Yeabin Moon, Ph.D.
BUS 212A-2 Spring 2023
January 5, 2023

Homework 2

Submission Instructions

1) You have to use Juppyter Notebook
2) Click the Save button at the top of the Jupyter Notebook.
3) Select Cell → All Output → Clear. This will clear all the outputs from all cells (but will keep the content of all cells).
4) Select Cell → Run All. This will run all the cells in order, and will take several minutes.
5) Once you've rerun everything, select File → Download as → PDF via LaTeX (If you have trouble using "PDF via LaTex", you can also save the webpage as pdf. Make sure all your solutions especially the coding parts are displayed in the pdf, it's okay if the provided codes get cut off because lines are not wrapped in code cells).
6) Look at the PDF file and make sure all your solutions are there, displayed correctly. The PDF is the only thing your graders will see!
7) Submit your PDF on Latte.

Question 1. Suppose we have a data set with five predictors,

- $X_1 = GPA$
- $X_2 = IQ$
- $X_3 =$ Level (1 for College and 0 for High School)
- $X_4 =$ Interaction between GPA and IQ
- $X_5 =$ Interaction between GPA and Level

The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50, \hat{\beta}_1 = 20, \hat{\beta}_2 = 0.07, \hat{\beta}_3 = 35, \hat{\beta}_4 = 0.01, \hat{\beta}_5 = -10$.

1) Which answer is correct, and why?
   (a) For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.
   (b) For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.
   (c) For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.
   (d) For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that the GPA is high enough.
2) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.
3) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

Question 2. I collect a set of data ($n = 100$ observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$

1) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.
2) Answer again using test rather than training RSS.

3) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.
4) Answer again using test rather than training RSS.

Question 3. Use *Auto* data set to answer the following questions

1) Explore data. That entails detailing the exploration of the data, the features, the feature types, and any possible missing data.
   - How would you identify the missing values?
   - How would you deal with them? Supply your intuition.
   - Apply your strategy to the data.
2) Perform a simple linear regression with **mpg** as the response and **horsepower** as the predictor.
   - Is there a relationship between the predictor and the response?
   - How strong is the relationship between the predictor and the response?
   - Is the relationship between the predictor and the response positive or negative?
   - What is the predicted mpg associated with a horsepower of 98? What are the associated 95 % confidence and prediction intervals?
3) Plot the response and the predictor to display the least squares regression line
4) Produce some diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

Question 4. Use *Auto* data set again to answer the following questions.

1) Produce a scatterplot matrix which includes all of the variables in the data set.
2) Compute the matrix of correlations between the numerical variables.
3) Perform a multiple linear regression with **mpg** as the response and all other variables except **name** as the predictors.
   - Is there a relationship between the predictors and the response?
   - Which predictors appear to have a statistically significant relationship to the response?
   - What does the coefficient for the **year** variable suggest?
4) Produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?
5) Perform linear regression models with interaction effects. Do any interactions appear to be statistically significant?

Question 5. Which linear regression training algorithm can you use if you have a training set with millions of features?