**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>

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

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

(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.

To start your work on this question, read the data in AutoLoss.csv to a data frame called AutoLoss. This data set has missing values, marked as “?” in the data file, so we will read the data to make sure that we identify the rows with ?s, so that we can remove them. To do so, run the following two lines of code: The first one replaces ?s with NA while reading the data from the .csv file, and the second one removes all the observations with any NA.

AutoLoss <- read.csv("AutoLoss.csv", na.strings = "?")

AutoLoss <- na.omit(AutoLoss)

1. 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?

Note: I will add extra information to help you understand one subtlety about the output. Please pay attention to the names of predictors displayed in the output – for example, the output will not show BodyStyle, which is one of the columns in your data, as a predictor. Instead, you will notice predictor names BodyStylehardtop, BodyStylehatchback, BodyStylesedan, BodyStylewagon. This is because BodyStyle is a qualitative variable, which can take the values “hardtop,” “hatchback,” “sedan,” “wagon,” and “convertible.” R is replacing the qualitative variable “BodyStyle” with four columns (BodyStylehardtop, BodyStylehatchback, BodyStylesedan, BodyStylewagon), which are 0-1 columns. For example, BodyStylehardtop would be 1 if the car’s BodyStyle is hardtop and 0 otherwise. BodyStylehatchback would be 1 if the car’s BodysStyle is hatchback, and 0 otherwise. Here, R is automating it for you when you run regsubsets().

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?

Hint: In part (a), you used regsubsets() to determine the best model with k predictors for k = 1, 2, …, 15. Then, compare the Cp of the best model with 1 predictor versus the Cp of the best model with 2 predictors versus …. the Cp of the best model with 15 predictors. (For guidance, check how we made similar comparisons for Hitters data set in class.) The model with the lowest Cp is the best model.

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.)
2. 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.)
3. How do the models in (c) and (d) compare to the model obtained by the best subset selection method?

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

**APPENDIX: Description of the AutoLoss data set**

Each row represents one particular type of vehicle. The columns in the data set are as follows:

* Losses: The losses covered by the insurance company for this vehicle
* Fuel type: whether the vehicle has gas or diesel engine
* Aspiration: shows if the vehicle is standard or turbo
* NumDoors: whether the vehicle has two or four doors
* BodyStyle: whether the vehicle is convertible, hardtop, hatchback, sedan, or station
* DriveWheels: whether the vehicle is front-wheel drive, rear-wheel drive, or four-wheel drive
* Length: length of the vehicle
* Width: width of the vehicle
* Height: height of the vehicle
* Weight: weight of the vehicle
* EngineSize: Engine size of the vehicle
* Horsepower: Horsepower of the vehicle
* PeakRPM: The peak RPM the vehicle can reach
* Citympg: The mpg of the vehicle in city driving
* Price: The price of the vehicle