**BU.510.650 Homework #4**

**Data Analytics Page 1 of 2**

**Spring 2021 The Johns Hopkins University**

**Arnab Bisi Carey Business School**

**Homework #4**

Due: 02/27/21, 11:59pm

**Attention:** Homeworks #1 to #6 are all individual assignments but you are always free to ask the instructor/teaching assistant questions as needed. Copy your e-mail to both of us.

For all homework assignments prepare two files. The first is an MSWord file for your answers including any necessary figures. Name this file based on your name as in “*LastName.FirstName.HWNumber.docx”.* This file should be both concise and complete. We should not have to open your **R** file to find your answers. On the other hand, this file should only contain what is necessary to answer the question.

The second file is your **R** Script. Make this file name *“LastName.FirstName.HWNumber.R*”. We will use this file if we need more information about your answers. Both files are to be submitted via Blackboard.

1. In this question, you will use the LASSO method on the AutoLoss data set from Assignment 3. Our goal is to predict the losses paid by an insurance company as a function of the predictors, which are several features of a vehicle.

To begin your work on this question, as in the previous homework, 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. Fit a LASSO model to the AutoLoss data set, using Losses as the response (output variable) and all other variables as predictors (input variables), for  values ranging from 1 to 10 in increments of 1. (See, in particular, the code we used to perform Task 1 and Task 3 on Slide 36 in Session 3, when we applied LASSO to Hitters data set.) State the coefficient estimates you obtained for  = 1 and = 10.
2. For  = 1: What predictors are included in the resulting model? For  = 10: What predictors are included in the resulting model?
3. What do you observe about the coefficient estimates you obtain as  increases?

(Total 3 points for parts (a)-(c))

1. In this question, you will continue to use the LASSO method on the AutoLoss data set. This time, you will use 5-fold cross-validation to find the best value for . (This follows up on the last question, where you ran LASSO for  values ranging from 1 to 10 in increments of 1. Now, you will be determining the best possible value for . For guidance, see the code we used to perform Tasks 8 and 10 on Slide 41 in Session 3.)
2. Fit a LASSO model to the AutoLoss data set, using Losses as the response and all other variables as predictors, using 5-fold cross-validation on the entire data to find the best value for . (R Hints: To make it a five-fold cross-validation, include nfolds=5 as an argument inside cv.glmnet(). \*\*Remember to include set.seed(1) before cv.glmnet(), so we all end up making the same split.\*\*) State the best value for 
3. Run LASSO one more time, this time using the best value for State the coefficient estimates you obtained. What predictors are included in the resulting model?

(Total 2 points for parts (a)-(b))

1. Suppose that we collect data for a group of students at the Johns Hopkins Carey Business School who recently took the written test and the road test in their driver’s license applications. The data include written test score (X1), hours of supervised practice driving (X2) and the road test results (Y, can be pass or fail). We want to predict the probability of passing the driving test based on the written test score (X1) and hours of supervised practice driving (X2). After running the logistic regression, we obtain the coefficients: .
2. Suppose that Sam will take the road test next week. His written test score is 25; he practiced 50 hours of supervised driving. Estimate the probability that Sam will pass the road test. (1 point)
3. Another student Shengqi just passed the written exam with score 24. How many hours should he practice to have 50% chance of passing the road test? (2 points)
4. (a) For the Smarket data set from “ISLR” library, using year 2005 data as test data and remaining as training data, what is your prediction accuracy for market Direction if Lag 1, Lag 2, Lag 3, Lag 4, Lag 5 and Volume are used as predictor variables? (1 point)

(b) For the same setting, if you use only Lag 1 and Lag 2 as predictor variables, what is your prediction accuracy for market Direction? (1 point)