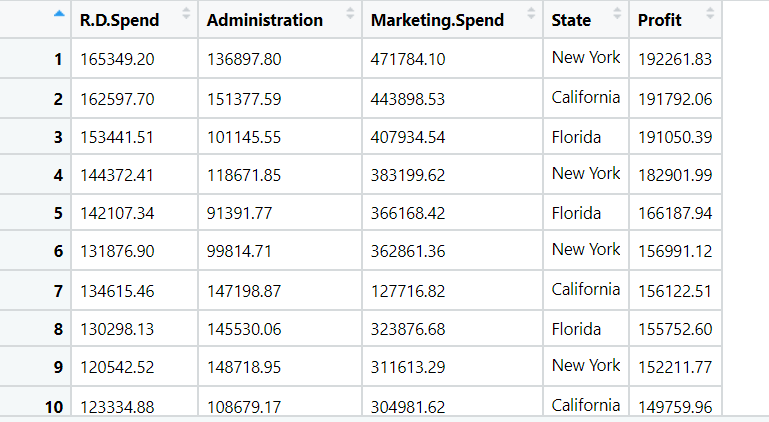
# Lasso & Ridge Regression (Module - 8)

## You have already done the Multiple Linear Regression for the following attached data. Now perform the Lasso & Ridge Regression on all of these.

1.) Prepare a prediction model for profit of 50\_startups data.



**Ans:**

**Analyzing the input and output variables:**

* Input Variables (x) = R.D Spend, Administration, Marketing Spend, State
* Output Variable(y) = Profit

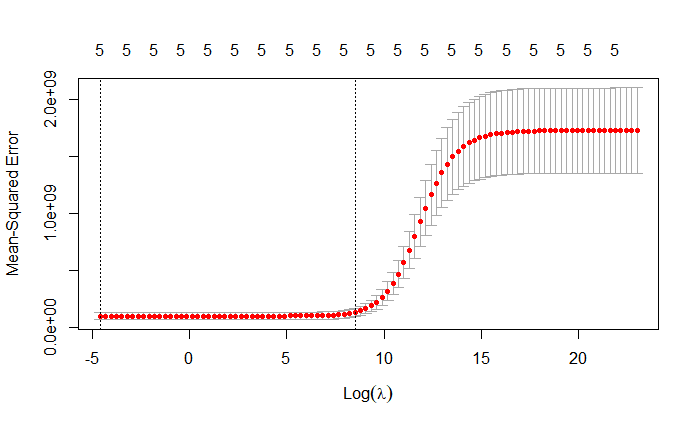
**Data Preprocessing:**

* To make the easy access of variables of input and output, columns are rearranged.
* In R , dummy variable are created automatically when object(x) is created as model. Matrix command has inbuild feature of converting dummy variables whereas, in python we need to write a code for creation of dummy variables.

**Creating the grid:**

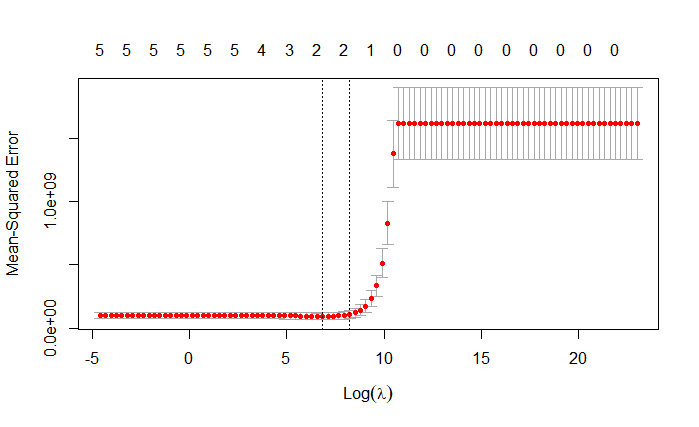
* In R , grid is created for lambda values and the range can be anything and tuning hyper parameter function is to minimize the error value.
* But in python the alpha value will represent as a lambda value.

**Ridge Regression:**

****

* The red dotted line is the cross validation and the upper and lower grey horizontal lines are std deviation lines (Error bars)
* The 2 vertical lines gives the range of Mean squared error, the first line is the minimum and the second line is the maximum lambda value and one std error away from the min lambda value gives the right model and max give the most regularized model
* As there is no much information from the model summary, we need to do the cross validation or K-fold cross validation where k values can be any, as the K value increases the best will be the R^2 increased.
* The R square value in Ridge regression is 0.95
* The Alpha value in R will always be ZERO and in python we can take any value it will act as a Lambda.
* The P- values are insignificant for RD & State

**Lasso Regression:**

****

* The red dotted line is the cross validation and the upper and lower horizontal lines are std deviation lines (Error bars)
* The 2 vertical lines gives the range of Mean squared error, the first line is the min and the second line is the max lambda value and one std error away from the min lambda value gives the right model and max give most regularized model
* As there is no much information from the model summary, we need to do cross validation or K-fold cross validation where k values can be any, as the K value increases the best will be the R^2 increased.
* The R square value in Ridge regression is 0.94
* The Alpha value in R will always be One and in python I can be taken any value and alpha value will act as a Lambda in python
* The P- values are insignificant for RD & State

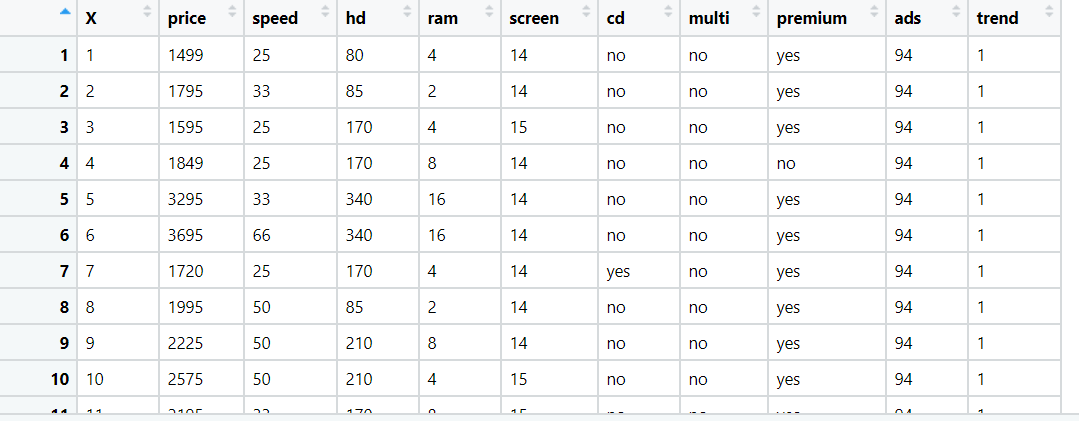
**Conclusion:**

* **After the regularization with Ridge, the Rmse value has been reduce from 8854.76 to 8019.57 which indicates that the Ridge model is the best fit model.**

**Note: 1) The R^2 value of both Lasso and Elastic-net regression is almost same.**

**2) The Lambda values are taken as log values to do scaling of the numbers.**

2.) Predict the sales of the computer



**Ans:**

**Analyzing the input and output variables:**

* Input Variables(x) = speed, hd, ram, screen, cd, multi, premium, ads, trend
* Output Variable(y) = Sales Price

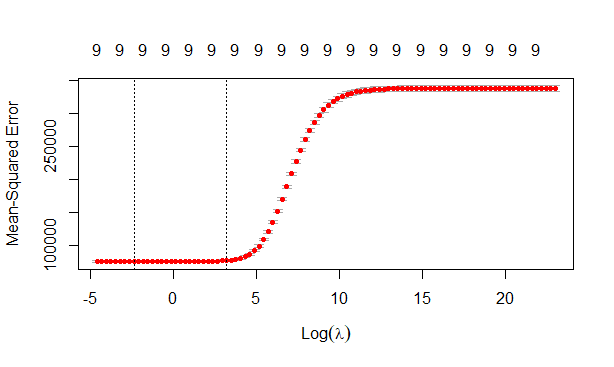
**Data Preprocessing:**

* In this problem, the variables are already arranged in order but the unnecessary columns are removed so that they should not impact on the output.
* In R, dummy variable is created automatically when object(x) is created as model. Matrix command has inbuild function of converting dummy variables but in python we need to write a code for creation of dummy variables.

**Creating the grid:**

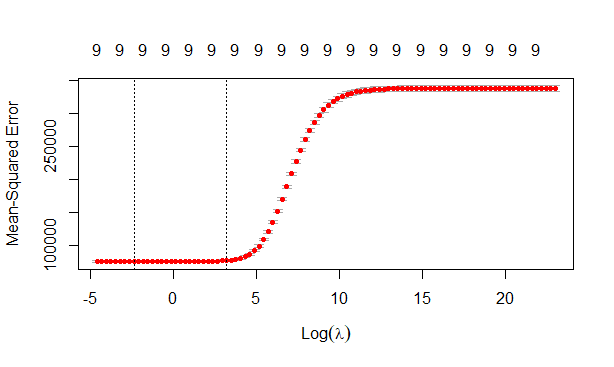
* In R grid is created for lambda values and the range can be any and this lambda value will be a tuning hyper parameter to minimize the error value
* But in python the alpha value will act as a lambda value

**Ridge Regression:**



* The red dotted line is the cross validation and the upper and lower horizontal lines are std deviation lines (Error bars)
* The 2 vertical lines gives the range of Mean squared error, the first line is the min and the second line is the max lambda value and one std error away from the min lambda value gives the right model and max give most regularized model
* As there is no much information from the model summary, we need to do cross validation or K-fold cross validation where k values can be any, as the K value increases the best will be the R^2 increased.
* The optimum lambda value = 0.09 and R square value in Ridge regression is 0.77
* The Alpha value in R will always be ZERO and in python we can take any value and it act as a Lambda.

**Lasso Regression:**



* The red dotted line is the cross validation and the upper and lower horizontal lines are std deviation lines (Error bars).
* The 2 vertical lines gives the range of Mean squared error, the first line is the min and the second line is the max lambda value and one std error away from the min lambda value gives the right model and max give most regularized model.
* As there is no much information from the model summary, we need to do cross validation or K-fold cross validation where k values can be any, as the K value increases the best will be the R^2 increased.
* The optimum lambda value = 0.01 and R square value in Ridge regression is 0.77
* The Alpha value in R will always be One and in python I can be taken any value and it will act as a Lambda.

**Conclusion:**

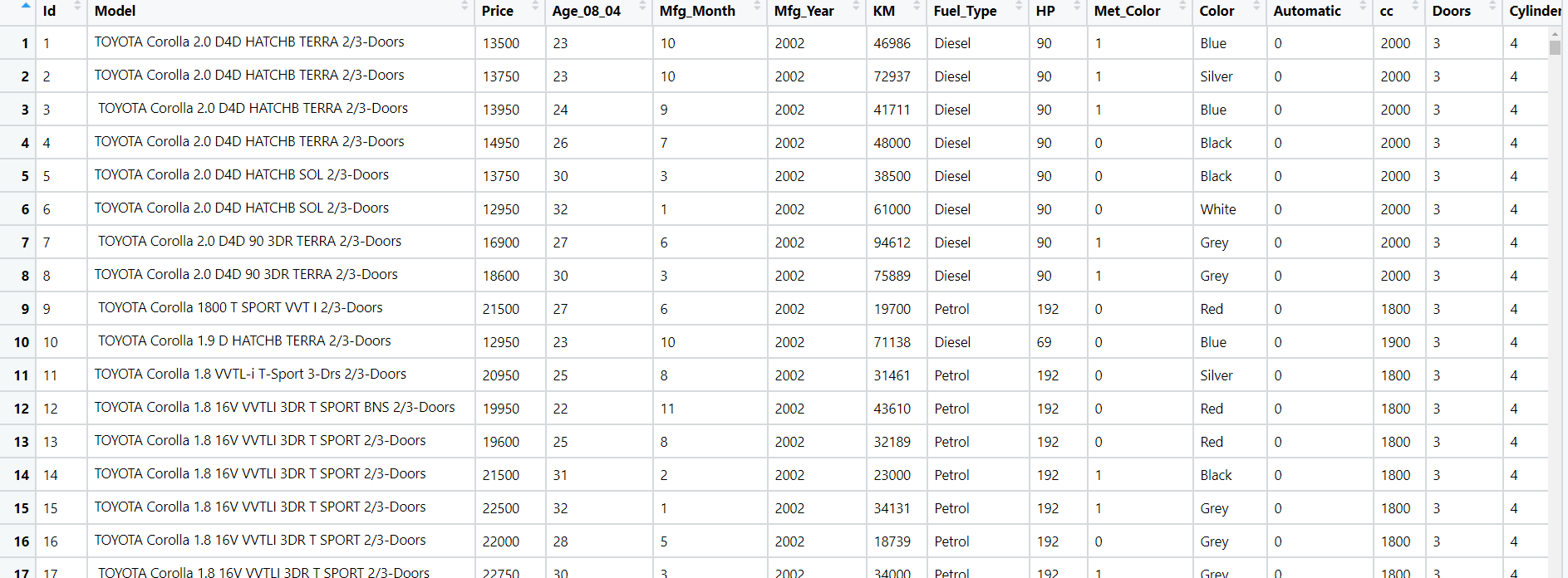
**Rmse value is more for both Ridge and lasso regression.**

**Rmse value is less in Multilinear model.**

* **As a Conclusion after comparing 3 model, Multilinear seems to be the best fit model.**

**Note: 1) The R^2 value of both Lasso and Elastic-net regression is almost same.**

**2) The Lambda values are taken as log values to do scaling of the numbers**

3.) Consider only the below columns and prepare a prediction model for predicting Price.

Corolla < Corolla[c("Price","Age\_08\_04","KM","HP","cc","Doors","Gears","Quarterly\_Tax","Weight")]

**Analyzing the input and output variables:**

* Input Variables(x) = Other Variables
* Output Variable(y) = Sales Price

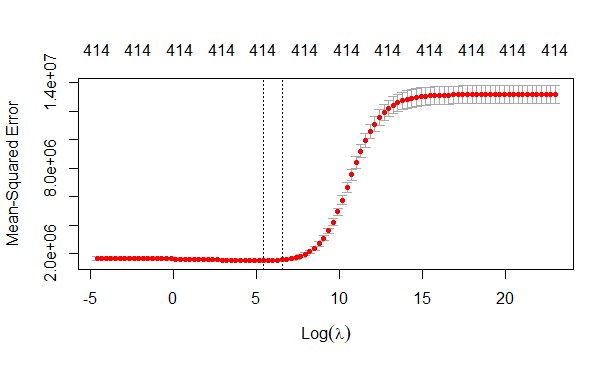
**Data Preprocessing:**

* To make the easy access of variables of input and output, columns are rearranged.
* The unnecessary columns are removed so that they should not impact on the output.
* In R, dummy variable is created automatically when object(x) is created as model. Matrix command has inbuild nature of converting dummy variables but in python we need to write a code for creation of dummy variables.

**Creating the grid:**

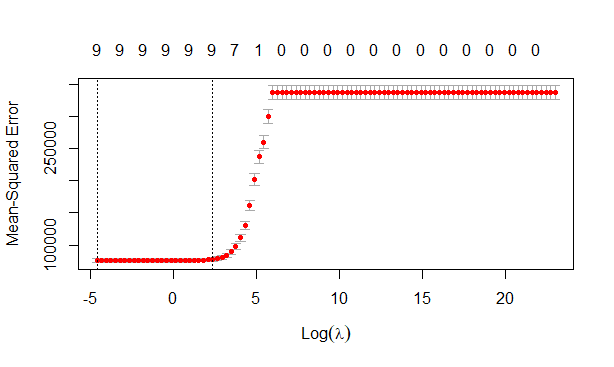
* In R grid is created for lambda values and the range can be any and this lambda value will be a tuning hyper parameter to minimize the error value
* But in python the alpha value will act as a lambda value

**Ridge Regression:**



* The red dotted line is the cross validation and the upper and lower horizontal lines are std deviation lines (Error bars)
* The 2 vertical lines gives the range of Mean squared error, the first line is the min and the second line is the max lambda value and one std error away from the min lambda value gives the right model and max give most regularized model
* As there is no much information from the model summary, we need to do cross validation or K-fold cross validation where k values can be any, as the K value increases the best will be the R^2 increased.
* The optimum lambda value = 231.013 and R square value in Ridge regression is 0.96
* The Alpha value in R will always be ZERO and in python I can be taken any value and it will act as a Lambda.

**Lasso Regression:**

****

* The red dotted line is the cross validation and the upper and lower horizontal lines are std deviation lines (Error bars)
* The 2 vertical lines gives the range of Mean squared error, the first line is the min and the second line is the max lambda value and one std error away from the min lambda value gives the right model and max give most regularized model
* As there is no much information from the model summary, we need to do cross validation or K-fold cross validation where k values can be any, as the K value increases the best will be the R^2 increased.
* The optimum lambda value = 75.64 and R square value in Ridge regression is 0.92
* The Alpha value in R will always be One and in python we can take any value and it will act as a Lambda.

**Conclusion:**

**Rmse value is more for both Ridge and lasso regression.**

**Rmse value is less in Multilinear model.**

* **As a Conclusion after comparing 3 model, Multilinear seems to be the best fit model.**

**Note: 1) The R^2 value of both Lasso and Elastic-net regression will be almost same**

**2) The Lambda values are taken as log values to do scaling of the numbers.**

**Hints:**

1. Business Problem
   1. Objective
   2. Constraints (if any)
2. Data Pre-processing

2.1 Data cleaning, Feature Engineering, EDA etc.

1. Model Building
   1. Partition the dataset
   2. Model(s) - Reasons to choose any algorithm
   3. Model(s) Improvement steps
   4. Model Evaluation
   5. Python and R codes
2. Deployment

4.1 Deploy solutions using R shiny and Python Flask.

1. Result Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.

**Note:**

1. For each assignment the solution should be submitted in the format
2. Research and Perform all possible steps for improving the model(s) accuracy.

Ex: Transformations, Feature Engineering, Hyper Parameter tuning, Outlier treatment, etc.

1. All the codes (executable programs) are running without error
2. Documentation of the module should be submitted along with R & Python codes, elaborating on every step mentioned here.