DS 301 Homework 2
Due: Feb. 9, 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: 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 half the `Boston` data set into a training set and the remaining half into the 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?

h. Fit the following multiple linear regression model:

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

Use this model to answer the following question: A consultant thinks that if all predictors are held equal, a tract on the Charles River (`chas`) has an effect on median home values (`medv`). Is there evidence to reject the consultant's claim? Carry out a hypothesis test (using the above linear regression model) to answer this. Be sure to report your null/alternative hypothesis, test statistic, null distribution, p-value, and conclusion. What additional assumption do we need to make to carry out this hypothesis test?

i. Same setup as part (h), but now the consultant claims a tract on the Charles River increases median home values by \$5,000. Note that `medv` is in \$1,000's of dollars. Carry out a hypothesis test to test their claim. Be sure to report your null/alternative hypothesis, test statistic, null distribution, p-value, and conclusion.

Problem 2: Concept Review

a. List the assumptions needed just to fit a least squares regression model (there should be four). In general, is there an assumption you believe to be particularly problematic?

b. When asked to state the true population regression model, a fellow student writes it as follows:

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

Is this correct? Justify your answer.

c. Your classmate wants to test whether a regression coefficient is equal to zero or not at $\alpha = 0.05$. They set up the null and alternative as follows:

$$H_0 : \hat{\beta}_j = 0 \text{ versus } H_1 : \hat{\beta}_j \neq 0.$$

Is this correct? Justify your answer.

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

Problem 3: Multiple Testing Problem

Design and implement a simulation study to illustrate the multiple testing problem. Generate 1000 observations for 200 predictors $(X_1, X_2, \ldots, X_{200})$. Then generate 1000 Y observations such that Y has a relationship with only 5 of the 200 predictors. Explicitly:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \epsilon_i \quad (i = 1, \ldots, n), \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Decide on the values the parameters and report them (do not forget to report σ). Fit a multiple linear regression model on all 200 predictors and report the number of individual t-tests that are significant at $\alpha = 0.05$. Use this example to explain (in plain language, no statistics terminology), why we cannot depend on individual t-tests to tell us whether or not there is a relationship between at least one of the predictors and the response Y. Discuss the implications of the multiple testing problem on real applications outside the context of supervised learning. What tools are available to us to resolve this issue? Please make sure to submit your `R` code to receive full credit.