ISYE 6501, Week 5 HW

**Question 1**

Describe a situation or problem from your job, everyday life, current events, etc., for which a linear regression model would be appropriate. List some (up to 5) predictors that you might use

**Response –**

Linear regression would be appropriate to estimate the credit or loan worthiness of a customer for a bank by calculating a “credit score” as a response for various factors which represent the financial situation of the customer.

The following are few factors the bank could consider –

1. Total credit the customer owes at the time of application
2. Income of the customer
3. Average expenditure of the customer
4. Past credit/loan history – Any missed payments or defaults
5. Duration of relationship with the customer.

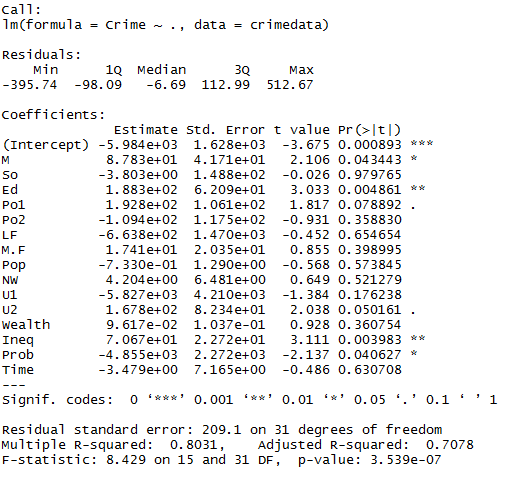
**Question 2**

Using crime data from http://www.statsci.org/data/general/uscrime.txt (description at http://www.statsci.org/data/general/uscrime.html ), use regression (a useful R function is lm or glm) to predict the observed crime rate in a city with the following data: M = 14.0 So = 0 Ed = 10.0 Po1 = 12.0 Po2 = 15.5 LF = 0.640 M.F = 94.0 Pop = 150 NW = 1.1 U1 = 0.120 U2 = 3.6 Wealth = 3200 Ineq = 20.1 Prob = 0.04 Time = 39.0 Show your model (factors used and their coefficients), the software output, and the quality of fit.

**Response –**

Please see the R script attached and explanation in the subsequent pages.

First, the linear regression model is built using all the factors and the results of the model are as follows.



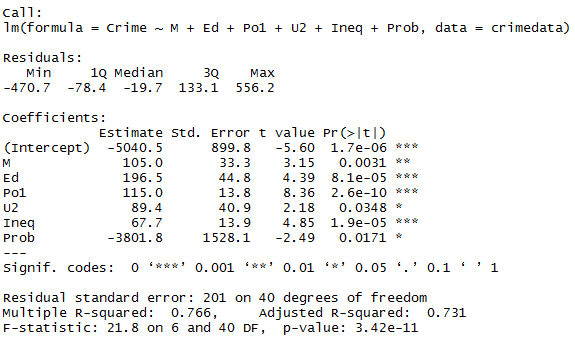
The adjusted R-squared value is 0.7078.

Upon testing with the test data point, the crime rate value came out to be 155 which is less than most of the crime values in the data set. Also, from the summary above it is evident that not all the factors are considered significant in the regression model.

On performing cross-validation, the R^2 value is observed to be 0.413, which is low to accept the model as a good fit

A number of iterations are performed, eliminating the factor with highest p value and building a regression model on the remaining factors until all the considered factors turn out to be significant.

The final regression model (model6 in code) is summarized as below –



The adjusted R-squared value is 0.731, which is higher than the first regression model created. The crime rate value using this regression model is 1304 which appears to be reasonable considering the input dataset.

On cross-validating the regression model, the R^2 value came out to be **0.638** which is much better than the earlier regression model and can be considered acceptable. The possible reasons for low value is small dataset that is used to build the regression model.

The final equation to find crime rate using the regression model is –

Crime rate = -5040.5 + 105 \*M + 196.5\*Ed + 115\*Po1 + 89.4\*U2 + 67.7\*Ineq – 3801.8\*Prob