**Statistical Learning 5**

1. 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 of length n = 100.

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

where , , , and are constants of your choice.

(c) Use the *regsubsets()* function to perform best subset selection in order to choose the best model containing the predictors , , . . .,. What is the best model obtained according to Cp, BIC, and adjusted R2? 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.

1. We have seen that as the number of features used in a model increases, the training error will necessarily decrease, but the test error may not. We will now explore this in a simulated data set.
2. Generate a data set with p = 20 features, n = 1,000 observations, and an associated quantitative response vector generated according to the model

where has some elements that are exactly equal to zero.

(b) Split your data set into a training set containing 100 observations and a test set containing 900 observations.

(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. If it takes on its minimum value for a model containing only an intercept or a model containing all of the features, then play around with the way that you are generating the data in (a) until you come up with a scenario in which the test set MSE is minimized for an intermediate model size.

(f) How does the model at which the test set MSE is minimized compare to the true model used to generate the data? Comment on the coefficient values.

1. This question uses the variables *dis* (the weighted mean of distances to five Boston employment centers) and *nox* (nitrogen oxides concentration in parts per 10 million) from the *Boston* data. We will treat *dis* as the predictor and *nox* as the response.
2. Use the *poly()* function to fit a cubic polynomial regression to predict *nox* using *dis*. Report the regression output, and plot the resulting data and polynomial fits.
3. Plot the polynomial fits for a range of different polynomial degrees (say, from 1 to 10), and report the associated residual sum of squares.
4. Perform cross-validation or another approach to select the optimal degree for the polynomial, and explain your results.
5. Use the *bs()* function to fit a regression spline to predict *nox* using *dis*. Report the output for the fit using **four** degrees of freedom. How did you choose the knots? Plot the resulting fit.
6. Now fit a regression spline for a range of degrees of freedom, and plot the resulting fits and report the resulting RSS. Describe the results obtained.