DS 301: Homework 6
Due: March 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` code or raw `R` output** (unless it is asked for specifically or needed for clarity).

Code should be submitted with your homework as a separate file (for example, an R file, text 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: Final Project Team Members

Please read carefully the final project instructions and form groups of 4 for your final project. Please report the following information for your team:

a. Team member names/student ID.

b. Team name.

c. The dataset you plan to analyze. Please include a link to the data. If it is your own data, you can upload the data to Google Drive/Box/Dropbox and send me link to a shared folder.

d. Anticipated responsibilities of each team member (in other words, how do you plan to divide the work?).

Problem 2: Concept Review

a. Subset selection will produce a collection of $p+1$ models $M_0, M_1, M_2, \ldots, M_p$. These represent the 'best' model of each size (where 'best' here is defined as the model with the smallest RSS). Is it true that the model identified as M_{k+1} must contain a subset of the predictors found in M_k? In other words, is it true that if $M_1 : Y \sim X_1$, then M_2 must also contain X_1. And if M_2 contains X_1 and X_2, then M_3 must also contain X_1 and X_2? Explain your answer.

b. Same question as part (a) but instead of subset selection, we now carry out forward selection.

c. Suppose we perform subset, forward, and backward selection on a single data set. For each approach, again we can obtain $p + 1$ models containing $0, 1, 2, \ldots, p$ predictors.
As we know, best subset will give us a best model with k predictors. Call this $M_{k,subset}$. Forward selection will give us a best model with k predictors. Call this $M_{k,forward}$. Backward selection will give us a best model with k predictors. Call this $M_{k,backward}$. Which of these three models would we expect to have the smallest training MSE? Explain your answer. Hint: Consider the case for $k = 0$ and $k = p$ first. Then the case for $k = 1$. Then the case for $k = 2, \ldots, p - 1$.

d. Same setup as part (c). Which of these three models would we expect to have the smallest test MSE? Explain your answer.

Problem 3: Model Selection

We will use the `College` data set in the `ISLR2` library to predict the number of applications (`Apps` each university received. Randomly split the data set so that 90% of the data belong to the training set and the remaining 10% belong to the test set.

a. Implement forward and backward model selection. Did you implement this on the full dataset or on your training set only? Explain your reasoning.

b. For both forward and backward selection, report the best model based on AIC and BIC. How do these models compare?

c. Implement best subset selection. Did you implement this on the full dataset or on your training set only? Explain your reasoning. Report the best model you obtained using AIC and BIC. How do these results compare with part (b)?

d. Implement forward selection and report the model with the smallest test MSE. Did you implement this on the full dataset or on your training set only? Explain your reasoning. Report your final model.

e. Repeat (d), but using a different random split to the dataset. Report the model with smallest test MSE. Is this the same as the model you obtained in (d)? Discuss how this reveals one advantage of using forward selection with AIC (or BIC) as our criteria.