STAT 4533/5533 - Homework 4 - Due March 10, 2024

Instructions: Everyone must answer Questions 1 and 2. Graduate students are also required to answer Question 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 6 Exercise 8 excluding parts (c) and (d)

2. The *diabetes.csv* file on D2L contains concentrations of 131 metabolites measured from a sample of 198 patients with diabetes. One of the metabolites, creatinine, is an indicator of kidney function. This problem will investigate the problem of predicting creatinine using the values of the other 130 metabolites.

   (a) Divide the data set into a training and test set. Your training set should consist of the first 150 observations while the test set is made up of the remaining observations.

   (b) Using the training data, fit a ridge regression model to predict creatinine using the other metabolites. Choose the value of lambda using cross-validation. What value of lambda is chosen?

   (c) Using the training data, fit a lasso regression model to predict creatinine using the other metabolites. Choose the value of lambda using cross-validation. What value of lambda is chosen?

   (d) How many of the regression coefficients from the lasso model in part (c) shink to zero?

   (e) Using the training data, fit a PCR model to predict creatinine using the other metabolites. Use cross-validation to choose the number of principal components. Explain your choice (there might be more than one correct answer here).

   (f) Calculate the MSE on the test data for each of the three models.

   (g) Which model performed the best? Explain your answer.

3. GRADUATE STUDENTS ONLY: Chapter 6 Exercise 2

The following question is recommended but will not be graded:

4. Chapter 6 Exercise 9