**Module 3 – Report**

**Name: Sunil Raj Thota**

**NUID: 001099670**

**Date: 02/04/2021**

**Title: Regularization Assignment**

**ALY 6015 – Intermediate Analytics**

**Prof. Roseanna Hopper**

**Introduction**

In this task, I used R and R studio to perform Descriptive Statistics, Linear Regression, and Regularization techniques on various preloaded datasets provided by CRAN. We can notice a list of all data sets in R using the function called data().

The dataset used for the analysis is the “mtcars”. It has the information of 11 variables and 32 observations. I have mostly focused on Descriptive Statistics, Exploratory Data Analysis, Linear Regression, Inferential Statistics, and Plots.

Descriptive Statistics provides us the basic information and synopsis of the given data set. Inferential Analysis provides us with insights about the performance of the data and helps us to forecast the movements for the future.

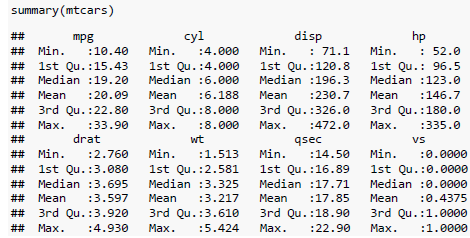
Linear Regression is mainly utilized for performing projecting analysis. The outcomes are for understanding the relation between 1 dependent and 1 or more independent variables. I have used the regression techniques and visualized the data using different plotting techniques that are available in R.

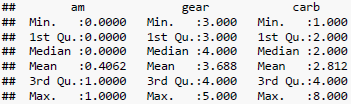
The dataset used for the analysis is the “mtcars” to discover the cross validation, regularization, and lm models on its variables. Also, used additional packages such as glmnet, biglasso, bigmemory, and ncvreg libraries.

**Analysis**

The “mtcars” dataset is comprised of R’s dataset package and loaded for the Data Analysis. It contains 32 observations and 11 variables.

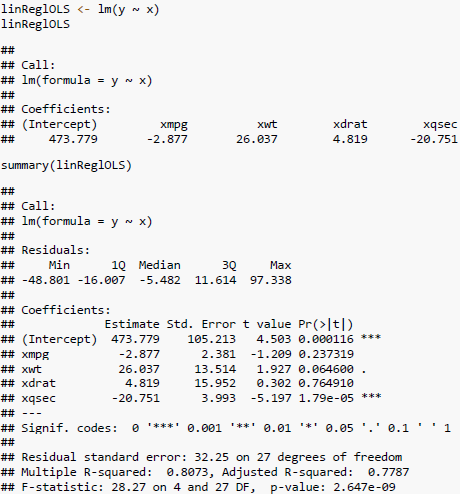
“mtcars” dataset is loaded using the command data(mtcars). We can view the dataset by using the command View(mtcars). The first six rows are printed in the console by using the command head(mtcars) and the last 6 rows by using tail(mtcars). Summary of the Dataset has been used to breakdown the complex understanding of the dataset into simple and easy to analyze way. It can be attained using the command summary(mtcars). Also, observed the structure of the dataset

****

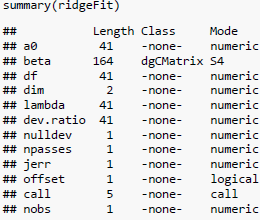
****

Let's perform some Regularization analysis and techniques using "mtcars" data set. This data set is readily available in the R Studio and can be loaded to the work space in R Studio. Or we can also install the packages by using install.packages("packagename") command. Once it is loaded, we can use it in the code for further analysis and calculations.

Here "y" variable is taken as the response variable. Here "x" is assigned with a matrix of predictor variables. In this, we need to regress "y" on the predictors in "x" using Ordinary Least Squares(OLS). The regression model was taken between "y" and "x". Summary() gives us the descriptive stats and hypothesis testing values like Standard Error, p-Value, t-Value, r-squared value, f-Statistic, Degrees of Freedom, and etc., This model is used as a baseline model to collate with the next upcoming models



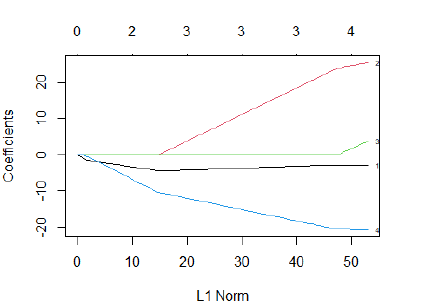
If the alpha is = 0 is corresponding to Ridge regression and alpha between 0 and 1 is corresponding to Elastic Net regression.



Setting the range of lambda values and Using glmnet() method to build the ridge regression in R. Checking the model using the Summary()

LASSO regression is performed and for that to happen we use "glmnet" package from the packages tab to install or simply use install.packages("glmnet") command. Now, let's load the "glmnet" in our work space to regularize the model using LASSO and plot it using plot().

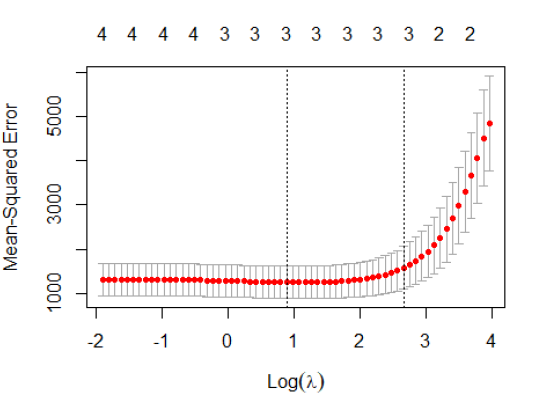
This plot indicates at which stage each coefficient shrinks to 0. and the lines depicts the values used by various other coefficients



Lasso Regression is a technique to fit a regression model if multicollinearity is present in the data. The least squares regression attempts to discover the coefficient approximations that minimize the SSR. In this, we select a value for λ that gives the lowermost likely trial MSE

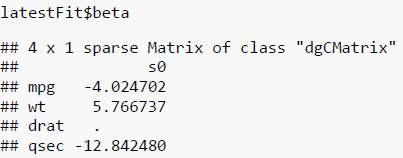
Here, Cross Validation is used to get the best value of lambda and plot the curve using plot(). It is possible with cv.glmnet() method. nlambda signifies the number of lambda values in sequence. In general, nlambda values must be above 100.

We will perform K Fold Cross Validation to know the lambda value that gives the lowest trial MSE.

****

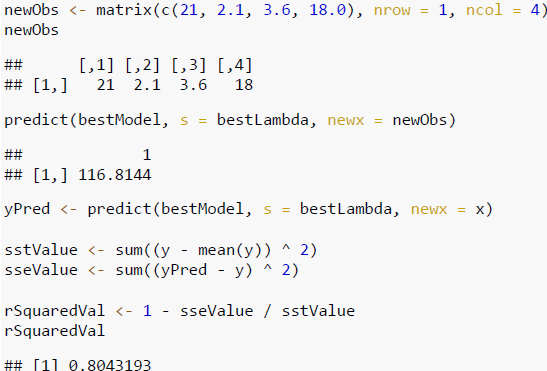
The lambda value that minimizes the test MSE turns out to be 4.662789.

Let's find optimal lambda value that minimizes test MSE and perform K-Fold Cross validation to find optimal lambda value. At last, let's produce the plot of test MSE by lambda value. From the plot we can depict that the value of lambda increased when the number of selected variables narrows down. This tells that higher the value of lambda, more shrink the selection is. Now, we find the min. value of lambda to get the best fit



Here, we use the minimum lambda value again in glmnet() function to get the best latest fit. Now we use a higher value of lambda that is within one standard error of the minimum to check its effect on shrinkage. There is 1 coefficient namely "drat" whose values have become 0. It's clear that this variable is not so necessary to determine the value of "y". LASSO tells that only 3 variables are necessary on which y depends. Thus, the shrinkage increases.

For example, we have a new car with the following attributes:



This model is used to predict the value for “hp”. The model predicts this car to have an “hp” value of 116.8144. The R-squared turns out to be 0.8043193. The best model was intelligent to explain 80.43% of the difference in the response values of the dataset.

To find the coefficients of best model, let's define a new observation and use LASSO regression model to predict response value. Use fitted best model to make predictions. Let's find SST, SSE, and R-Squared values for the new observation

**Conclusion**

To perform multiple analytical operations, we use R on widespread range of data like mtcars dataset. R is a very powerful tool to perform analysis which is mainly built by researchers, statisticians, and developers. In the beginning we did Descriptive Statistics Analysis, Exploratory Data Analysis, and continued with the Linear Regression and Linear Regression and Regularization Techniques to get understandings from the provided data. Plotted relevant parameters, to helps us to understand the data in a clear way. Used glmnet, biglasso, bigmemory, and ncvreg libraries.

The 1-sample test, 2-sample t-test, paired t-test, a test of equal or given proportions, and f-test for the mtcars datasets to solve given business problems. Ho and Ha are essential in performing the Hypothesis testing for the datasets to make and take data-driven business decisions.

**References**

[1] J H Maindonald, Using R for Data Analysis and Graphics was retrieved from https://cran.r-project.org/doc/contrib/usingR.pdf

[2] https://github.com/YaohuiZeng/biglasso/blob/master/vignettes/biglasso.pdf

[3] Yaohui Zeng, Patrick Breheny (11th March, 2018) The biglasso Package: A Memory- and Computation-Efficient Solver for Lasso Model Fitting with Big Data in R was retrieved from https://arxiv.org/pdf/1701.05936.pdf