Assignment 5

Your Name

26 October 2023

Question 1: Generate a data set with $p = 20$ features, $n = 1000$ observations, and an associated quantitative response vector generated according to the model $Y = X\beta + \epsilon$, where β has some elements that are exactly equal to zero. Split your data set into a training data set containing 100 observations and a test set containing 900 observations.

1. Perform best subset selection on training set and plot the training set MSE associated with the best model of each size.
2. plot AIC (or Cp) for the best model of each size.
3. Plot the test MSE associated with the best model of each size. For which model size does the test MSE takes on its minimum value?
4. Compare your results to the true model used to generate the data.

Question 2: Suppose that we have n distinguishable samples and that we perform a bootstrap sampling once. Mathematically show that the expected value of the fraction of unique samples is roughly 2/3. Simulate this process in `R` and verify your answer.