DS 3010 Homework 2

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 in your solutions.** (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: Concept Review

(a) We fit a linear regression model $Y \sim \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_p X_p$ to a dataset and obtain our least square estimates: $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$. Supposes 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 $Y \sim \hat{\alpha}_0 + \hat{\alpha}_1 Z_1 + \ldots + \hat{\alpha}_p Z_p$ 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.

The relationship is $\beta_j = c * \alpha_j$. Students could have argued this mathematically or via simulations.

```
c = 100
x = runif(100)
z = c*x
beta0 = 1
beta1 = 2
error = rnorm(100)
y = beta0 + beta1*x + error

summary(lm(y~x))
summary(lm(y~z))
```

(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.

We do not need to scale the predictors. Each coefficient will adjust for the scale of each predictor. In fact, it's easier for interpretation if we do not scale. That way, age and income get to stay on their original units (i.e. years and dollars).

(c) 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).

We need the following assumptions to fit a linear regression:

(a) approximately linear relationship between response and predictor,

(b) $E(\epsilon_i) = 0$ for i = 1, 2, ..., n, - this means that, on average, we expect our random error to be centered around 0.

(c) constant variance ($Var(\epsilon_i) = \sigma^2$) - each observation comes from a distribution with the same variability

(d) uncorrelated ϵ_i - the observations do not depend on each other.

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

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

Is this correct? Justify your answer.

This is incorrect. The true population regression line involves random error since we assume that there is inherent noise/randomness in the population: $Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \epsilon_i$.

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

Since we randomly split the data into training and test sets, it is possible that the training MSE is larger than the test MSE. Nothing guarantees that the training MSE must be smaller than the test MSE, although in general that is what we expect.

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

Suppose we had an infinite number of datasets, and for each of these datasets we can fit a model and obtain least square estimates. Then the average of these (infinite) number of least square estimates would exactly equal the true population parameters.

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.

Training MSE should be 42.49 and test MSE should be 41.19.
Note students should use `crim` *as their response here. If they used a different response variable but their code/logic is correct, please just deduct one point.*

(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)?

Training MSE should be 43.97 and test MSE should be 39.62. Students should point out that this test MSE is smaller than the one they achieved in part (k).

(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.

Students should provide some reasonable intuition here. Since the full model in part (a) involves more predictors, it should have a smaller training MSE compared to the model in part (b). This is because with more predictors, it is a more complex model and can achieve a smaller RSS. Therefore, we would expect it to have a better fit on the training set.

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

Despite the fact that the model in part (b) is more complex, that does not guarantee that it will perform better on the test set. In fact, if we only utilize predictors that are relevant in predicting crime, we can achieve a smaller test MSE. Therefore we would expect the model in part (b) to have a smaller test MSE compared to the model in part (a).

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?

True values are $\beta_0 = 2$, $\beta_1 = 3$, and $\beta_2 = 5$.

(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.
```

No output necessary. Students' code should look something like:

```
 n = 100
beta_0 = 2
beta_1 = 3
beta_2 = 5
X1 = seq(0,10,length.out =100)
X2 = runif(100)

error = rnorm(n,0,1)
Y = beta_0 + beta_1*X1 + beta_2*log(X2) + error
```

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

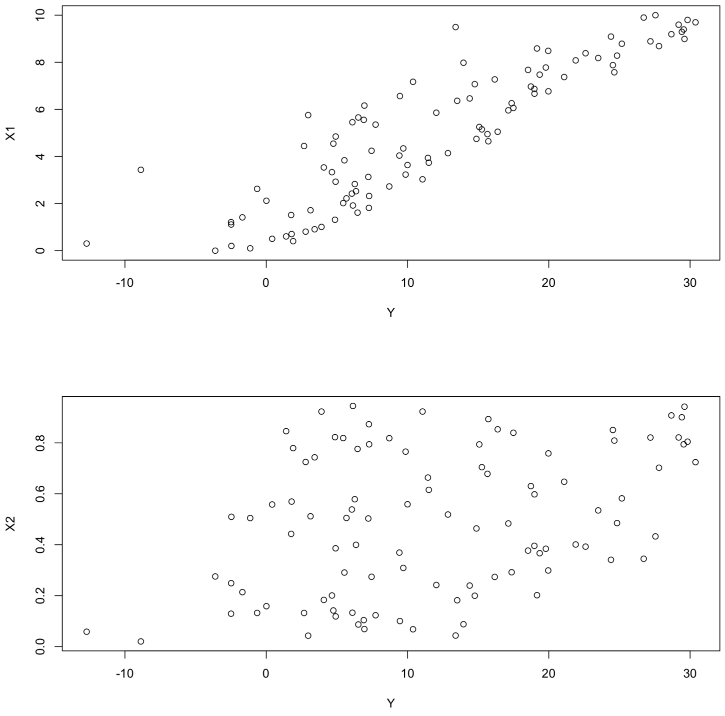

Figure 1: Scatterplot of relationship between X_1 versus Y and X_2 versus Y.

From Figure **??**, *we can observe a linear relationship between Y and X_1 and a non-linear (or weaker linear) relationship between Y and X_2.*

```
 B = 5000
beta1hat = beta2hat = rep(NA,B)

for(i in 1:B){
  error = rnorm(n,0,1)
  Y = beta_0 + beta_1*X1 + beta_2*log(X2) + error
  fit = lm(Y˜X1+log(X2))
  beta1hat[i] = fit$coefficients[[2]]
  beta2hat[i] = fit$coefficients[[3]]
}

mean(beta1hat)
mean(beta2hat)

hist(beta1hat, main="Histogram of beta1_hat")
abline(v=3,col=’red’)

hist(beta2hat, main="Histogram of beta2_hat")
abline(v=5,col=’red’)
```

(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.

```
 B = 5000
beta1hat = beta2hat = rep(NA,B)

for(i in 1:B){
  error = rnorm(n,0,1)
  Y = beta_0 + beta_1*X1 + beta_2*log(X2) + error
  fit = lm(Y˜X1+log(X2))
  beta1hat[i] = fit$coefficients[[2]]
  beta2hat[i] = fit$coefficients[[3]]
}

mean(beta1hat)
mean(beta2hat)

hist(beta1hat, main="Histogram of beta1_hat")
abline(v=3,col=’red’)

hist(beta2hat, main="Histogram of beta2_hat")
abline(v=5,col=’red’)
```

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

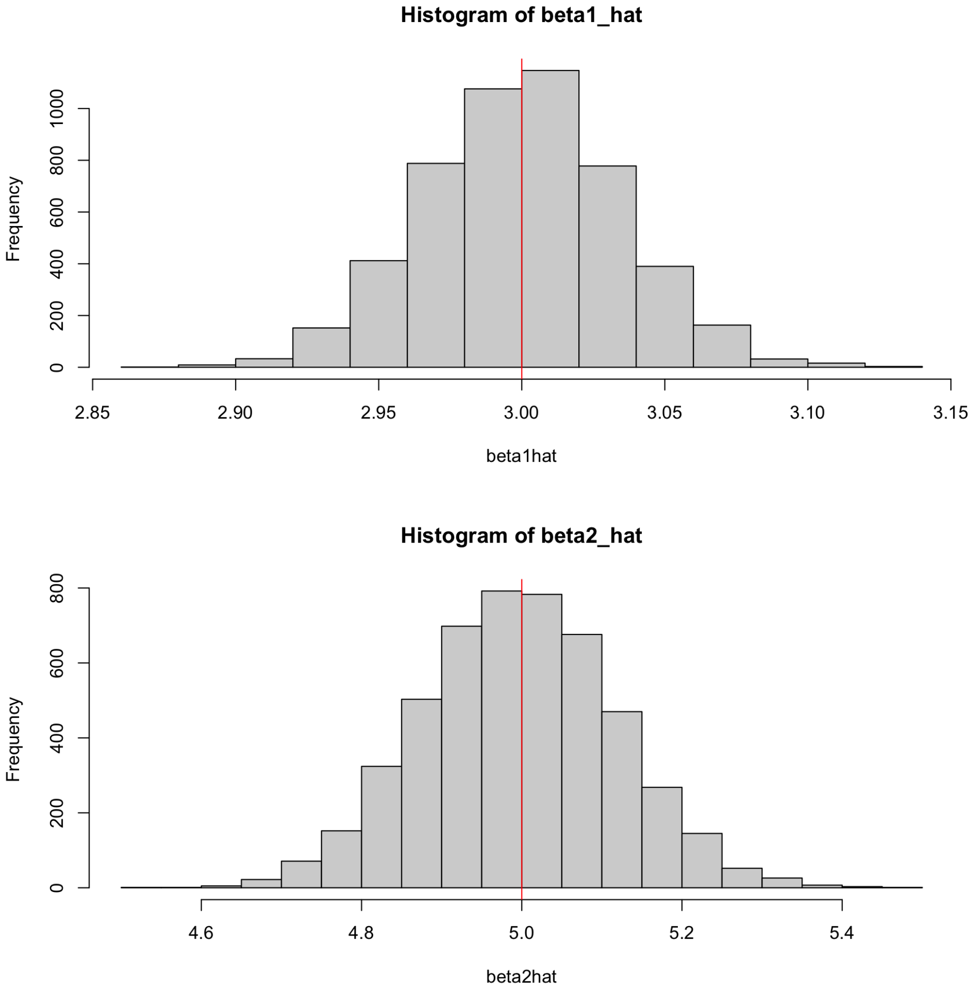

Figure 2: Histogram of the sampling distribution of $\hat{\beta}_1$ and $\hat{\beta}_2$.

(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 with both X_1 and X_2.

See code above.

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

See Figure **??**.

End of assignment.