ESI 4606 Analytics I - Foundations of Data Science
Homework 5
Due: Nov. 9st (11:00AM), 2022

Problem 1 (1 point)

Explain how to utilize 10-fold cross-validation to estimate test error of multiple linear regression. (Use words and mathematical notations to provide a clear description)

Problem 2 (3 points)

Questions in this problem should be answered using the data set file of "HM5.txt". $x_1, x_2, ..., x_{10}$ are predictors and y is the response variable.

(a) Perform the best subset selection to choose the best model from the possible predictors, $x_1, x_2, ..., x_{10}$. What is the best model obtained according to C_p, BIC and adjusted R^2, respectively? Show some plots to provide evidence for your answer, and report the coefficients of the corresponding best model obtained.

(b) Perform the forward stepwise selection to choose the best model from the possible predictors, $x_1, x_2, ..., x_{10}$. What is the best model obtained according to C_p, BIC and adjusted R^2, respectively? Show some plots to provide evidence for your answer, and report the coefficients of the corresponding best model obtained.

(c) Perform the backward stepwise selection to choose the best model from the possible predictors, $x_1, x_2, ..., x_{10}$. What is the best model obtained according to C_p, BIC and adjusted R^2, respectively? Show some plots to provide evidence for your answer, and report the coefficients of the corresponding best model obtained.

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

Problem 3 (1 point)

Let $p = 9$ and $n = 50$, where p is the number of predictors and n is the sample size. When linear regression is considered for data fitting, answer the following questions.

(a) To determine the best model by selecting the best subset of relevant predictors, how many models in total need to be compared if best subset selection method is used?

(b) To determine M_3, i.e., best model with 3 predictors, how many models with 3 predictors need to be compared if best subset selection method is used?

(c) To determine M_3, i.e., best model with 3 predictors, how many models with 3 predictors need to be compared if forward stepwise selection method is used?

(d) To determine M_3, i.e., best model with 3 predictors, how many models with 3 predictors need to be compared if backward stepwise selection method is used?