IEOR 4574: Business Analytics for OR
Spring 2016

Professor Adam Elmachtoub
Professor Jacob Leshno

Assignment 2: Linear Regression

Due date: February 18, 11:59pm

Attention: Please prepare two files for each homework assignment: a .pdf file for your answers including relevant figures, and a .R file for your relevant R scripts. File names should be `Last_First_hw.pdf` and `Last_First_hw.R`, e.g., `Obama_Barack_2.pdf` and `Obama_Barack_2.R`. Your submissions must be based on your own original work. Late submissions will not be accepted.

In this problem you will need to analyze the `CollegeData.csv` data set. This data set from 2013 represents all colleges that grant graduate degrees in the United States. We have significantly already trimmed down the original data set which can be found online on data.gov. In this assignment, we will try to figure out what factors can be used to predict the quality of a school, using SAT_AVG as our measure of quality. You can look up the meanings of the columns in `CollegeDataDictionary.csv`.

1. Download `College.csv` to your computer and read it into R. Be aware that the rows and columns are labeled. Remove any rows that have missing entries. The function `na.omit(...)` is also useful. How many rows of data do you have?

2. To make our models potentially richer, we will add more columns to our data set. Several of the columns give a numerical dollar amount. For each of these add another column corresponding to the log of the dollar amount. For each pair of original columns giving a numerical dollar amount, add the interaction term. How many covariates are now in your data set? What is the mean of each column?

3. Randomly divide your data into two parts: a training set (75%) and a test set (25%). Initialize the random generator with `set.seed(4574)`. What is the mean SAT_AVG in each set?

4. Using 5-fold cross-validation on the training set, use best subset selection to find the best model. Consider subsets of sizes from 1 to 8. Which subset size is best? What is your final model?

5. Using 5-fold cross-validation on the training set, use lasso regression to find the best model. Consider the values 0, .001, .01, .1, 1, 10, 100, 1000 for λ. Which choice of λ is the best? What is your final model?

6. Pick one of the two models final models from the previous two questions. What is the MSE of your model on the test data?

7. What insights can you take away from your final model? What kind of further investigations would you like to do based on what you have learned from your model?