**BU.510.650 Homework #3**

**Data Analytics Page 1 of 3**

**Spring 2021 The Johns Hopkins University**

**Arnab Bisi Carey Business School**

**Homework #3**

Due: 02/21/21, 11:59pm

1. (a) Download the Hitters data set from R library “ISLR” and run the first part of the R codes provided in the file “Hitters-Subset-Selection-withAuto.R”, and submit your output. (1 point)

[Hitters-Subset-Selection-withAuto.docx](file:///C:\Users\duter\AppData\Roaming\Microsoft\Word\Hitters-Subset-Selection-withAuto.docx)

(b) For the Hitters data set, using “Adjusted R-square” as the model selection criterion, which models would you select using Best Subset and Forward Stepwise selection methods?

As shown below, both Best Subset and Forward Stepwise selection methods yielded the same best model: (Salary ~ AtBat + Hitsm + Walks + CAtBat + CRuns + CRBI + CWalks + LeagueN + DivisionW + PutOuts + Assists)

regfit.bss <- **regsubsets**(Salary **~** .,  
 data = Hitters.data,  
 nvmax = **ncol**(Hitters.data) **-** 1)  
  
regfit.fwd <- **regsubsets**(Salary **~** .,  
 data = Hitters.data,  
 nvmax = **ncol**(Hitters.data) **-** 1,  
 method = 'forward')  
  
**coef**(regfit.bss, **which.max**(**summary**(regfit.bss)**$**adjr2))

## (Intercept) AtBat Hits Walks CAtBat CRuns   
## 135.7512195 -2.1277482 6.9236994 5.6202755 -0.1389914 1.4553310   
## CRBI CWalks LeagueN DivisionW PutOuts Assists   
## 0.7852528 -0.8228559 43.1116152 -111.1460252 0.2894087 0.2688277

**coef**(regfit.fwd, **which.max**(**summary**(regfit.fwd)**$**adjr2))

## (Intercept) AtBat Hits Walks CAtBat CRuns   
## 135.7512195 -2.1277482 6.9236994 5.6202755 -0.1389914 1.4553310   
## CRBI CWalks LeagueN DivisionW PutOuts Assists   
## 0.7852528 -0.8228559 43.1116152 -111.1460252 0.2894087 0.2688277

Similarly, using “BIC” as the model selection criterion, which models would you select using Best Subset and Backward Stepwise selection methods? (2 points)

Best Subset: (Salary ~ CRBI)

Backward Stepwise: (Salary ~ CRuns)

regfit.bwd <- **regsubsets**(Salary **~** .,  
 data = Hitters.data,  
 nvmax = **ncol**(Hitters.data) **-** 1,  
 method = 'backward')  
  
**coef**(regfit.bss, **which.max**(**summary**(regfit.bss)**$**bic))

## (Intercept) CRBI   
## 274.5803864 0.7909536

**coef**(regfit.bwd, **which.max**(**summary**(regfit.bwd)**$**bic))

## (Intercept) CRuns   
## 259.0822757 0.7664116

1. Suppose we have a data set, which has *p* predictors (input variables), and we perform model selection using (i) best subset selection (BSS), (ii) forward stepwise selection (FwSS), and (iii) backward stepwise selection (BwSS). Specifically, using each of these three approaches, we determine the best model with *k* predictors for all possible values of *k*, that is, *k* = 1, 2, …, *p*.
2. For a given *k*, suppose we are comparing the models obtained by these three methods, that is, the model with k predictors obtained by BSS, FwSS, BwSS. Which of the three models will have the smallest RSS on the data we used to perform model selection? Explain your answer. (1 point)

Since the BSS method considers more possible predictor combinations than FwSS and BwSS, it is more likely to yield a lower RSS.

1. Answer the following True or False questions:
2. The predictors in the k-predictor model identified by FwSS are a subset of predictors in the (k+1)-predictor model identified by FwSS. ***(True)***
3. The predictors in the k-predictor model identified by BwSS are a subset of predictors in the (k+1)-predictor model identified by BwSS. ***(True)***
4. The predictors in the k-predictor model identified by BSS are a subset of predictors in the (k+1)-predictor model identified by BSS. ***(False)***
5. The predictors in the k-predictor model identified by BwSS are a subset of predictors in the (k+1)-predictor model identified by FwSS. ***(False)***

(2 points)

1. In this question, you will perform model selection using AutoLoss data set (adapted from a data set of loss payments made by insurance companies – see the following url for details: <https://archive.ics.uci.edu/ml/datasets/Automobile>). An explanation of the variables in this data set is included in the Appendix to this assignment. Our goal is to predict the loss payment for a vehicle (the payments made by an insurance company to cover claims) as a function of vehicle characteristics.
2. Using the best subset selection method and allowing up to 15 predictors, use regsubsets() to determine the best model with k predictors for k = 1, 2, …, 15. Use the output to answer the following question: Which predictors are included in the best model with 10 predictors?

(Losses ~ NumDoorstwo + BodyStylehatchback + BodyStylesedan + BodyStylewagon + DriveWheelsrwd + Height + EngineSize + Horsepower + PeakRPM + Citympg)

Auto = **na.omit**(**read.csv**("AutoLoss.csv",  
 header = T,  
 na.strings = "?",  
 stringsAsFactors = T))  
  
auto.bss <- **regsubsets**(Losses **~** .,  
 data = Auto,  
 nvmax = 15)  
**coef**(auto.bss, 10)

## (Intercept) NumDoorstwo BodyStylehatchback BodyStylesedan   
## 309.0160644 25.6751714 17.9176511 30.2551608   
## BodyStylewagon DriveWheelsrwd Height EngineSize   
## 13.6514238 17.9667900 -4.7496012 0.2201527   
## Horsepower PeakRPM Citympg   
## -0.1953871 0.0107142 -1.2483672

1. Allowing up to 15 predictors, what is the best model according to Cp criterion? State the predictors in the best model and their coefficients. Comment on predictors: What types of cars tend to have higher losses? What types of cars tend to have lower losses?

Best model according to C\_p criterion: (Losses ~ NumDoorstwo + BodyStylesedan + DriveWheelsrwd + Height + Citympg)According to predictor coefficients, cars with two doors, sedan body style, and rear-wheel drive tend to have higher losses. Cars with greater height and city mpg tend to have lower losses.

**coef**(auto.bss, **which.min**(**summary**(auto.bss)**$**cp))

## (Intercept) NumDoorstwo BodyStylesedan DriveWheelsrwd Height   
## 406.180677 26.523835 16.078951 16.373892 -5.136322   
## Citympg   
## -1.214444

1. Using the forward stepwise selection method, what would be the best model with 5 predictors? (State the predictors included in the model and their coefficients.)

auto.fwd <- **regsubsets**(Losses **~** .,  
 data = Auto,  
 nvmax = 15,  
 method = 'forward')  
**coef**(auto.fwd, 5)

## (Intercept) NumDoorstwo BodyStylesedan DriveWheelsrwd Height   
## 406.180677 26.523835 16.078951 16.373892 -5.136322   
## Citympg   
## -1.214444

1. Using the backward stepwise selection method, what would be the best model with 5 predictors? (State the predictors included in the model and their coefficients.)

auto.bwd <- **regsubsets**(Losses **~** .,  
 data = Auto,  
 nvmax = 15,  
 method = 'backward')  
**coef**(auto.bwd, 5)

## (Intercept) NumDoorstwo BodyStylesedan DriveWheelsrwd Height   
## 406.180677 26.523835 16.078951 16.373892 -5.136322   
## Citympg   
## -1.214444

1. How do the models in (c) and (d) compare to the model obtained by the best subset selection method?

If we are comparing all three models using the C\_p criterion, all these three models are the same.

(Parts (a)-(b) 1 point each, Parts (c)-(e) total 2 points)