ILLINOIS INSTITUTE OF TECHONOLOGY

MATH 564 Applied Statistics

Statistical Analysis and Modelling of Real-Estate Price Prediction

Shriya Prasanna
sprasanna@hawk.iit.edu
A20521733

Naveen Raju Sreerama Raju Govinda Raju
nsreeramarajugovinda@hawk.iit.edu
A20516868

Raghunath Babu
rbabu@hawk.iit.edu
A20511598

Prof. Lulu Kang

Submission Date: 2nd December 2022

IT IS DECLARED WITH MUTUAL AGREEMENT THAT EACH MEMBER HAS EQUALLY CONTRIBUTED IN THE PROJECT

Tables of Contents

1. Abstract

This project presents the concepts of statistical modeling to a real-estate dataset which has a significant volume of influential outliers, unequal variance and non-linearity up to a certain degree. Further the methods of regression are explored to increase the R-square and decrease RMSE. The dataset is about predicting the cost of occupied homes based of several features. Upon eliminating the features with high multicollinearity and removing the influential outliers using VIF, Cooks distance and DFBETAS, this project proceeds to split the dataset into two different population samples based on the qualitative variable. Then statistical modelling techniques like Ridge Regression, Robust Regression, Regression tree, Random Forest were employed and analyzed in interest of obtaining maximum R-square and minimum RMSE which would aid in choosing the best model.

2. Introduction

We have considered a real-estate dataset, where we have to predict the cost of occupied homes (MEDV). The MEDV is considered as response variable (Y). For modelling the data, we consider 13 features that would help in predicting the response value Y. The features include crime rate per capita, number of rooms, accessibility of highways, property tax, student- teacher ratio in the town and few more factors that would play a vital role in predicting the cost of homes.

Below are the challenges faced with respect to data:

1. Missing values.
2. Eliminating Multicollinearity.
3. Rectifying curvilinear predictor variables.
4. Residual Analysis.
5. Dealing with residual which are not normally distributed.
6. Dealing with unequal error variances and non-linearity of regression function.
7. Identifying outlying observations with respect to Y (response variable) and X's (predictor variables).
8. Identifying and eliminating influential outlying observation.
9. Identifying observation that has a strong influence on the coefficients.
10. Dealing with automated regression calibration.
11. Random Forest which mtry to choose and cp.
12. In presence of qualitative variable not satisfying the test of identity.

Performing various statistical analysis, hypothesis testing and building regression models such as ridge regression, regression tree, robust regression and random forest would help to overcome the challenges and predict the cost.

3. Data Sources

The real-estate dataset is referred from $kaggle$[1], it has 506 instances, 13 continuous attributes and 1 binary-valued attribute. There are some missing attribute values.

Attribute Name	**Description**
CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq. ft.
INDUS	proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centers

RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
B	1000(Bk - 0.63) ^2 where Bk is the proportion of blacks by town
LSTAT	% Lower status of the population
MEDV	Median value of owner-occupied homes in $1000's

4. Proposed Methodology and Analysis

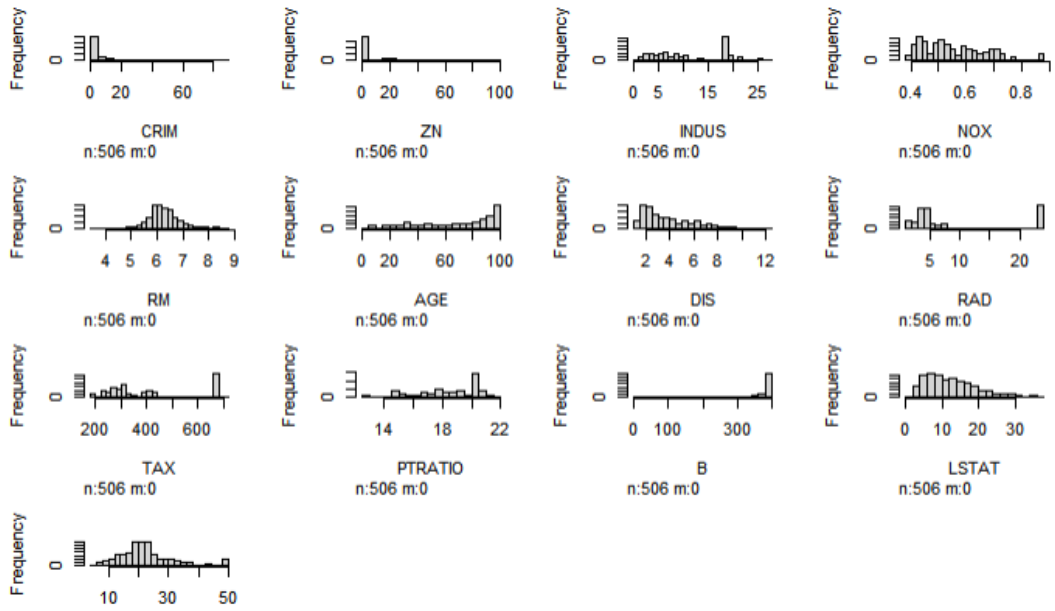
4.1. Data Analysis and Data Pre-processing

1) Predictor variable “RM” has NA values.

   Most common way for treating it is:
   A. Deleting its corresponding row in data frame
   B. Replacing with mean value of that predictor variable. If there exists outlier in data set then mean will not be appropriate.
   C. Replacing with median value of that predictor variable. It is better than mean because it will not get affected by outliers in the data set.

   Hence, we used median to replace NA values in the data set.

2) Analyzing the distribution of data of each predictor variable by plotting respective histogram.



4.2. Statistical analysis and modelling

1. Fit linear regression model with MEDV as response variable and all other variables in data set as predictor variable.
   Residual standard error is 4.75 and Adjusted R-squared is 0.7332
   Here, Residual standard error is the measures of standard deviation of the residuals in a regression model. Adjusted R-squared is a modified R-square formula, the value of it increases

only when newly added predictor variable contributes in explaining variation in dependent variable that is response variable.

2. Eliminating variable with high multi-collinearity.

a) Multicollinearity diagnosis

The second part of data collection and preparation is diagnosing if there are multicollinearity and interaction within the predictor variables. Initially a correlation matrix between the predictor variables is obtained.

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
CRIM	1	-0.198	0.406	-0.055	0.421	-0.22	0.351	-0.377	0.626	0.583	0.277	-0.384	0.405	-0.38
ZN	-0.198	1	-0.534	-0.041	-0.516	0.307	-0.568	0.665	-0.307	-0.312	-0.393	0.176	-0.39	0.34
INDUS	0.406	-0.534	1	0.062	0.764	-0.39	0.643	-0.708	0.592	0.719	0.379	-0.357	0.557	-0.463
CHAS	-0.055	-0.041	0.062	1	0.091	0.09	0.086	-0.098	-0.005	-0.034	-0.124	0.049	-0.057	0.165
NOX	0.421	-0.516	0.764	0.091	1	-0.301	0.729	-0.768	0.609	0.667	0.186	-0.38	0.54	-0.411
RM	-0.22	0.307	-0.39	0.09	-0.301	1	-0.237	0.199	-0.213	-0.293	-0.338	0.128	-0.552	0.667
AGE	0.351	-0.568	0.643	0.086	0.729	-0.237	1	-0.745	0.452	0.503	0.258	-0.272	0.53	-0.368
DIS	-0.377	0.665	-0.708	-0.098	-0.768	0.199	-0.745	1	-0.488	-0.53	-0.238	0.291	-0.467	0.233
RAD	0.626	-0.307	0.592	-0.005	0.609	-0.213	0.452	-0.488	1	0.91	0.439	-0.442	0.422	-0.379
TAX	0.583	-0.312	0.719	-0.034	0.667	-0.293	0.503	-0.53	0.91	1	0.441	-0.441	0.482	-0.459
PTRATIO	0.277	-0.393	0.379	-0.124	0.186	-0.338	0.258	-0.238	0.439	0.441	1	-0.175	0.393	-0.447
B	-0.384	0.176	-0.357	0.049	-0.38	0.128	-0.272	0.291	-0.442	-0.441	-0.175	1	-0.34	0.318
LSTAT	0.405	-0.39	0.557	-0.057	0.54	-0.552	0.53	-0.467	0.422	0.482	0.393	-0.34	1	-0.563
MEDV	-0.38	0.34	-0.463	0.165	-0.411	0.667	-0.368	0.233	-0.379	-0.459	-0.447	0.318	-0.563	1

Instead of tracing the highest multicollinearity among the predictor variables, a formal method for detecting the multicollinearity is used which is Variance Inflation Factor

b) Variance Inflation Factor

Variance inflation factor is the measure of variance that inflates the estimated regression coefficients compared to the predictor variables that are not linearly related, which comes as a result of correlation among the predictor variables.

It is obtained from the variance covariance matrix of the standardized regression coefficient

$r_{xx}{}^{-1}\rightarrow$ correlation among pairwise predictor variables

$$VIF_k = \frac{1}{1 - R_k{}^2}$$

Where $R_k{}^2\rightarrow$ multiple determination coefficient among the n-p predictor variables
When $R_k{}^2$ is zero VIF=1 which means X_k is not linearly associated with any predictor variables.
Increase in $R_k{}^2$ increases the VIF value.

In our dataset the VIF values for the predictor variables are

CRIM	ZN	INDUS	CHAS	NOX	RM
1.780779	2.296466	3.984262	1.073206	4.383370	1.652380

AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT

2.824032	3.94065	7.395601	9.005566	1.754597	1.342209	2.19816

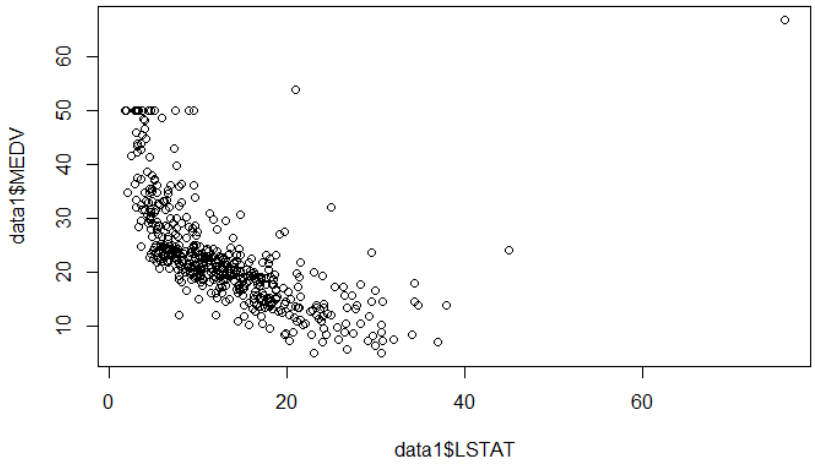
From the VIF table it is evident that the predictor variable TAX has severity of multicollinearity with a high VIF value of 9.005566.(influencing the least square estimates)

Thus, we remove the predictor variable TAX from the dataset, without any significant change in Adjusted R-square and residual error

Dataset	Adjusted R-square	Residual error
Before removing var TAX	0.7332	4.75
After removing var TAX	0.7279	4.798

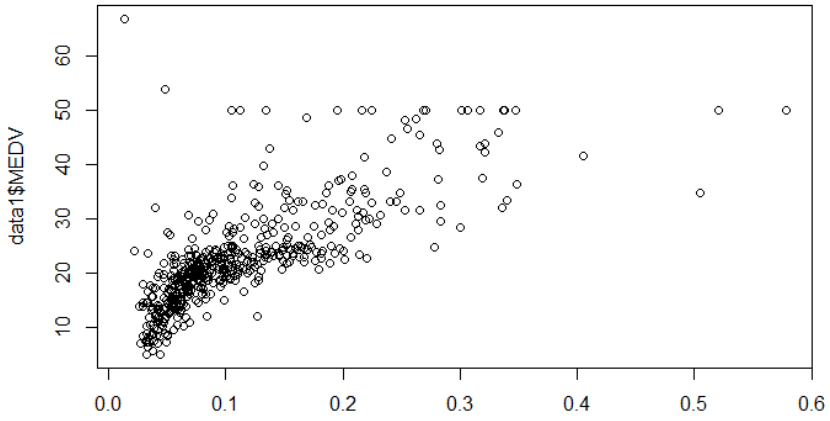
3. Linearity check between predictor variables and response variable.

Individually plotting the predictor variables with the output, it is found that the predictor variable LSTAT has a curvilinear relationship with the response variable.



To fix the collinearity, a transformation is done to the predictor variable LSTAT such that it is linear to the response variable MEDV.

After doing an inverse transformation to LSTAT i.e., substituting the predictor variable LSTAT as $\frac{1}{LSTAT}$

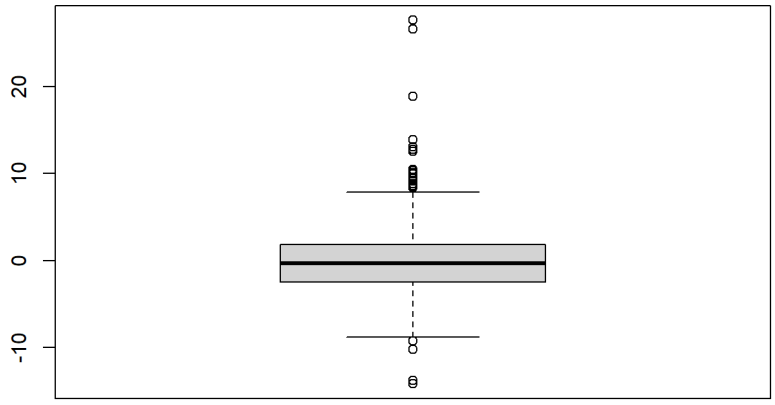


After transformation of LSTAT

Dataset	Adjusted R-square	Residual error
Before transforming LSTAT	0.7279	4.798
After transforming LSTAT	0.7771	4.342

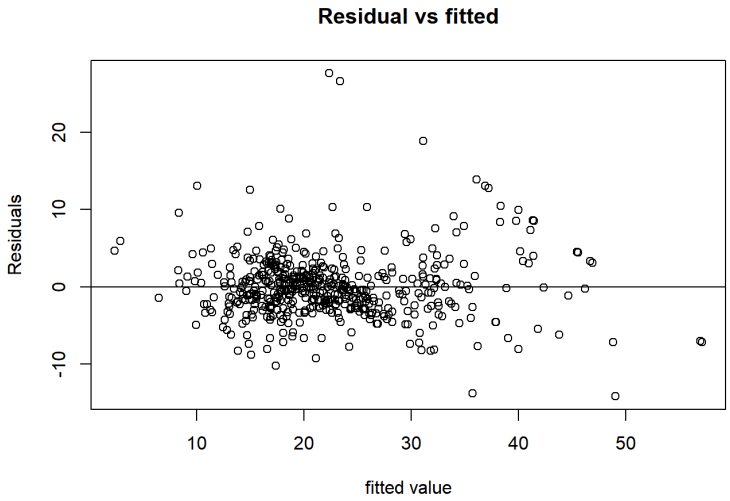
4. Analyzing residuals distribution using box plot.

   We observe there are lot of outliers beyond minimum and maximum points of box plot.



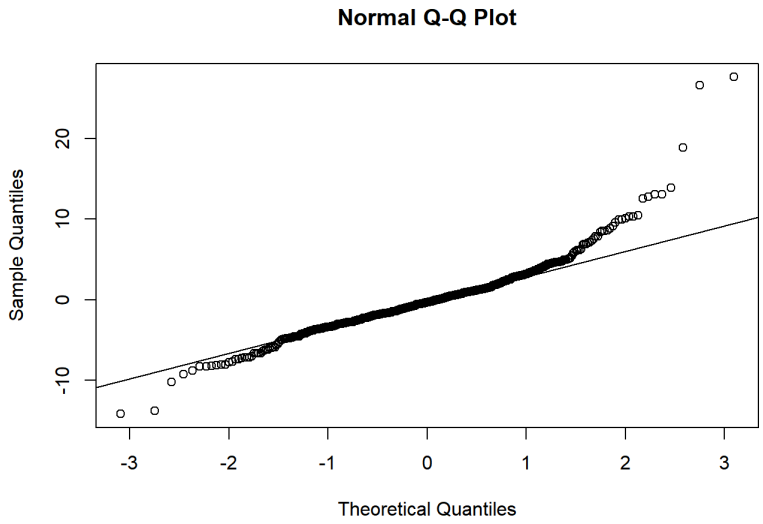
5. Plotting residuals vs fitted values graph to check if residuals follow homoscedasticity.

   Here analyzing the graph, we observe small curvature and outliers towards higher end of fitted values.



6. Distribution of Residuals.

   We can see in the below Normal QQ plot that residuals are deviating from the normal distribution



7. Test to check whether there is a regression relation.

   Using $\alpha = .05$
   H0: $\beta 1 = \beta 2 = \cdots = \beta p-1 = 0$
   Ha: not all $\beta i = 0$, $i = 1, \ldots, p - 1$

   F*=MSR/MSE
   F*=3559.6916/18.8503=188.8400

   F (0.95; 4, 501) = 2.389731

   If $F* \leq F\ (1 - \alpha; p - 1, n - p)$ conclude H0
   If $F* > F\ (1 - \alpha; p - 1, n - p)$ conclude Ha

   Since F*>f we can conclude Ha.

8. Perform the check for the Constancy of error Variance.

   We have used the Breusch Pagan test to check constancy of the error variance.
   Assuming log σi^2 = γ0 + γ1Xi1 +.... + γn Xin; use α = .01

   The alternatives are
   H0: γ1 = γ2 = 0
   Ha: γ1 != 0 or γ2 != 0 or ..... γn != 0

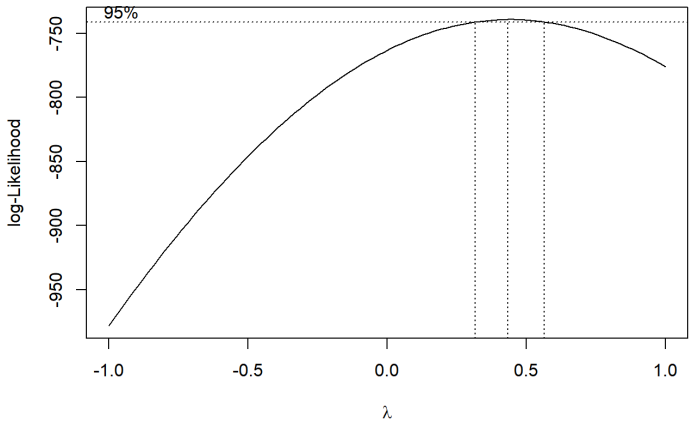
   Using bptest function of library lmtest.
   BP = 295.68, df = 12, p-value < 2.2e-16

   Since P value is less than 0.05, we reject Null hypothesis. So there exists non-Constance in error variance.

9. Dealing with unequal variance and non-linearity.

   Box Cox find which transformation on Y will be appropriate to correct skewness of the distributions of error terms, unequal error variances, and non-linearity of the regression function. Box cox automatically finds a transformation from family of power transformations on Y such that SSE reduces.

   Used Boxcox function of MASS library to check best transformation of values starting from -1 to 1 with step size 0.1.



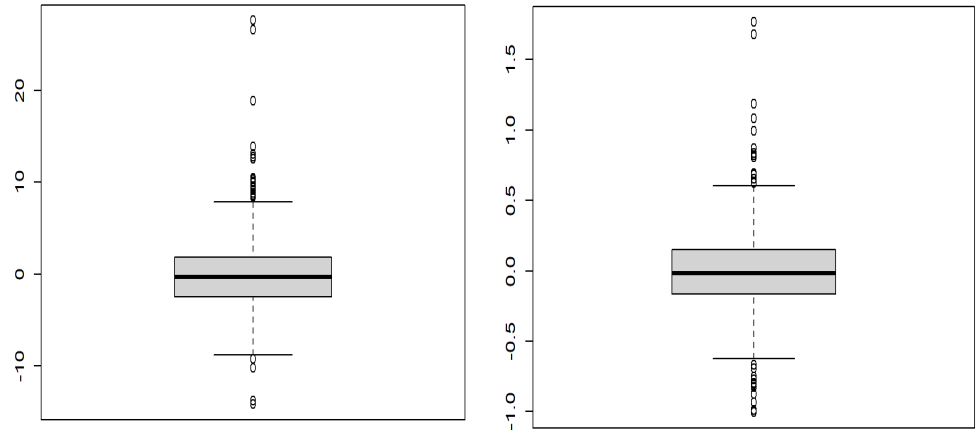
   Hence, Transformation of response variable using lambda=0.4343434 is optimum with minimum SSE.

   Fit linear regression model on data set with aforementioned transformation applied.
   Therefore, after applying Box Cox transformation Adjusted R square marginally decreased from 0.7771 to 0.7743. However, residual standard error decreased from 4.342 to 0.315.

***Box plot of residuals before and after Box cox transformation*.**

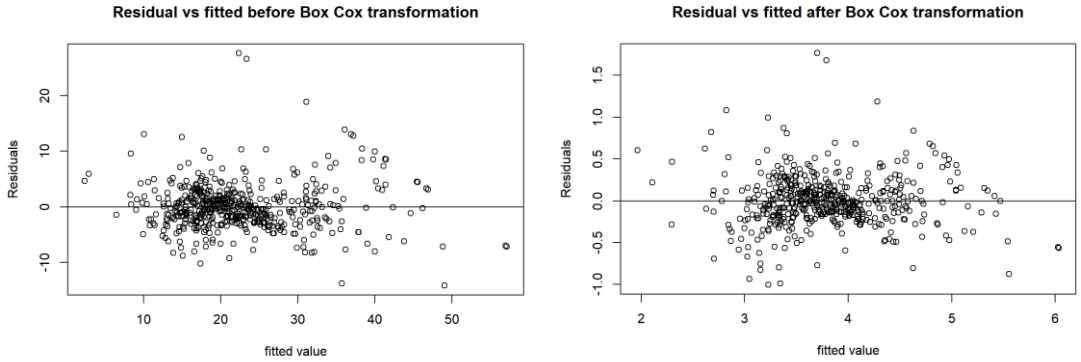
We can see that residuals out liers patterns have been distributed on either side of minimum and maximum values.

Before Box cox transformation After Box cox transformation

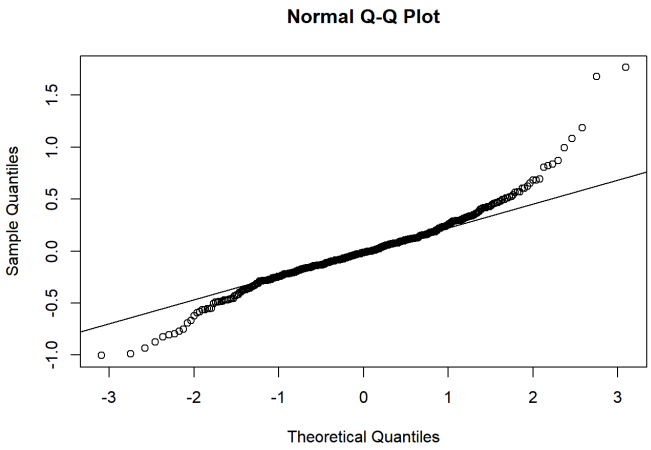


Residuals vs fitted values before and after Box Cox transformation.

We observe the change in distribution of residuals around zero after Box Cox transformation is applied.



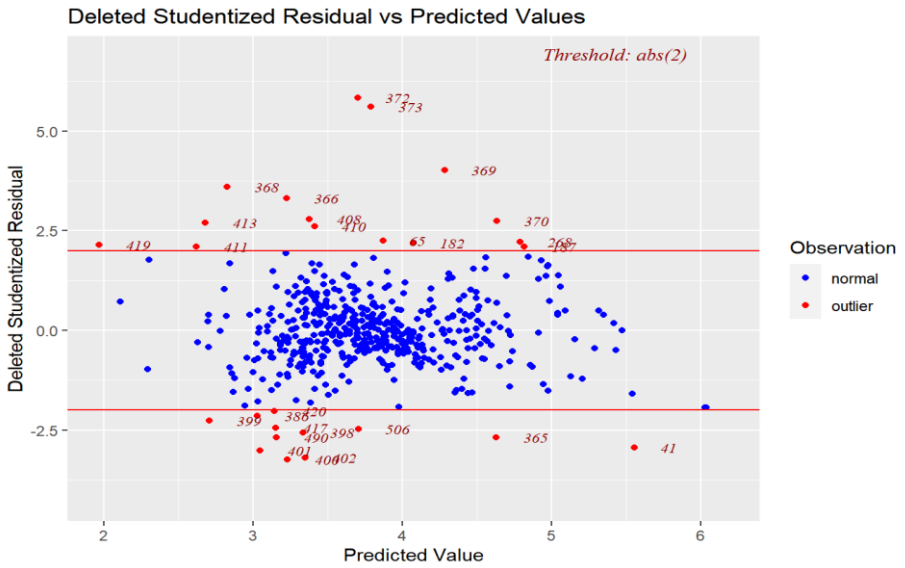
Analyzing QQ plot: We observe that still the residuals are not following normal distribution due to outliers, as observed on top and bottom corner of the QQ plot.



10. Identifying outlying observations:

a) *Identifying outlying Y observations based on Studentized deleted residuals*

Used “ols_plot_resid_stud_fit” function from “olsrr” library.

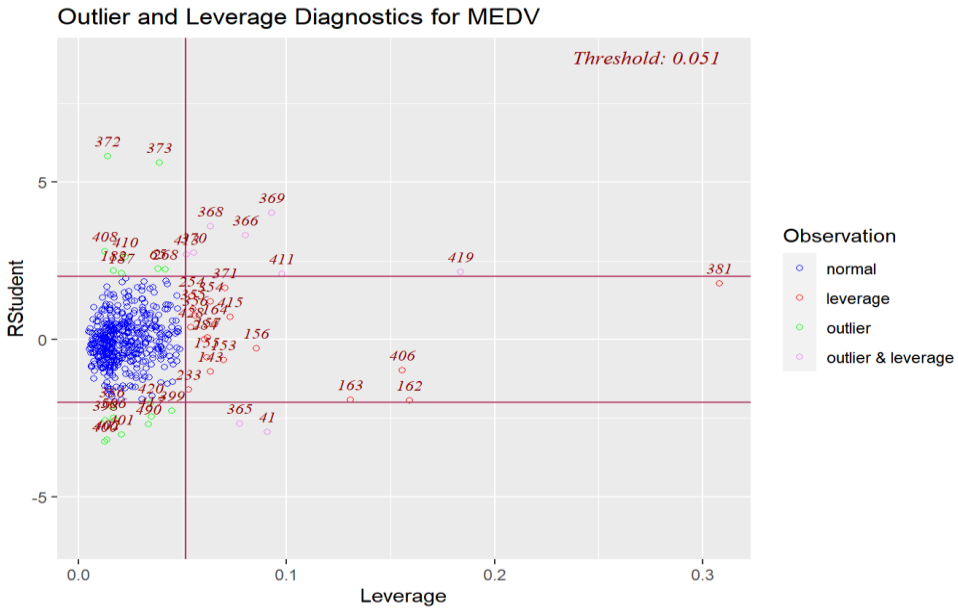


Threshold = t distribution (1-0.05/2*n; n - p - 1)

We conduct a formal test using Bonferroni test procedure of whether the case with the largest absolute studentized deleted residual is an outlier. Using the aforementioned test we were able to find 4 observations where outliers.

b) *<u>Identifying outlying X observations based on HAT matrix Leverage values:</u>*

Using “ols_plot_resid_lev” function from “olsrr” library.



By a rule that any leverage values greater than 2p/n are considered as outlying observations with respect to X values. From this formula we found leverage threshold as 0.05533597. Using the aforementioned test, we were able to find 23 observations were outliers.

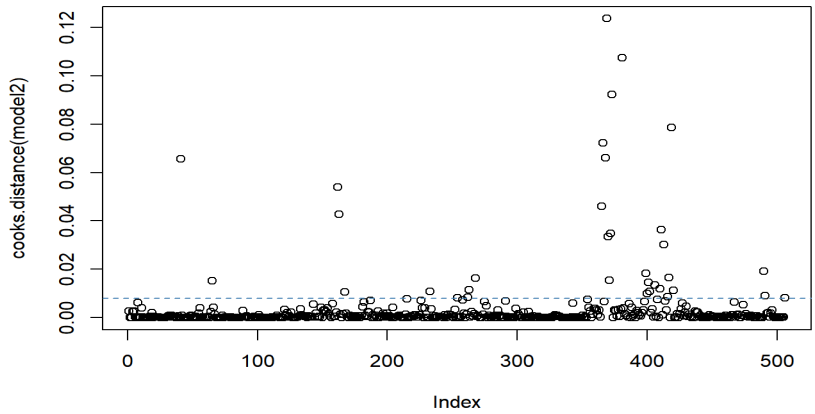
11. Eliminating influential outliers.

a) Cook's distance:

- After identifying cases that are outlying with respect to their Y values and or with their X values, the next step is to find whether or not these outlying cases are influential. We shall consider a case to be influential if its exclusion causes major changes in the fitted regression function

- Cook's distance measure considers the influence of the i^{th}case on all n fitted values. Cook's distance measure, denoted by Di, is an aggregate influence measure, showing the effect of the i^{th}case on all n fitted values

- Cook's distance greater than 4/n, where n is total number of observations in data set. Then it is considered as outlier.



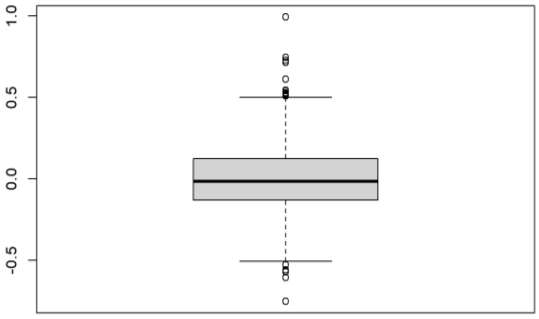
- Now, we remove all outliers based on Cook's distance threshold as said above. Now we fit the model using the data set with outlying observations removed.

- Adjusted R square improved from 0.7743 to 0.8469.
  Residual standard error decreased from 0.315 to 0.2283.

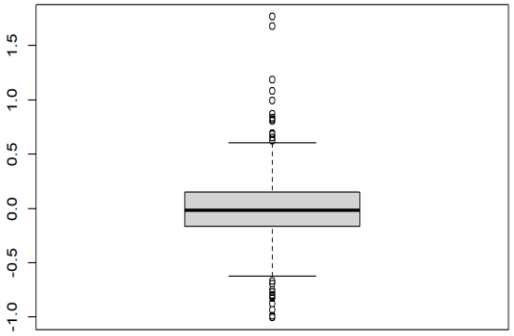
Analyzing following mentioned plot before and after removing outliers, plots are residuals vs fitted, residuals box plot and QQ plot

- In residuals vs fitted plot that outliers are not present now.
- In box plot we can see number of outliers has reduced as we have removed influential outliers.
- In QQ plot we can see that before points were deviating from the tails but now it's less deviating.

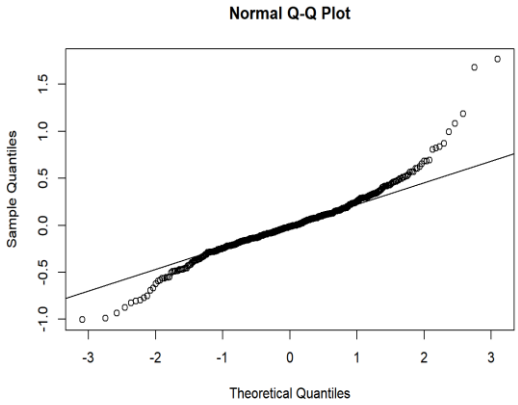
Before Transformation



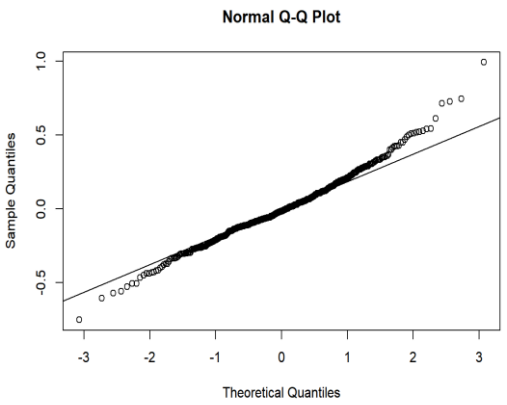
After Transformation



Before Removal of Outlier

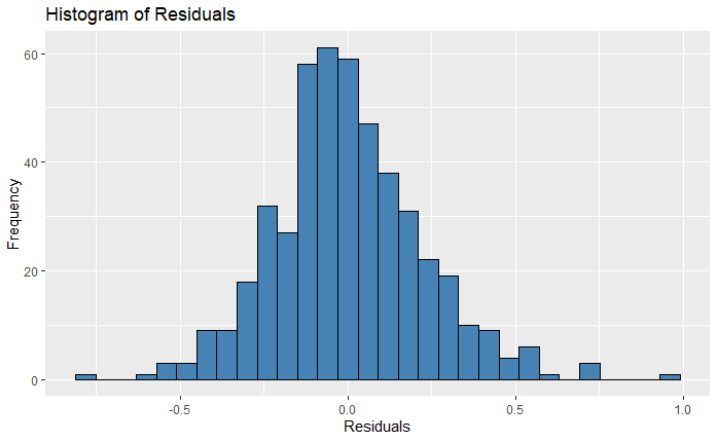


After Removal of Outlier



Histogram of residuals

After eliminating the influential outliers from the data and applying suitable transformation for the predictor and response variable, it is necessary to check the nature of the distribution of residuals. To examine this, histogram of residuals plot is used to find the nature of the distribution. From the transformed model and the updated data set the histogram of residuals comes as



From the plot it is evident that the error in the data is normally distributed. This inference helps to explore other regression techniques to predict the outcomes accurately.

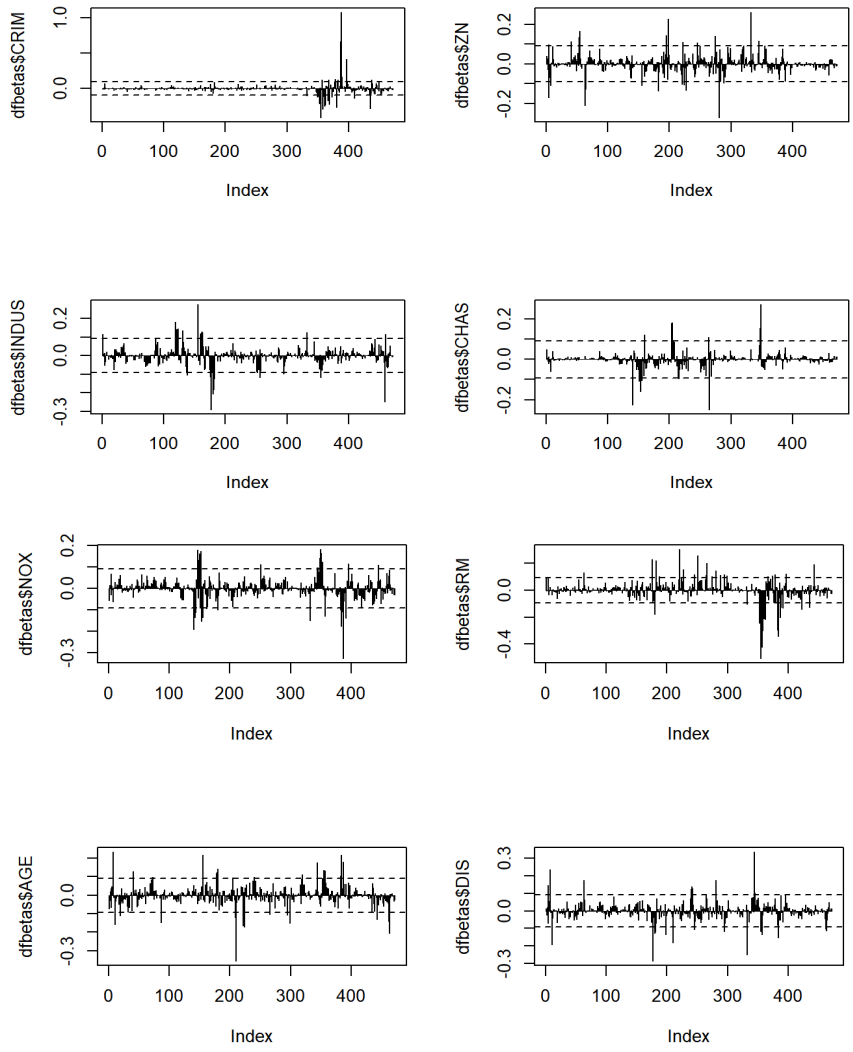
b) DFBETAS

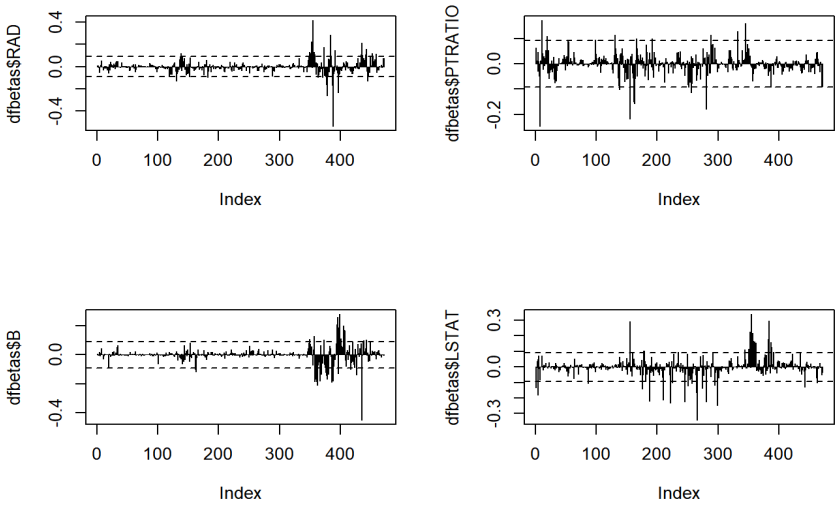
DEBETAS is one of the ways of calculating the influence of observations; it tells us about the standardized effect on each coefficient of individual observation that is being deleted. In a given regression model, this metric gives an idea on how influential each observation is on each coefficient estimate.

As a guideline for identifying influential cases, we consider a case influential if the absolute value of DFBETAS exceeds 1 for small to medium data sets and 2/ $\sqrt{n}$ for large data sets.

DFBETAS analysis on data where outliers are removed using COOK's distance.

Thresh = 2/sqrt(nrow(trans_data1)) = 0.09205746





Predictor variables	Number of influential observations based on DFBETAS
CRIM	12
ZN	13
INDUS	13
CHAS	7
NOX	13
RM	23
AGE	17
DIS	13
RAD	20
PTRATIO	12
B	13
LSTAT	20

12. Analysis of Qualitative variable

In the data set there is a qualitative variable, CHAS. It is recorded as 1 if the houses are in the track bounds of the Charles River and 0 otherwise.

Number of records in CHAS=0→ 442 (near river)
Number of records in CHAS=1→30 (away river)

(After eliminating all influential outliers)

To check if the two populations can be fitted in one model, it has to satisfy the test of identity and the predictor variable CHAS has to be significant under t test.

T test

t*=bk/s{bk}
t*(CHAS)= 2.396007001
t (0.95,472) =1.648088
Here t*>t
Additionally, the p value of the test of identity is found to be 0.3356.

- From the above T test and identity test it is inferred that, there is a necessity to split the qualitative variable into 2 populations and carry out statistical analysis individually.

- For the near river with 30 records, the Adjusted R-square is 0.9668 and the Residual standard error 0.1115. Whereas for the away river population Adjusted R-square is .8385 and the Residual standard error 0.2324.

- Thus, it is sufficient to carry out statistical analysis and apply additional modeling techniques for the 'away river' population whose CHAS value is 0.

- From this part of the project, we have taken the away river population as our interest and carried out further statistical modeling.

13. Remedy for the problem of multicollinearity

High multicollinearity among the predictor variables results in an unreliable and high variance.

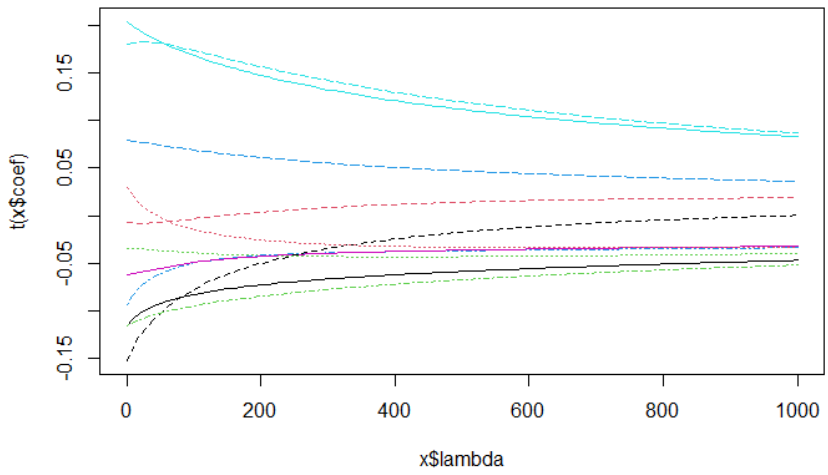
A remedy to overcome multicollinearity without eliminating any predictor variable is done by ridge regression. This is achieved by allowing biased estimators through modifying the method of least squares.

A shrinkage penalty term is introduced along with the regression coefficient

RSS + λΣβj2

- When λ=0 there is no penalty to the regression coefficient, as λ approaches infinity the ridge regression estimate goes to zero.

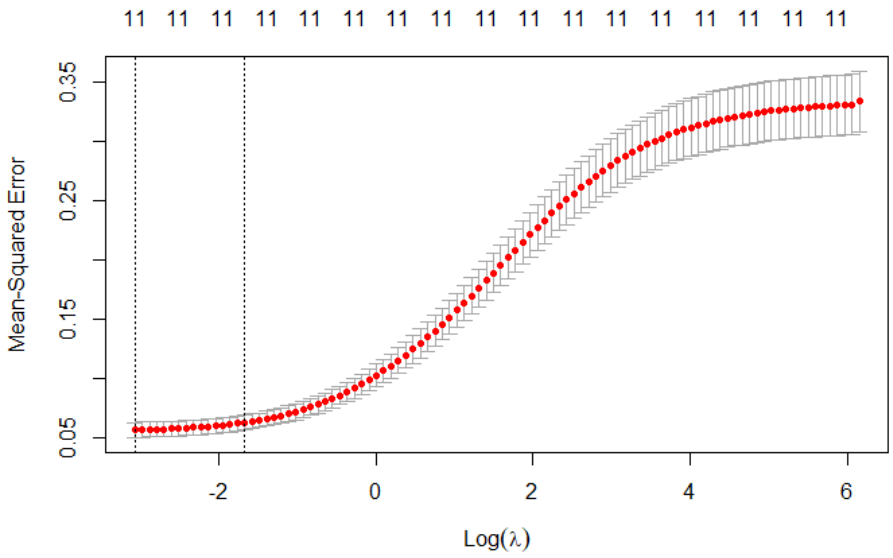
- Since a ridge regression is all about introducing bias in the model, we choose λ that produces the lowest MSE.



- Proceeding with the 'away river' dataset, we estimated the optimal lambda value minimizing the MSE which turns out to be 0.04676432

- It is clear that in below graph there is no significant multicollinearity exists in the 'away river' dataset.

- Adjusted R-square→ 0.8395315



14. Regression tree

When the relationship between the predictor variables and response variable is not linear, regression helps in partitioning the predictor variable's space into rectangular regions and estimates a regression surface. (cart)

Initially the regression tree is built by passing the model into 'rpart' function.

Regression tree output is as shown below:

1) root 442 147.4317000 3.767866

2) LSTAT< 0.1029343 259 41.1577700 3.440425

4) LSTAT< 0.06607237 135 17.1920100 3.194393

8) CRIM>=5.76921 65 5.4073890 2.959399 *

9) CRIM< 5.76921 70 4.8621670 3.412601 *

5) LSTAT>=0.06607237 124 6.8972510 3.708282 *

3) LSTAT>=0.1029343 183 39.2023300 4.231295

6) RM< 6.941 131 8.3941870 4.008055

12) RM< 6.543 85 3.0865300 3.893622 *

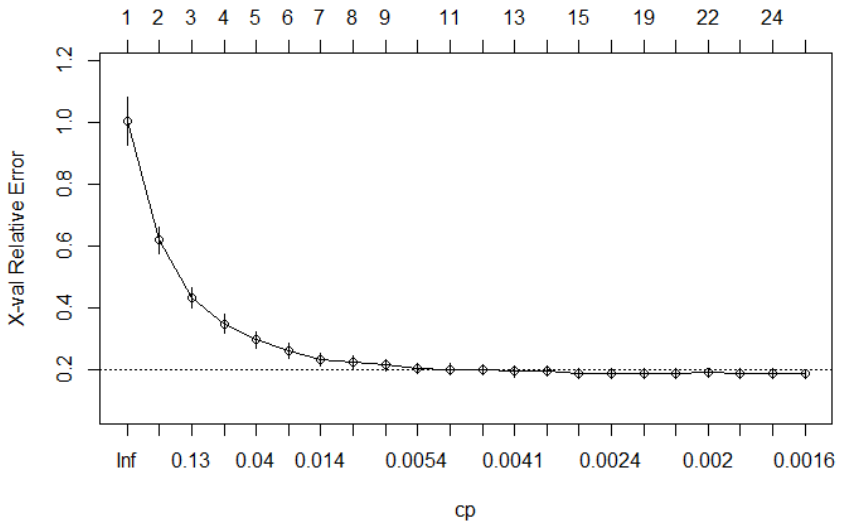
13) RM>=6.543 46 2.1378380 4.219507 *

7) RM>=6.941 52 7.8327270 4.793688

14) RM< 7.437 35 1.9795890 4.574837 *

15) RM>=7.437 17 0.7254568 5.244264 *

Then an optimal value of the cp (complexity parameter) is found by fixing the lowest test error.

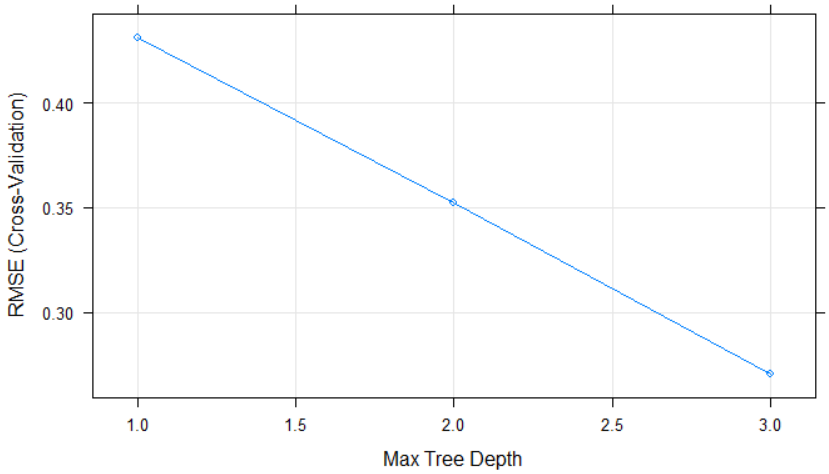


From the graph it is found that the optimal cp value is 0.00157294.

Then a regression tree is built using the 'caret' package.

To find the efficiency of the model, the data is splitted into 80 and 20 percent as training and testing data.

Then the regression tree is modeled with the caret package.



After the modeling using the training data, the model is allowed to predict the inputs from the test data and the residuals are found.

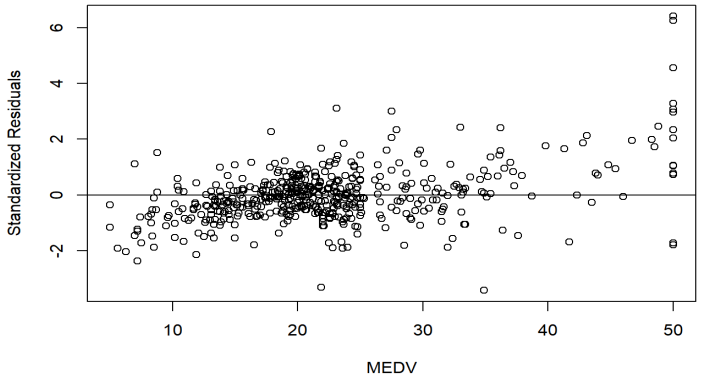
Summary of Regression tree:

RMSE	Adjusted R-square
0.25767	0.812813

15. Robust regression

When there is no time for a thorough identification of outlying cases and an analysis of their influence, nor for careful consideration of remedial measures. Instead, an automated regression calibration must be used. Robust regression procedures will automatically guard against the undue influence of outlying cases in this situation.

Considering the initial original raw data set. From the plot we can see that there are observations with standardized residuals as outliers.



Using "rlm" function of "MASS" library.

In Robust regression, the weight goes up as the absolute residual goes down. We can also say that the cases with large residuals tend to be down-weighted. Therefore, in the robust regression, if more cases have a weight close to 1, OLS and robust regressions results would be closer.

After fitting ordinary linear regression model and Robust regression model we can observe that:

Data set used below- Predictor variable "TAX" removed based on VIF and predictor variable LSTAT rectified by applying transformation to correct curvilinear with respect to response variable.

- From Ordinary least square regression the Adjusted R square is 0.7743 and Residual standard error is 0.315.
- From Robust regression model using Huber weights the Adjusted R square is 0.8542 and Residual standard error is 0.2243.

Hence, Residual standard error is reduced and Adjusted R square is increased

Applying Robust Regression on previous mentioned data set on which response variable is transformed using Box Cox method.

Applying Robust Regression on data set of which response variable is transformed using Box Cox method and outlying observations removed using Cook's distance.

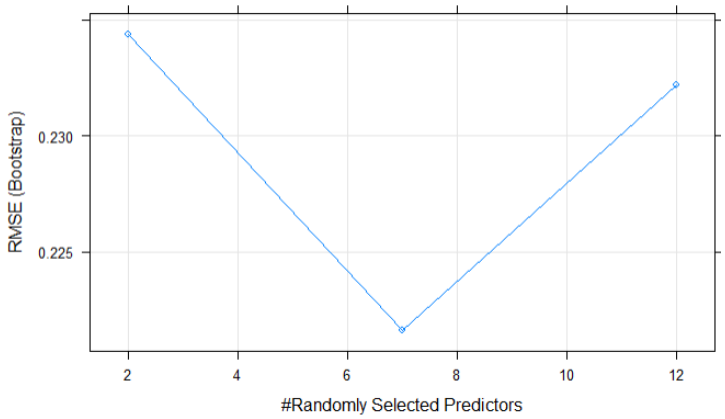
- From Ordinary least square regression the Adjusted R square is 0.8469 and Residual standard error is 0.2283.
- From Robust regression model using Huber weights the Adjusted R square is 0.881 and Residual standard error is 0.1888.

Hence, Residual standard error is reduced and Adjusted R square is increased.

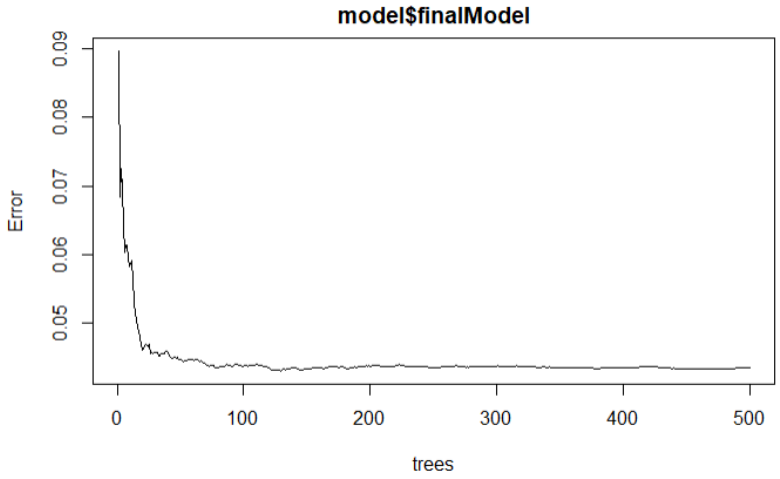
16. Random Forest

Random Forest is collection of multiple decision trees. A decision tree consists of decision nodes and leaf nodes. Each question creates a decision node in the tree which would split the data. When each node makes a decision, the result would reside in leaf node. Decision trees are trained through an algorithm CART (Classification and Regression Tree) which would try giving the best result. There are problems like overfitting with respect to decision tree which can be solved using random forest. The accuracy in random forest is much better than decision tree when the individual trees are uncorrelated.

- In random forest to create decision trees there is an addition of randomness required which is controlled by mtry any given point in time.
- The mtry captures the number of input feature attributes and is denoted by whole number which falls in the range 1 to total number of feature attributes.
- Here mtry at each split considers the number of available features. As shown in the below graph the final value used in the model is mtry = 7.



- We can observe that the error is not reducing further for the number of trees approximately beyond 150. Thus, we can construct random forest using approximately 150 trees to reduce complexity of the model.



5. Conclusions

In this project, we dealt with issues such as: missing values by replacing them with median values; eliminating multicollinearity by identifying it with VIF; rectifying curvilinear predictor variables by applying an appropriate transformation to the predictor variable; performing residual analysis by plotting different residual plots and the QQ plot. Performed specific hypothesis testing for regression relationships by using the F test; checking the constancy of the error variance by conducting the Breusch Pagan test. We also corrected the skewness of the distributions of error terms, unequal error variances, and non-linearity of the regression function by implementing Box Cox transformation. The outlying Y observations are delt by using Studentized deleted residuals. The outlying X observations are identified using HAT matrix Leverage values and eliminated using techniques Cook's distance and DFBETAS. Then this project furthers to analyse the qualitative variables. Robust Regression is used to automatically handle the outlying cases. In the process of implementing the aforementioned strategy, we show the proof of how the linear regression model is improving as we find remedies for all the statistical challenges. We then build Regression Tree and Random Forest model on the dataset in interest to improving the efficiency. As far as regression modelling is concerned Random Forest is applied on the dataset after eliminating the influential outliers and suitable response variable transformation is observed to be effective.

6. Bibliography and Credits

[1] https://www.kaggle.com/datasets/arslanali4343/real-estate-dataset?resource=download

[2] K. Ayinde, A. F. Lukman, and O. Arowolo, "Robust Regression Diagnostics of Influential Observations in Linear Regression Model," Open Journal of Statistics, vol. 05, no. 04, pp. 273–283, 2015, doi: 10.4236/ojs.2015.54029.

[3] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, Applied linear statistical models. New Delhi: Mcgraw-Hill Education (India) Private Limited, 2013.

[4] S. Sanyal, S. Kumar Biswas, D. Das, M. Chakraborty, and B. Purkayastha, "Boston House Price Prediction Using Regression Models," IEEE Xplore, Jun. 01, 2022.

[5] https://rdocumentation.org/

[6] https://www.statology.org/

[7] https://stackoverflow.com/

[8] Dr. Lulu Kang. "Applied Statistics." *Lecture Notes, Illinois Institute of Technology* (2022).