STAT 435: Homework 2

Due 10/14

This homework should be done independently. All work should be done in Rmarkdown.

Q1. Do Question 14. of Chapter 3 of the ISLR book. (Page 125).

Q2. Before attempting this question, set the seed number in R by using `set.seed(1)` to ensure consistent results

a. Simulate a training data set of $n = 25$ observations as

$$y = \exp(x) + \epsilon$$

where x and ϵ are generated via a normal distribution with mean zero and standard deviation one. (use `rnorm()` to simulate these variables). Then do the following,

b. Fit the following four linear regression models to the above training data set (using the `lm()` function in R), (i) $y = \beta_0 + \beta_1 x + \epsilon$, (ii) $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$, (iii) $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$, (iv) $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \epsilon$.

c. Now simulate a testing data set with $n = 500$ observations from the model in part (a), by generating new values of x and ϵ.

d. Use the estimated coefficients in Part (b) to compute the test error, i.e. the $MSE = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$ of the testing data set for each of the four models computed in part (b).

e. Based on your results of Part (b), which model would you reccommend as the 'best fit model'? is the conclusion suprising?

Q3. Consider the `Hitters` data in the `ISLR` package, our objective here is to predict the `salary` variable as the response using the remaining variables.

a. Split the data into a training and testing data set.

b. Fit a linear model using least squares on the training set and report the test error obtained.

c. Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained.

d. Fit a lasso model on the training set, with λ chosen by cross validation. Report the test error obtained, along with the number of non-zero coefficients estimates.

e. Commment on the results obtained. How accurately can we predict the number of college applications recieved? Is there much difference among the test errors resulting from these three approaches?