DS 303 Homework 3
Due: Sept. 20, 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: Wrapping up Multiple Linear Regression

Suppose you are going to teach a friend the fundamentals of multiple linear regression. Summarize the following topics in a way that your friend, who has a minimal statistics/CS background, could understand. Each summary should be no more than 3-5 sentences.

a. What is a linear regression model attempting to estimate? Does it it require any assumptions?

b. Conceptually, how do we obtain the estimates for a multiple linear regression model?

c. Are these estimates trustworthy? How do we know? Explain any key concepts in plain language.

d. How do we obtain a prediction for Y based on our model? How do we quantify any uncertainty related to our predictions?

e. How can we evaluate how good our model is at prediction?

f. What is statistical inference and why is it useful in the context of linear regression models?

g. What are 3 potential issues that may arise with our multiple linear regression model? For each of these issues, explain 1. why the issue can cause problems and 2. what can be done to resolve the issue.

Problem 2: Interaction terms

We will use the `Credit` dataset for this problem. It is part of the library `ISLR2`.

a. This data set contains a few categorical predictors. As we already discussed in lecture, these predictors should be stored as `factors` so that `R` can handle them properly. Using the `str` function, check that all the qualitative predictors in our dataset are stored correctly in `R` as factors. Copy and paste your output.

b. Fit a model with the response (Y) as credit card balance and X_1= `Income` and X_2 = `Student` as the predictors. Call this model `fit`. Summarize your output.

c. Based on our results from part (b), write out the fitted model for students and write out the fitted model for non-students.

d. Interpret the regression coefficient related to `Income` for both models.

e. Notice that our model says that regardless of student status, the effect of `Income` on average `Balance` is the same. Do you think this is a reasonable constraint of our model? Construct some plots to back up your answer.

f. One way we could relax this assumption is by incorporating *interaction terms* into our model. Specifically:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_i 3 + \epsilon_i,$$

where X_1= `Income`, X_2 = `Student`, and X_3 = `Income` × `Student`. Fit a model with an interaction term using the following code:

```
lm(Balance ~ Income + Student + Income:Student, data=Credit)
```

Based on this model, write out the fitted model for students and write out the fitted model for non-students.

g. Interpret the regression coefficient related to `Income` for the fitted models obtained in part (f).

h. The model from part (f) has a significant F-test statistic, which tells us the overall model is jointly significant and at least one of the regression coefficients is significantly different from zero. However, the R^2 is quite low. Are these results contradictory? Explain.

Problem 3: A Puzzling Problem

When fitting a linear regression model on a data set, you encounter the following `R` output. You notice there is something strange about the results. Point out what is strange in this output and **explain clearly** how this could happen.

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3700 -1.6364 -0.1208  1.4261  5.2558

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2936     0.5217   4.396 2.82e-05 ***
x1            1.2600     2.3006   0.548    0.585
x2            1.8968     2.5509   0.744    0.459
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.376 on 97 degrees of freedom
Multiple R-squared:  0.09896,Adjusted R-squared:  0.08038
F-statistic: 5.326 on 2 and 97 DF,  p-value: 0.006385
```

Problem 4: Predictions in the presence of multicollinearity

a. Is multicollinearity a problem for making accurate predictions? If you're unsure, make an educated guess based on what we have learned in class.

b. Let's carry out a simulation study to answer this. We will simulate data with and without multicollinearity. This is our true model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

where $i = 1, \ldots, 100$, $\epsilon \sim N(\mu = 0, \sigma^2 = 4)$, $\beta_0 = 3$, $\beta_1 = 2$, and $\beta_2 = 4$.
To generate your predictors **with** multicollinearity use the following code:

```
set.seed(42)
x1 = runif(100)
x2 = 0.8*x1 + rnorm(100,0,0.1)
```

Using these predictors, generate your `Y` values in `R`. Check the correlation between `x1` and `x2` using the `cor()` function. Report that value here.

c. Split your data into a training set and test set. Train your model on the training set.

```
lm(Y ~ x1+x2, data = train)
```

Report the test MSE for this model.

d. Repeat this process 2,500 times (use a for-loop). This means for each iteration, you'll need to generate a random set of Y's, fit a model on your training set, and then obtain the test MSE for that model. We do not need to generate new predictor values (think about why). **Remember to store the test MSE for each iteration and do not set seed**. What is the mean test MSE in this setting when the predictors are highly correlated? Plot a histogram of your 2,500 test MSEs and comment on what you see.

e. Now generate predictors **without** multicollinearity using the following code:

```
set.seed(24)
x1 = runif(100)
x2 = rnorm(100,0,1)
```

Using these predictors, generate your `Y` values in `R`. Check the correlation between `x1` and `x2`. Report that value here.

f. Again run 2,500 simulations to obtain the test MSE of our model when the predictors are not correlated. What is the mean test MSE in this setting when the predictors are not correlated? Plot a histogram of your test MSE and comment on what you see.

g. Based on our simulation study, is multicollinearity a problem for making accurate predictions? Comment on your findings.

Problem 5: Model Diagnostics

We will use the `Auto` dataset for this problem. It is part of the library `ISLR2`. We will treat `mpg` as the response and all other variables except `name` as the predictors.

a. Produce a scatterplot matrix which includes all of the variables in the data set. Hint: use the `plot()` function. Describe any relationships you observe. For which variables, if any, is there evidence of a non-linear relationship with the response?

b. Perform a multiple linear regression with `mpg` as the response and all other variables except `name` as the predictors. Summarize the output.

c. Is there a significant relationship between at least one of the predictors and the response? Justify your answer.

d. What does the coefficient for the `year` predictor suggest?

e. Is multicollinearity an issue in our model? Justify your answer. You do not need to implement a solution.

f. Check that the constant variance and linearity assumption hold for the model. Comment on any problems you see.

g. Propose and implement some transformations on the variables. This should include fitting some of the non-linear models we've covered. Report your final model. Is this an improvement over the model from part (b)? Justify your answer.