STA 545 Statistical Data Mining I, Fall 2020

Homework 3, due: Wednesday 9/23/2020 (1PM)

Please submit the pdf file generated by R markdown in UBlearns. Please use tables, figures, or a few sentences to answer data analysis questions.

1. (40 points) This problem involves the Boston data set, which we discussed in the data analysis example about subset selection. We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

   (a) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response?

   (b) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

   (c) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.

   (d) Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form

   $$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon,$$

   and test $H_0 : \beta_2 = \beta_3 = 0$.

2. (30 points) We perform best subset, forward selection, and backward elimination selection on a single data set. For each approach, we obtain $p+1$ models, containing $0, 1, 2, \ldots, p$ predictors. Explain your answers:

(a) Which of the three models with k predictors has the smallest training error?

(b) Which of the three models with k predictors has the smallest test error?

(c) True or False:

(i) The predictors in the k-variable model identified by forward selection are a subset of the predictors in the (k+1)-variable model identified by forward selection.

(ii) The predictors in the k-variable model identified by backward elimination are a subset of the predictors in the (k + 1)-variable model identified by backward elimination.

(iii) The predictors in the k-variable model identified by backward elimination are a subset of the predictors in the (k + 1)-variable model identified by forward selection.

(iv) The predictors in the k-variable model identified by forward selection are a subset of the predictors in the (k+1)-variable model identified by backward elimination.

(v) The predictors in the k-variable model identified by best subset are a subset of the predictors in the (k + 1)-variable model identified by best subset selection.

3. (30 points) In this exercise, we will generate simulated data, and will then use this data to perform best subset selection, forward selection, backward elimination. Make sure to use **set.seed(1)** prior to starting part (a) to ensure consistent results.

(a) Use the rnorm() function to generate a predictor X of length $n = 100$, as well as a noise vector ϵ of length $n = 100$.

(b) Generate a response vector Y of length $n = 100$ according to the model

$$Y = 1 + 2.5X + 2X^2 + X^3 + \epsilon.$$

(c) Use the regsubsets() function to perform best subset selection, forward selection, and backward elimination in order to choose the best model containing the predictors $X, X^2, ..., X^{10}$. Report the coefficients of the selected models.