IEOR 4650: Business Analytics
Spring 2019 Professor Adam Elmachtoub

Assignment 2: Linear Regression and Model Selection

Due date: February 22, 11:59am

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 be penalized at 10% per hour.

1. Consider the stock prices of GE in the years 2016 and 2017 in the file `GE.csv`. You will replicate the strategy from the financial analytics lecture.

   (a) Compute the 1D and 5D returns for each day. Here a day refers to a business day where the stock was traded. What is the average 1D and 5D return in the dataset?

   (b) Using 2016 as your training data, develop a linear regression model to predict the next day's return based on the 1D and 5D return. What is the final model using both independent variables?

   (c) Implement the long-short strategy on the 2017 data. Report your average 1D return from the strategy, as well as your final return on investment. Evaluate how the strategy performed, and what could be improved, if anything.

2. 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. I have already significantly trimmed down the original data set which can be found online on data.gov. In this assignment, you 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`.

   (a) 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?

   (b) 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 square root 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?

(c) 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?

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

(e) 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 prediction model?

(f) What is the MSE of your model (from the previous question) on the test data?

(g) What insights can you take away from your final model? As a future parent, student, taxpayer, and/or secretary of education, what kind of further investigations would you like to do based on what you have learned so far?