fProblem 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.

Keven Lin 851529703

Adam Banwell 375725362

Caleb Purcell 603399821

Brian Sayre 807690904

b. Team name.

Anti - Heart Disease

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.

https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

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

We will all share responsibility in data cleaning, calculation, and conclusion.

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.

Yes, because for each addition to the M, it will also contain the previous Xi plus the most recent Xi.

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

Yes, because for forward selection, any predictors that are added during each iteration will stay in the model. So as long as the iteration does not stop, then a new predictor will be added.

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 $k = 2, \ldots, p - 1$.

K = 0: Backward selection as it contains all the predictors. MSE is biased to more predictors due to bias variance trade off.

K = 1: Backward selection as it contains all the predictors. MSE is biased to more predictors due to bias variance trade off.

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

The smallest training RSS will be for the model with best subset approach. This is because the model will be chosen after considering all the possible models with k parameters for best subset. This is not true for either backward stepwise or forward stepwise.

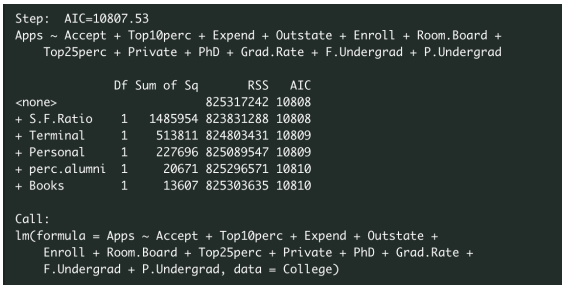
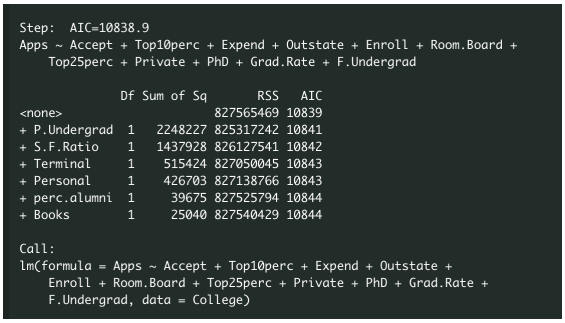
Problem 3: Model Selection

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.

I implemented the full data set because the prof told us in ED.

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

Forward Selection with AIC	Forward Selection with with BIC!!! (picture said bic but it is ran with K= log(n) to produce BIC)
	

```
Step:  AIC=10807.53
Apps ~ Accept + Top10perc + Expend + Outstate + Enroll + Room.Board +
    Top25perc + Private + PhD + Grad.Rate + F.Undergrad + P.Undergrad

               Df Sum of Sq       RSS   AIC
<none>                     825317242 10808
+ S.F.Ratio    1   1485954 823831288 10808
+ Terminal     1    513811 824803431 10809
+ Personal     1    227696 825089547 10809
+ perc.alumni  1     20671 825296571 10810
+ Books        1     13607 825303635 10810

Call:
lm(formula = Apps ~ Accept + Top10perc + Expend + Outstate +
    Enroll + Room.Board + Top25perc + Private + PhD + Grad.Rate +
    F.Undergrad + P.Undergrad, data = College)
```

```
Step:  AIC=10838.9
Apps ~ Accept + Top10perc + Expend + Outstate + Enroll + Room.Board +
    Top25perc + Private + PhD + Grad.Rate + F.Undergrad

               Df Sum of Sq       RSS   AIC
<none>                     827565469 10839
+ P.Undergrad  1   2248227 825317242 10841
+ S.F.Ratio    1   1437928 826127541 10842
+ Terminal     1    515424 827050045 10843
+ Personal     1    426703 827138766 10843
+ perc.alumni  1     39675 827525794 10844
+ Books        1     25040 827540429 10844

Call:
lm(formula = Apps ~ Accept + Top10perc + Expend + Outstate +
    Enroll + Room.Board + Top25perc + Private + PhD + Grad.Rate +
    F.Undergrad, data = College)
```

Backward Selection with AIC	Backward Selection with BIC!!! (picture said bic but it is ran with K= log(n) to produce BIC)

```
Step:  AIC=10807.53
Apps ~ Private + Accept + Enroll + Top10perc + Top25perc + F.Undergrad +
    P.Undergrad + Outstate + Room.Board + PhD + Expend + Grad.Rate

               Df  Sum of Sq        RSS   AIC
<none>                       825317242 10808
- P.Undergrad  1    2248227  827565469 10808
- F.Undergrad  1    3629713  828946955 10809
- Grad.Rate    1    9850583  835167825 10815
- Room.Board   1   10699017  836016260 10816
- Top25perc    1   12037817  837355059 10817
- PhD          1   12708568  838025810 10817
- Private      1   15691081  841008323 10820
- Enroll       1   24676722  849993965 10828
- Outstate     1   26201946  851519188 10830
- Expend       1   43734225  869051468 10846
- Top10perc    1   89928332  915245574 10886
- Accept       1 1696846612 2522163854 11674

Call:
lm(formula = Apps ~ Private + Accept + Enroll + Top10perc + Top25perc +
    F.Undergrad + P.Undergrad + Outstate + Room.Board + PhD +
    Expend + Grad.Rate, data = College)
```

```
Step:  AIC=10838.9
Apps ~ Private + Accept + Enroll + Top10perc + Top25perc + F.Undergrad +
    Outstate + Room.Board + PhD + Expend + Grad.Rate

               Df  Sum of Sq        RSS   AIC
<none>                       827565469 10839
- F.Undergrad  1    5704713  833270183 10840
- Grad.Rate    1    8648712  836214181 10842
- PhD          1   11976579  839542048 10846
- Room.Board   1   11983192  839548662 10846
- Top25perc    1   11989884  839555353 10846
- Private      1   16320275  843885744 10850
- Enroll       1   25831876  853397345 10858
- Outstate     1   26496767  854062236 10859
- Expend       1   44598808  872164277 10875
- Top10perc    1   88092182  915657651 10913
- Accept       1 1695534802 2523100271 11700

Call:
lm(formula = Apps ~ Private + Accept + Enroll + Top10perc + Top25perc +
    F.Undergrad + Outstate + Room.Board + PhD + Expend + Grad.Rate,
    data = College)
```

We can see that AIC and BIC runs for both Forward and Backwards selection predictor picks are nearly identical.

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 the 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.