STT3851 Homework 10

Dr. Lasanthi Watagoda

Due – April 22

1) We will now try to predict per capita crime rate in the Boston data set.

   (a) Try out some of the regression methods explored in this chapter, such as best subset selection, forward selection, backward selection, the lasso and ridge regression Present and discuss results for the approaches (at least 3) that you consider.

   (b) Propose a model (or set of models) that seem to perform well on this data set, and justify your answer. Make sure that you are evaluating model performance using validation set error, crossvalidation, or some other reasonable alternative, as opposed to using training error.

2) (Knowing how to conduct a simulation can be valuable for an Statistician) In this exercise, we will generate simulated data, and will then use this data to perform best subset selection.

   (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

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

   where β_1, β_2, and β_3 are constants of your choice. Note that this is the TRUE model. In general we do not know this, but in a simulation we do!

   (c) Use the `regsubsets()` function to perform best subset selection in order to choose the best model containing the predictors $X, X^2, ..., X^{10}$. What is the best model obtained according to C_p, BIC, and adjusted R^2? Show some plots to provide evidence for your answer, and report the coefficients of the best model obtained. Note you will need to use the data.frame() function to create a single data set containing both X and Y

   (d) Repeat (c), using forward selection and also using backwards selection. How does your answer compare to the results in (c)?