STAT 435
SPRING QUARTER 2018

Homework # 6
Due Friday, May 18, 2018 at 12:00 PM (Noon)
Online Submission Via Canvas

Instructions: You may discuss the homework problems in small groups, but you must write up the final solutions and code yourself. Please turn in your code for the problems that involve coding. However, for the problems that involve coding, you must also provide written answers: you will receive no credit if you submit code without written answers. You might want to use Rmarkdown to prepare your assignment.

1. For this problem, you will analyze a data set of your choice, *not* taken from the `ISLR` package. I suggest choosing a data set that has $p \approx n$ or even $p > n$, since you will apply methods from Chapter 6 on this data.

   (a) Describe the data in words. Where did you get it from, and what is the data about? You will perform supervised learning on this data, so you must identify a response, Y, and features, $X_1, \ldots, X_p$. What are the values of n and p? Describe the response and the features (e.g. what are they measuring; are they quantitative or qualitative?). Plot some summary statistics of the data.

   (b) Split the data into a training set and a test set. What are the values of n and p on the training set?

   (c) Fit a linear model using least squares on the training set, and report the test error obtained.

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

   (e) 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 coefficient estimates.

   (f) Fit a principal components regression model on the training set, with M chosen by cross-validation. Report the test error obtained, along with the value of M selected by cross-validation.

   (g) Fit a partial least squares model on the training set, with M chosen by cross-validation. Report the test error obtained, along with the value of M selected by cross-validation.

(h) Comment on the results obtained. How accurately is the best model you obtained, in terms of test error? Is there much difference among the test errors resulting from these approaches? Which model do you prefer?

2. Define the basis functions $b_1(X) = I(-1 < X \leq 1) - (2X - 1)I(1 < X \leq 3)$, $b_2(X) = (X + 1)I(3 < X \leq 5) - I(5 < X \leq 6)$. We fit the linear regression model

$$Y = \beta_0 + \beta_1 b_1(X) + \beta_2 b_2(X) + \epsilon,$$

and obtain coefficient estimates $\hat{\beta}_0 = 2, \hat{\beta}_1 = -1, \hat{\beta}_2 = 2$. Sketch the estimated curve between $X = -3$ and $X = 8$. Note the intercepts, slopes, and other relevant information.