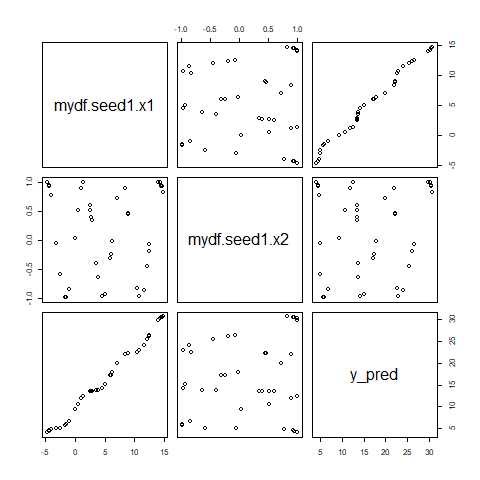
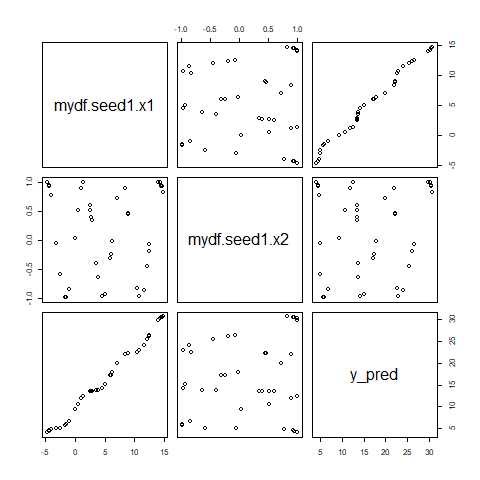
|  |  |  |  |
| --- | --- | --- | --- |
| Method  Data mining Assignment 2  Vivek Pandit 130010009 | RMSE | MAE | Regression Coefficients |
| Lm1 | 4.407997 | 3.594617 | Coefficients:  (Intercept) mydf.s1$x1 mydf.s1$x2  9.262 1.384 1.289 |
| tr.lm | 4.562436 | 4.030695 | Coefficients:  (Intercept) mydf.s1$x1 mydf.s1$x2  9.262 1.384 1.289 |
| tr.svmRadial | 5.41109  4.90079  4.792328 | 4.404792  3.966591  3.899442 | C = 0.25  0.5  1.00  The final values used for the model were sigma = 0.6722367 and C = 1. |
| tr.svmLinear | 4.678353 | 4.192767 | Tuning parameter 'C' was held constant at a value  of 1 |
| tr.ridge | 4.795537  4.795544  4.802368 | 4.069813  4.069816  4.072137 | Lambda = 0e+00  1e-04  1e-01  The final value used for the model was lambda = 0. |
| tr.lasso | 8.596943  5.461069  4.578189 | 7.135601  4.494207  3.841591 | Fraction = 0.1  0.5  0.9  The final value used for the model was fraction =  0.9. |
| tr.enet | 4.554249 | 3.842732 | The final values used for the model were fraction = 1 and lambda = 0.1. |

Plot of data.frame(x1,x2,ypred) for

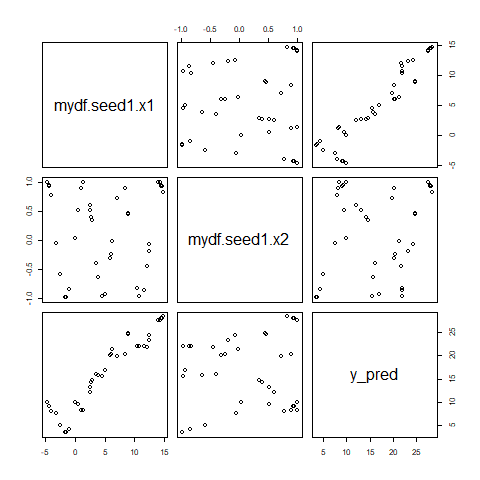
Lm1 ->



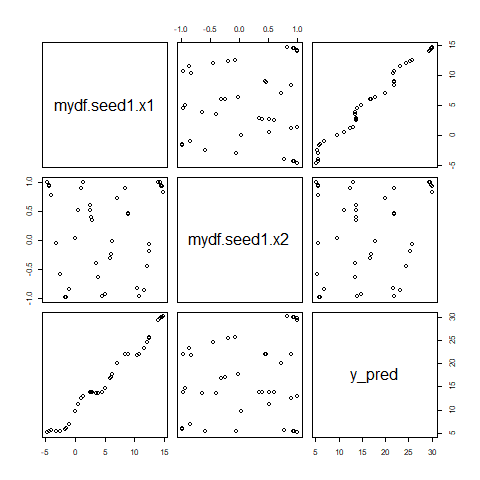
Tr.lm ->



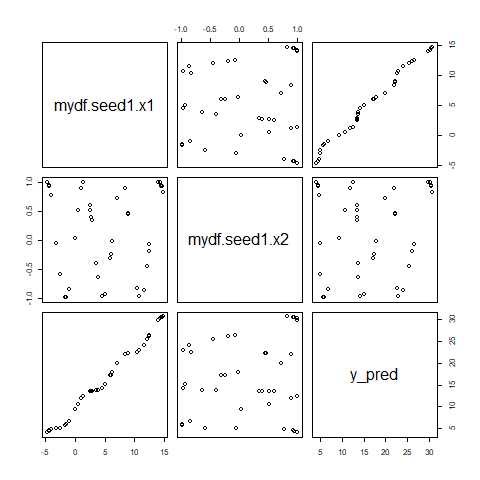
Tr.svmradial - >



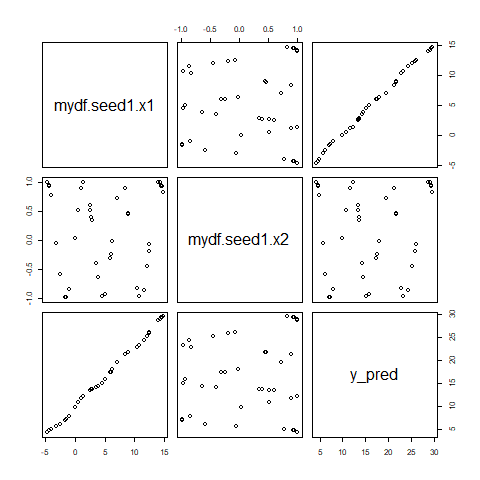
Tr.svmlinear - >



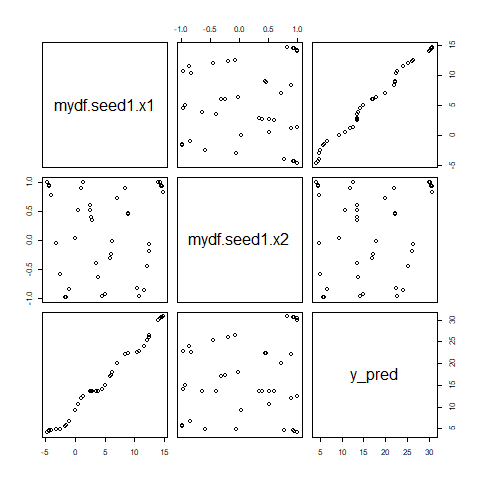
Tr.ridge ->



Tr.lasso - >



Tr.enet ->



For all the models, there seems to be a linear relationship, with curves in the middle, between y\_pred and x1. Also the plot between x2 and y\_pred is scattered, thereby indicating there is not much of a linear relationship between x2 & y\_pred

Q7.

Training the model means updating the parameters of the model such that a loss function, that is defined for each model, is optimized. This requires calculating the loss function for the training data over an initial value of weights and updating them using an iterative strategy. Finally after it has converged, the parameters obtained are the trained parameters and the model can be tested on the test data.

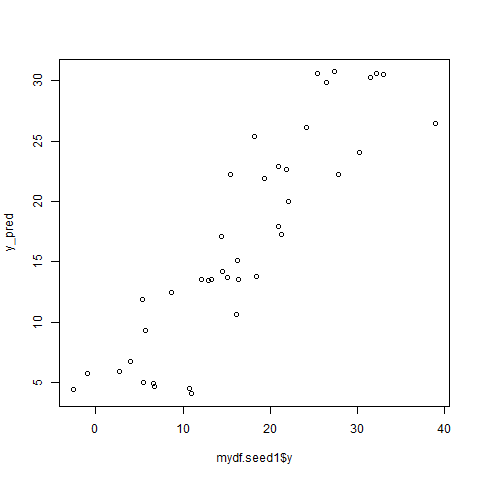
Train control object is an object for k fold cross validation training strategy. Here, the dataset is divided in to k parts and one part Is used as test data and other k-1 parts used as training dataset. This is repeated over all such parts becoming the test data at least once. The train control object helps train the model using k fold cross validation strategy.

Cross validation is basically dividing the dataset into training and testing. The model is trained on the training data set and validated on the test dataset. Then the model is trained using the previous test dataset and validated on the previous training data set. The parameters obtained from the above two cases are averaged out to create a final trained model.

Resampling is basically finding the precision using small groups of the dataset and validating the model using random subsets of the model. It is a type of method used to help estimate the sample statistics.

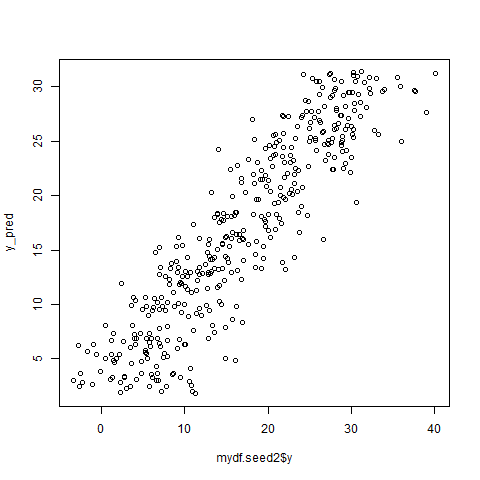
Q8. Plots of Y predicted vs Y given along with correlation coefficient for each plot

Lm1 ->



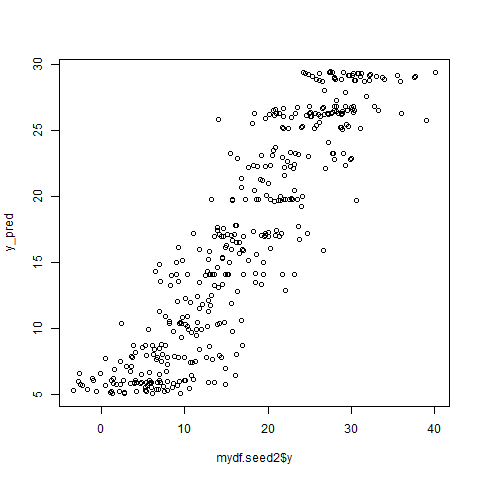
Correlation coefficient = 0.8921678

Tr.lm ->



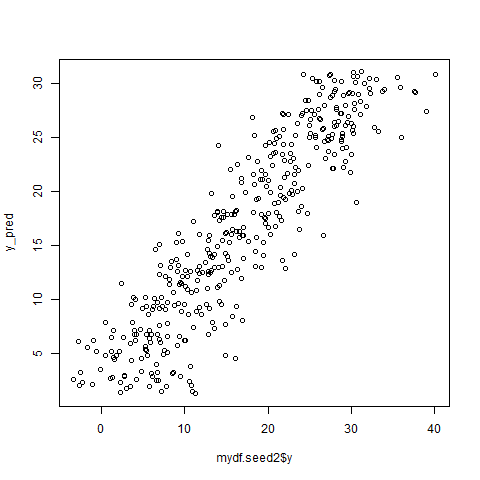
Corr coeff. = 0.8921678

Tr.svmRadial ->



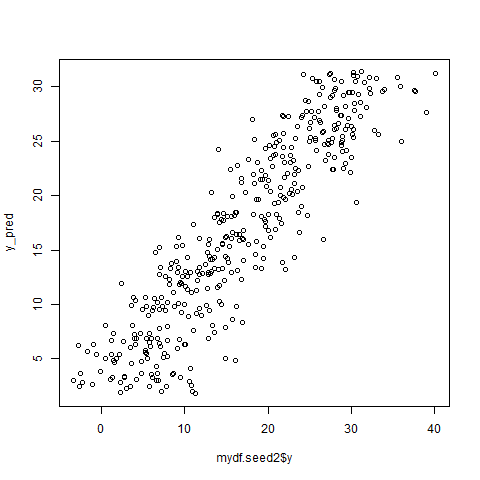
Corr coeff = 0.9025192

Tr.svmlinear ->



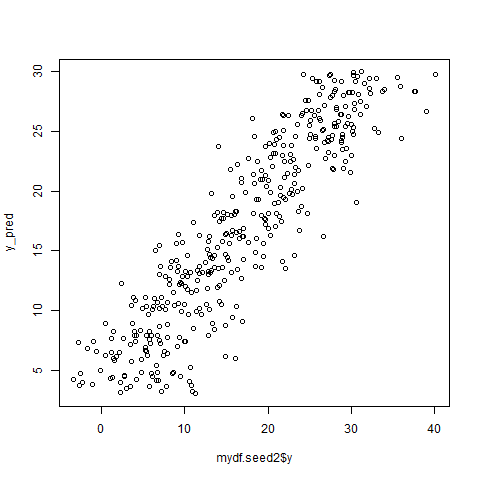
Corr coeff. = 0.8915858

Tr.ridge ->



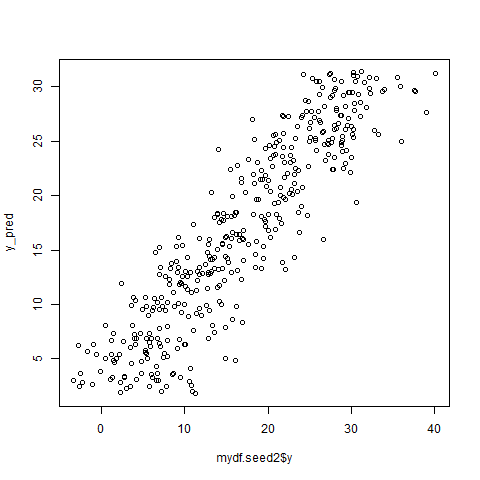
Corr coeff. = 0.8922

Tr.lasso ->



Corr coeff = 0.8910034

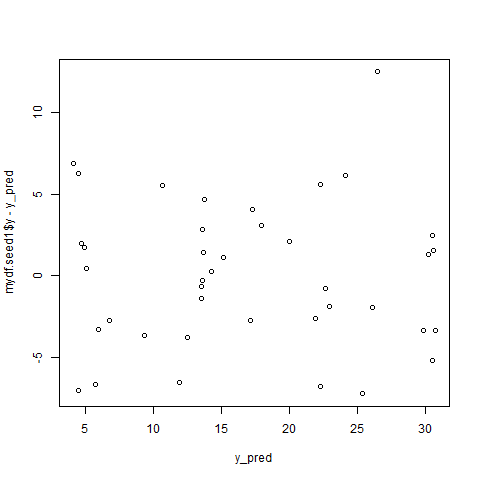
Tr.enet ->



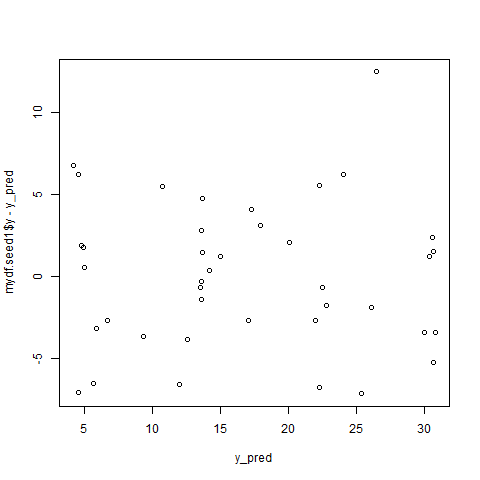
Corr coeff = 0.8921448

Y\_pred and y value are highlighy correlated and ideally, it should ideally show the corr coefficient to be 1.

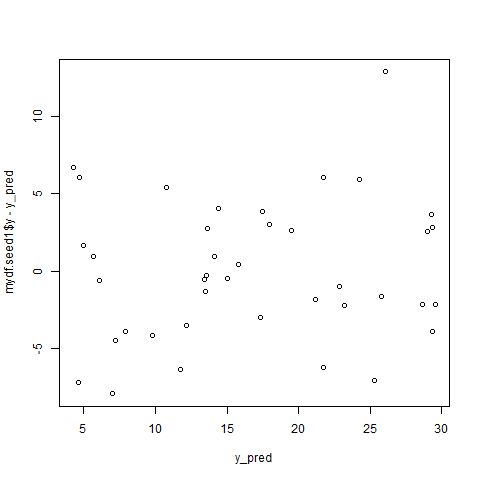
The plots shown below are that of residuals vs y\_pred. tr.T 🡨lm1.Here the residuals are scattered and random hence do not show any set pattern



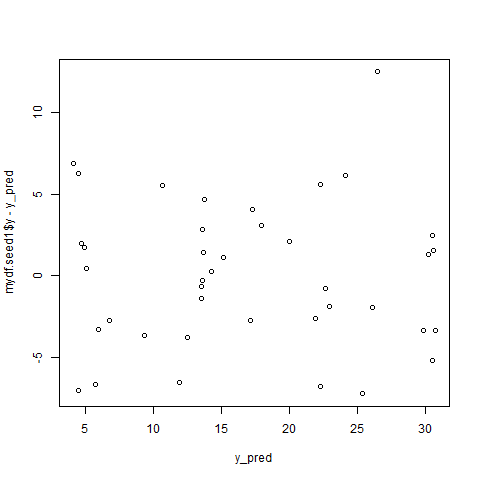
🡨tr.enet Here as well, the residuals are scattered.



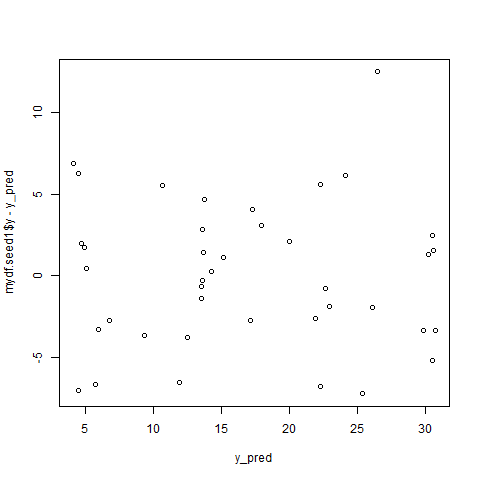
🡨 tr.lasso.



🡨tr.lm

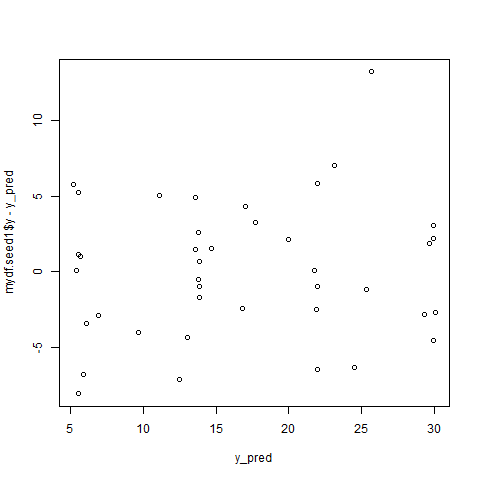


🡨 tr.ridge

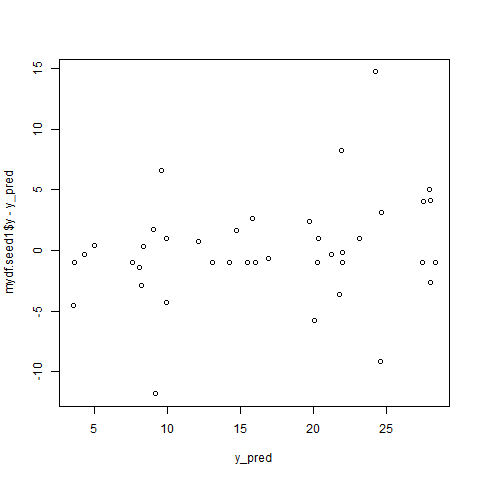


Lasso, lm and ridge, all have quite scattered residuals and do not follow a set pattern

🡨tr.svmLinear



🡨tr.svmRadial



Here the noise shows higher frequency at value = 0. Hence the noise can be estimated to follow a normal distribution with mean 0.

Q9. With change in seed and sample of the original dataset, the graphs slightly vary in terms of how linear it is or how much noise is available. The basic structure of the plots are still intact.

Q10. The best fit model looks slightly different compared to the original divided dataset and they have much more variance compared to other algorithms.