ESI 4606 Analytics I - Foundations of Data Science
Homework 4
Due: October 19st (11:00AM), 2022

Problem 1 (1.5 points)

This question involves using R to perform the multiple linear regression using the "Auto" data from R library of "ISLR".

(a) Fit a multiple linear regression model to predict "mpg" using all other variables except "name" as the predictors. With significance level of 0.05, for which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

(b) Based on answers of (a), re-fit a smaller model until all predictors in the model are significant based on significance level of 0.05. Provide an interpretation of each coefficient in the model.

(c) Perform the log transformation of the response variable "mpg". Based on the transformed "mpg", fit a multiple linear regression model using all other variables except "name" as the predictors. With significance level of 0.05, for which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

(d) Based on answers of (c), re-fit a smaller model until all predictors in the model are significant based on significance level of 0.05. How well do the final models in (d) and (b) fit the data?

Note: **To get full points, include R codes in the appendix sections**

Problem 2 (1.5 points)

75% of the light aircraft that disappear while in flight in a certain country are subsequently discovered. Of the aircraft that are discovered, 65% have an emergency locator, whereas 95% of the aircraft not discovered do not have such a locator. Suppose a light aircraft has disappeared.

(a) What is the probability that it has an emergency locator?

(b) If it has an emergency locator, what is the probability that it will not be discovered?

(c) If it does not have an emergency locator, what is the probability that it will be discovered?

Problem 3 (2 points)

Questions in this problem should be answered using the data set files of training data "HM4-train-2022.txt" and test data "HM4-test-2022.txt". After importing the data, please change response variable "y" in each data set as a factor object.

(a) Using "y" as response variable and all other variables as predictors, fit the logistic regression model based on **training data**. Compute the confusion matrix and the mis-classification error rate for **test data**.

(b) Repeat (a) using LDA.

(c) Repeat (a) using QDA.

(d) Repeat (a) using KNN with K=2 (Note: set your random seed as 2022).

(e) Which of these methods appears to provide the best prediction results? Why?

Note: **To get full points, include R codes in the appendix sections**