Stat 897 Spring 2017 Data Analysis Assignment 4

Penn State

Due September 17, 2017

1. Use the College data found in the ISLR library. It contains 18 variables for 777 different universities and colleges in the US. The list of variables and their full description can be found on p. 54 of the book.

(a) Load the dataset College. Our objective is to predict the number of applications received using the other variables in the data set.

(b) Split your data set into a training set containing 100 observations and a test set containing the rest of the observations. For reproducibility of results use set.seed() with a value of your choosing.

(c) Perform best subset selection on the training set, and plot the training set MSE associated with the best model of each size.

(d) Plot the test set MSE associated with the best model of each size.

(e) For which model size does the test set MSE take on its minimum value? Comment on your results.

Please check that none of the best models is the intercept only model, or the model with ALL predictors. If they are, try using a different seed value to avoid it.

(f) Fit a regression model with all features to the full data containing 777 observations. Let the regression coefficients for this model be denoted by β_j. Let $\hat{\beta}_j^r$ be the estimated regression coefficient for the best model containing r features. Create a plot displaying

$$\sqrt{\sum_{j=1}(\beta_j - \hat{\beta}_j^r)^2}$$

for a range of values of r. Comment on what you observe. How does this plot compare to the test MSE plot from (d).

(g) Now use forward and backward stepwise selection with the BIC and AIC to select models (so you will get up to four best models). How do the results compare to what you obtained above in part (c) and (d)?