

(c) **(practice)** List the assumptions needed just to fit a least squares regression model (there should be four). Explain each assumption in plain language (i.e. use limited stats/math terminology).

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

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \qquad (i = 1, \ldots, n).$$

Is this correct? Justify your answer.

(e) True or False: For a given model, the training MSE must always be smaller than the test MSE. Justify your answer.

(f) We say that our least square estimates ($\hat{\beta}$) are unbiased. Explain what the term 'unbiased' means in plain language.

Problem 2: Multiple linear regression

For this problem, we will continue using 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`.

(a) 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.

(b) On the training set you created in part (a), 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 (a)?

(c) Did you expect the model in part (b) to have a smaller or larger training MSE compared to the model in part (a)? Explain.

(d) **(practice)** Did you expect the model in part (b) to have a smaller or larger test MSE compared to the model in part (a)? Explain.

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):

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

(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 with both X_1 and X_2.

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

(f) **(practice)** 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 with both X_1 and X_2.

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