DS 301 Homework 4
Due: Feb. 23, 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: Insurance data

Download the `insurance.csv` file from Module 05. Then load your data into `R` using the following code:

```
insurance=read.csv("/.../insurance.csv")
```

where the `"...."` needs to be the pathway you've saved your file in.

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. Check that all the qualitative predictors in our dataset are stored correctly in `R` as factors using the `str()` function. If they are not, convert them to factors. Copy and paste your output here.

b. Fit a model with the response (Y) as health care charges and x_1= `age`, x_2 = `bmi`, x_3 = `gender` as your predictors. Call this model `fit`. Summarize your output here. Discuss any insights you can obtain from this model and its related output.

c. Based on our results from part (b), write out the fitted model for males only (when `gendermale = 1`). Then write out the fitted model for females only (when `gendermale = 0`).

d. Your classmate tells you that including `gender` as a dummy variable in the model is not necessary. Instead you can just fit a model for only those observations that are males and only those observations that are female. To see whether or not your classmate's approach makes sense, subset your data into two groups: data for males and data for females. To obtain data for males only, you can use the following code:

   ```
   males=insurance[insurance$gender==‘male’,]
   ```

   You can modify this code directly to obtain data for females only. Fit a model with `bmi` and `age` for the male group only. Call this model `fit_males`. Now do the same for the female group. Call this model `fit_females`. Write out your two models here with the estimated regression coefficients.

e. Compare your results in part (d) to part (c). Is the model you obtained for males only in part (c) the same as `fit_males`? What about for females? Explain in plain language to your classmate why these two approaches will not give the same results.

f. The model from part (b) 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 2: 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 (otherwise you will get the same values for each iteration)**. What is the mean of the 2,500 test MSEs 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 of the 2,500 test MSEs in this setting when the predictors are not correlated? Plot a histogram of your test MSEs 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 3: 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 and see if they improve your model. Report your final model.

Problem 4: Wrapping up Multiple Linear Regression

Let's review the fundamentals of multiple linear regression.

a. Write out the population multiple linear regression model.

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

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

d. What is our estimate for $E(Y)$ for specific values of X? Clearly define any quantities. How do we quantify any uncertainty about our estimate for $E(Y)$?

e. What is our prediction for Y for specific values of X? Clearly define any quantities. How do we quantify any uncertainty about our prediction?

f. How can we evaluate how good our model is at prediction? Explain what the bias-variance tradeoff tells us about model behavior.

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

h. 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.