STAT 4533/5533 - Homework 3 - Due February 18, 2024

Instructions: Everyone must answer Questions 1 - 3. Question 4 is optional and will not be graded. Solutions must be created using R Markdown, and you will upload your HTML file to D2L to submit your answers. Your markdown document must be logically organized and easy to read. Do not print entire data sets in your output.

1. Chapter 5 Exercise 8

2. This question uses the *newCredit.csv* data set on D2L. The data contains finanicial information for 300 individuals including credit limit, credit rating, number of credit cards, age, credit card balance, and a categorical indicator variable for high income. You will use this data to calculate the LOOCV error for predicting the binary high income variable without using the `cv.glm()` function. Use the full data set without training and test splits.

   (a) Fit a logistic regression model to predict the income group using all but the first observation. Include all of the independent variables in the model.

   (b) Use the model from part (a) to predict the income group for the first observation.

   (c) Write a for loop from $i = 1$ to n that performs the following steps:

      i. Fit a logistic regression model using all but the ith observation to predict income using the other variables.

      ii. Compute the posterior probability of the ith observation belonging to the high income group.

      iii. Use the probability from part ii to predict whether the ith observation is in the high income group.

      iv. Determine whether or not an error was made in the prediction for the ith observation. Save this result to access outside the loop.

   (d) Use the information saved in part (c) iv to determine the LOOCV estimate of the test error.

3. Two multiple regression models are fit using 5-fold cross-validation. The resulting MSE on each cross-validation set for each model is shown in the table below. Which model would you choose? Explain your choice.

Model 1		Model 2	
CV Set	MSE	CV Set	MSE
1	33,415	1	26,666
2	38,741	2	38,554
3	32,112	3	39,662
4	37,210	4	36,756
5	29,501	5	30,303

The following question is recommended, but will not be graded:

4. Chapter 5 Exercise 7