Indian Statistical Institute, New Delhi

Masters of Statistics

HOUSE PRICE PREDICTION

Data Analysis Project Work

Name: SUBHASIS SUR

Roll Number: MD2223

Supervisor : Prof. Deepayan Sarkar

Abstract

House is one of the most necessities in our humuan life.In this project our main focus is to build a prediction model based on some relevant factors. Among these factors some are continuous and discrete predictor and rest of them are categorical predictors.Also by this model we can get some insights about some relationships between the response variable i.e. price of house and the covariates such sqft lot size, location, no. of bedrooms etc.

Contents

Chapter 1

Introduction

In this Project our main goal is to build a prediction model for house price predictions.We divide whole data set in two parts train and test. Fit multiple linear model .robust linear model with loss function bisquare and huber and by this we use it on test data and find correlation between actual and predicted. Again we use lasso to select variable form all of the variable and also fit a linear model with this variable and find predictive R square, multiple R square.Also apply this model on the test dataset and find correlation between actual and predicted.At last we compare between all of them.

Chapter 2

Description of the Dataset

Name : Predicting House Price

Source : Kaggle

Link of the Dataset : https://www.kaggle.com/code/burhanykiyakoglu/predicting-house-prices

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015. Here we have total 21613 rows (units of data) and 17 columns

1.Price: Price of the house (**Response variable**)

2.id : a notation for a house

3.date: Date house was sold

4.bedrooms: Number of Bedrooms

5.bathrooms: Number of bathrooms

6.sqft living: square footage of the house

7.sqft lot: square footage of the lot

8.floors :Total floors (levels) in house

9.waterfront :House which has a view to a waterfront

10.view: Has been viewed by the buyers

11.condition :How good the condition is Overall

12.grade: overall grade given to the housing unit, based on King County grading system

13.sqft above :square footage of house apart from basement

14.sqft basement: square footage of the basement

15.yr built :built Year

16.yr renovated :Year when house was renovated

17.zipcode:zip code

18.lat: Latitude coordinate

19.long: Longitude coordinate

20.sqft living15 :Living room area in 2015(implies– some renovations).
This might or mightnot have affected the lotsize area

21.sqft lot15 :lotSize area in 2015(implies– some renovations)

Chapter 3

Data Cleaning And Modification

At first we import the data file in R.
Then I remove the id column from dataset because logically id of a house doesn't give any idea about the the house price.
I have 3 columns about the latitude, longitude and zip code of a house. Again in logical sense latitude and longitude don't give any idea about the house price. It may happen at some latitude and longitude it gives the price is very high but there doesn't exist any land.
The values of bathrooms and also the floors are fractional which also not seems to be very rational so we round this columns.
In the zipcode all the rows are have almost unique columns so we take first 4 digits of the zipcode and transform them into the categorical variable having 19 levels.
As grade ,condition and view are defined it is very reasonable to take this as categorical variables.
Now from the definition of sqft living ,sqft basement and sqft above we get
sqft living= sqft above + sqft basement
and by 3d plot we get this.So we remove this sqft living from the dataset.

In the yr renovated column , if the house was renovated then it year of renovation is given otherwise it was 0 .So, we change the column the into binary where
1 denotes **renovated** or 0 denotes **not renovated**
and also change the column name to renovated.
Waterfront is also binary so we change into factors

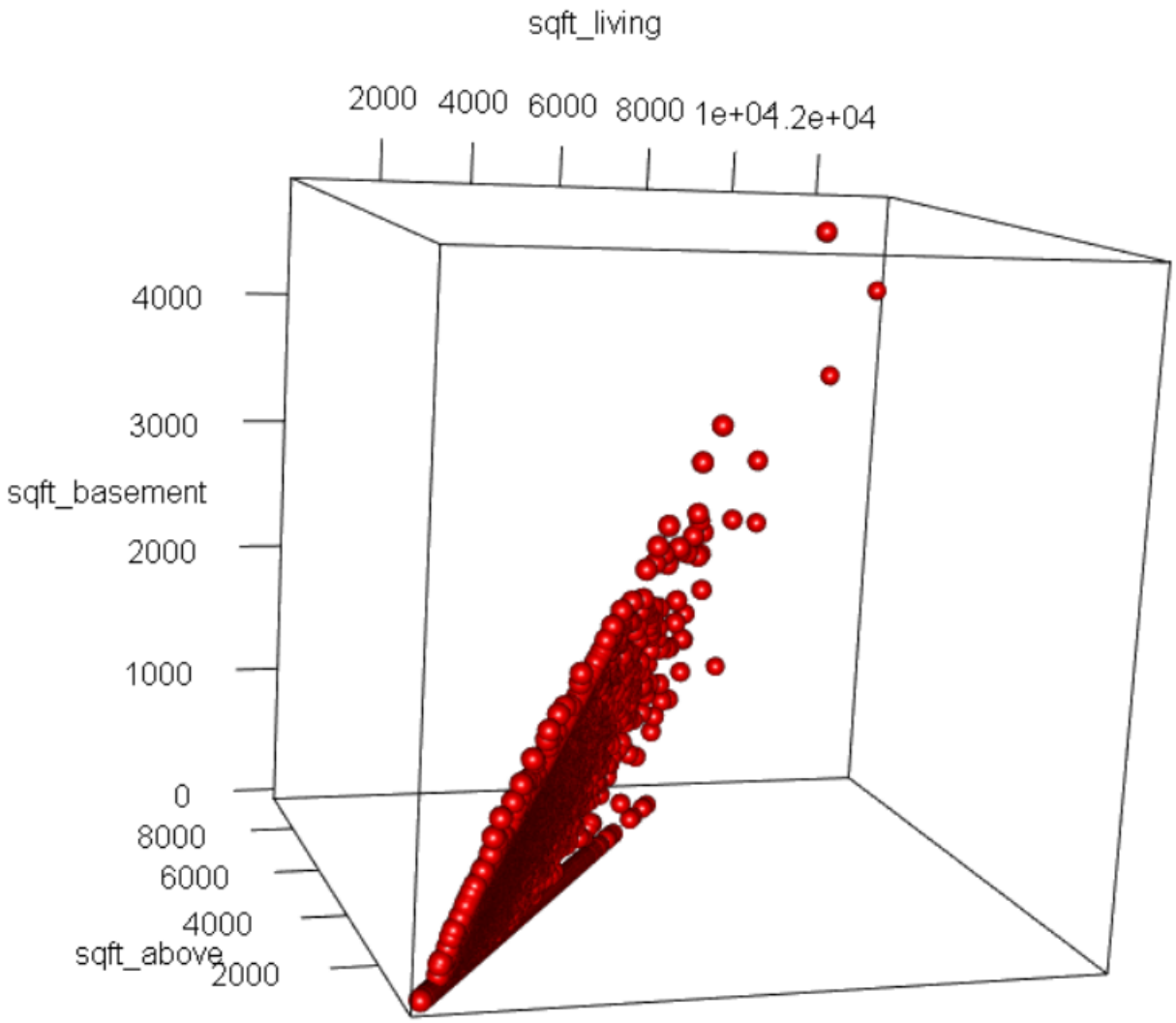

Figure 3.1: 3d plot between sqft of living, basement and above.

```
'data.frame':   21613 obs. of  18 variables:
 $ date         : Factor w/ 2 levels "2014","2015": 1 1 2 1 2 1 1 2 2 2 ...
 $ price        : num  221900 538000 180000 604000 510000 ...
 $ bedrooms     : int  3 3 2 4 3 4 3 3 3 3 ...
 $ bathrooms    : num  1 2 1 3 2 4 2 2 1 2 ...
 $ sqft_living  : int  1180 2570 770 1960 1680 5420 1715 1060 1780 1890 ...
 $ sqft_lot     : int  5650 7242 10000 5000 8080 101930 6819 9711 7470 6560 ...
 $ floors       : num  1 2 1 1 1 1 2 1 1 2 ...
 $ waterfront   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ view         : Factor w/ 5 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ condition    : Factor w/ 5 levels "1","2","3","4",..: 3 3 3 5 3 3 3 3 3 3 ...
 $ grade        : Factor w/ 12 levels "1","3","4","5",..: 6 6 5 6 7 10 6 6 6 6 ...
 $ sqft_above   : int  1180 2170 770 1050 1680 3890 1715 1060 1050 1890 ...
 $ sqft_basement: int  0 400 0 910 0 1530 0 0 730 0 ...
 $ yr_built     : int  1955 1951 1933 1965 1987 2001 1995 1963 1960 2003 ...
 $ zipcode      : Factor w/ 19 levels "9800","9801",..: 17 12 3 13 8 6 1 19 14 4 ...
 $ sqft_living15: int  1340 1690 2720 1360 1800 4760 2238 1650 1780 2390 ...
 $ sqft_lot15   : int  5650 7639 8062 5000 7503 101930 6819 9711 8113 7570 ...
 $ renovated    : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 1 ...
```

Figure 3.2: Data Frame

Chapter 4

Data Analysis

4.1 Relationship Between Response and Covariates

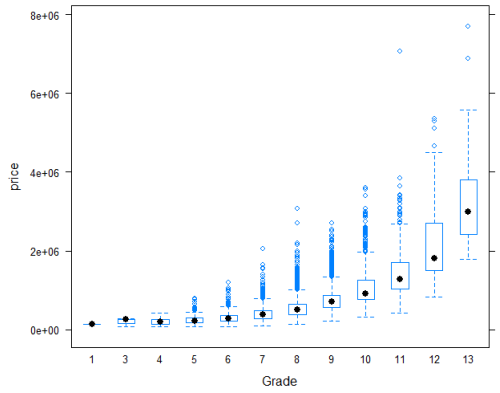

Figure 4.1: House Price vs Grade.

1. Price vs Grade : From Boxplot 4.1 given above we can see that there increasing relationship between the Price and Grade of the house which is also seem to very obvious since Grade is the overall grade of housing unit.

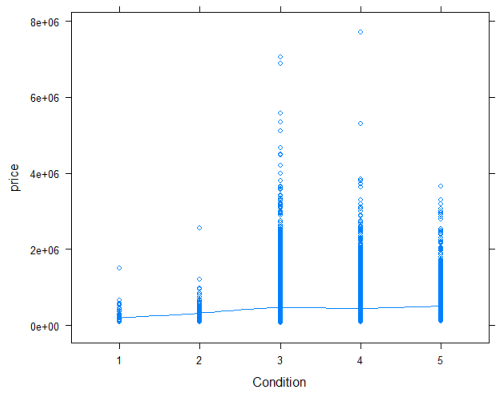

Figure 4.2: House Price vs Condition.

2. Price vs Condition : From the plot4.2 we get that there is little variation in price with the change in the condition of the house.

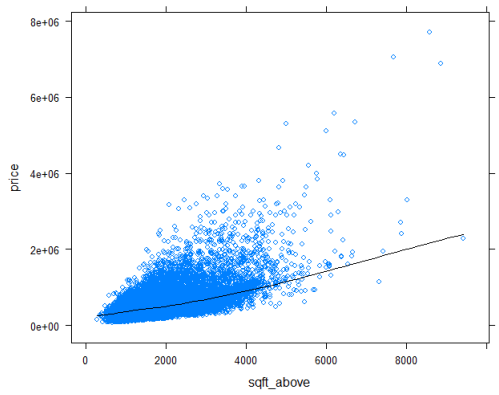

Figure 4.3: House Price vs sqft above.

3. Price vs Sqft Above : By this plot 4.3 we get that with the change in sqft above house price increases.

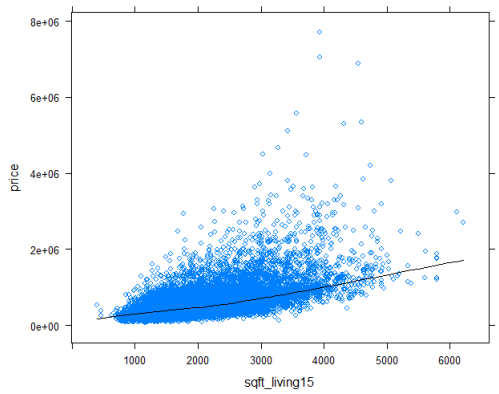

Figure 4.4: House Price vs sqft living15.

4. Price vs Sqft living15 : By this plot 4.4 we get also a increasing trend between price sqft living15.

4.2 Fitting of Multiple Linear Regression

Train and Testing Dataset

Here we at first split the two dataset in two disjoint parts as **train dataset**containing 0.8 proportion of total data and **test dataset**.Now we fit bulid a model based on the train data and then apply that model to the test data to know about model accuracy.We dumify all categorical variables.

Multiple Linear Regression

Here we fit a Multiple Regression model,say **fm1** to the training dataset.The output of the table are given value

coefficients	Estimate	$Pr(>\|t\|)$
(Intercept)	3478828	5.44E-42
date2015	28343.98	2.66E-18
bedrooms	-20696.6	2.32E-22
bathrooms	35465	4.39E-31
sqftlot	0.104685	0.054069
floors	692.6057	0.855512
waterfront1	545854.4	1.20E-134
view1	91316.26	2.63E-13
view2	36207.31	1.67E-06
view3	96279.74	2.86E-20
view4	226959.7	3.96E-46
condition2	47000.24	0.305926
condition3	81444.13	0.057765
condition4	108179.7	0.011746
condition5	151074.2	0.000465
grade10	203495	0.318135
grade11	439257.4	0.031464
grade12	887659.8	1.56E-05
grade13	2390484	7.04E-29
grade3	-73468.3	0.76653
grade4	-183314	0.375873
grade5	-214757	0.291686
grade6	-195262	0.337419

grade7	-144887	0.47662
grade8	-70708	0.72838
grade9	44534.82	0.826935
sqft above	146.3531	1.10E-260
sqft basement	130.5347	2.30E-160
yr built	-1705.84	3.44E-95
zipcode9801	-90253.9	4.30E-18
zipcode9802	-117054	1.25E-60
zipcode9803	-49024.4	2.05E-13
zipcode9804	-59260.5	1.59E-12
zipcode9805	-91410.1	4.33E-44
zipcode9806	-123862	1.06E-18
zipcode9807	-98123.3	9.11E-36
zipcode9809	-231029	1.52E-70
zipcode9810	83459.08	6.17E-27
zipcode9811	89220.21	5.12E-35
zipcode9812	19479.28	0.026107
zipcode9813	-15096.6	0.113127
zipcode9814	-32641.6	0.00128
zipcode9815	-54553.1	4.30E-06
zipcode9816	-139362	3.71E-36
zipcode9817	-93931.6	2.21E-17
zipcode9818	-153622	5.47E-15
zipcode9819	-22497.2	0.029828
sqft living15	55.71212	4.66E-46
sqft lot15	-0.29673	0.000347
reovated1	59489.46	4.32E-13

Here from the above table ,if we see the p value then all the variables are seems to significant.For a categorical variable some factors are seem to be insignificant but others are significant so at all, the categorical variable is significant.

Here,the **multiple R square** is **0.7108**

4.3 Influential Observations and Outliers

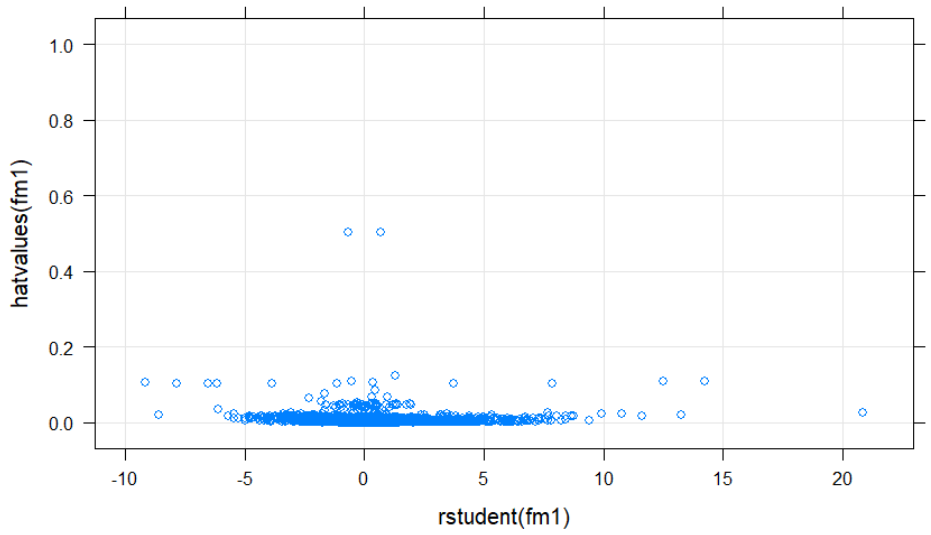

Figure 4.5: rstudent vs Hatvalue.

4.3.1 rstudent vs hatvalue

Here we plot the rstudent and hatvalues of the first model fm1 together to find the influential observations. From the plot 4.5 ,we take cutoff of hatvalue as 0.4 and the cutoff of rstudent is greater than 10.Then I remove those points which are greater than the cutoff and remove them from my dataset and fit the model again. The new **multiple R square** is **0.7072**.Since the multiple R square varies little bit so we keep those points in the data set as each observation carries significant amount of information about the dataset.

4.3.2 Bonferroni's test:

Here we do Bonferroni's test on the model fm1 taking α as 0.05 . And no. of points rejecting is **98**. As previously done we remove them from the dataset and fit a new model. Now , the **multiple R square** is **0.7175**.So we keep them in our dataset.

4.3.3 Cook's Distance:

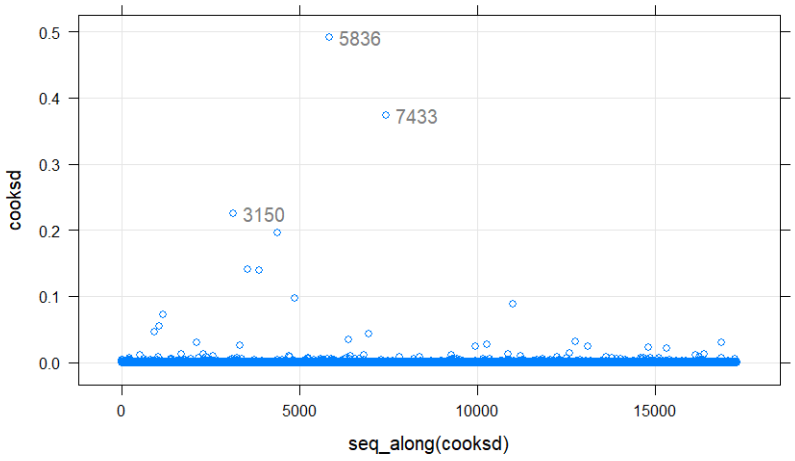

Figure 4.6: Cook's Distance.

Here We plot cook's distance of the each points of the model fm1 to detect the influential observations. From this plot 4.6 we get that observations' no. whose cook's distance are greater than 0.2. We reject those points and fit a new model having **multiple R square**, **0.7081**.So we don't remove them from the dataset.

4.3.4 Cov Ratio:

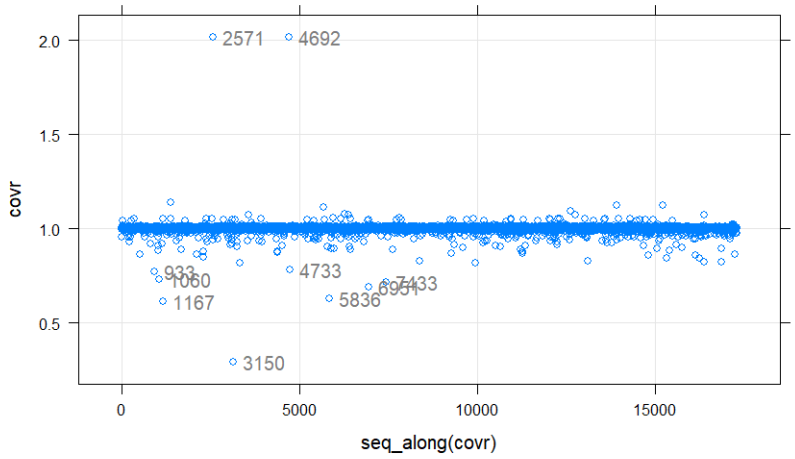

Figure 4.7: Cov Ratio.

After ploting 4.7 the Cov Ratio of the each point of model fm1, we take the cutoff the model as greater than 1.2 and less than 0.8. Again we fit a model without these 15 points and the **multiple R square** is **0.7078**. So we put this points in the datasets.

4.3.5 Observation:

From this we get very small amount of influential observation but it doesn't decrease the predictive power of the model so keep them in the model.

4.4 Validation of Model Assumptions

4.4.1 Non-Normality:

Cheaking of Non-Normality:

Here we do Kolmogorov-Smirnoff test for the test of normality. **p-value** of the test is $< 2.2e-16$.
Again,we plot different types of diagnostic plot of model fm1.From this plot we get that error distribution of model fm1 is not normal.

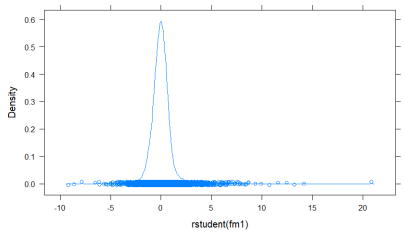

Figure 4.8: Density plot of rstudent of fm1.

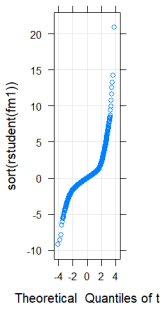

Figure 4.9: qqplot of fm1.

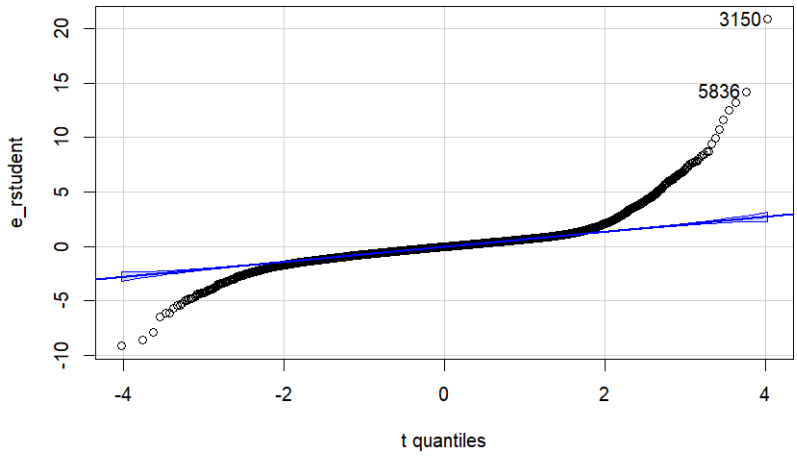

Figure 4.10: qqplot with C.I. of fm1.

Remedy of Non-Normality:

Here we have to do power transformation .So we have to find the estimate of λ and use Box-cox transformation. From the plot we get the estimation of λ as 0.045 so we do **log transformation** .Again from the histogram we get a positively skewed graph so log transformation is justified. So we take log of Price and create a new data frame.

Again I plot qqplot by fitting a new model say fm2 to check the normality.After transformation error distribution seems to be more or less normal.

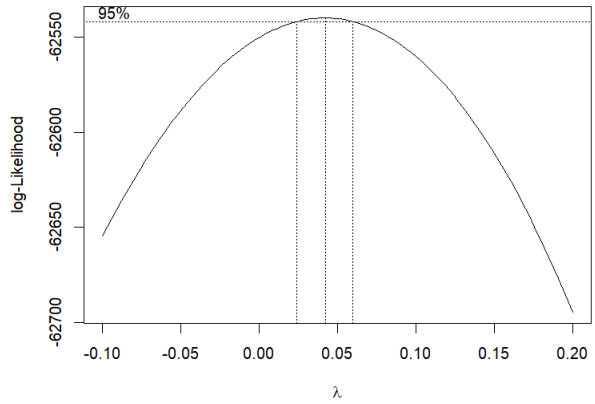

Figure 4.11: Estimation

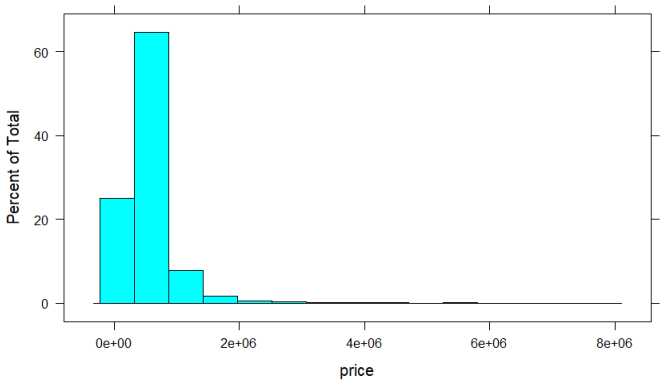

Figure 4.12: Histrogram of price.

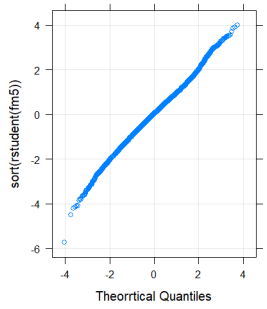

Figure 4.13: After transfomation qqplot

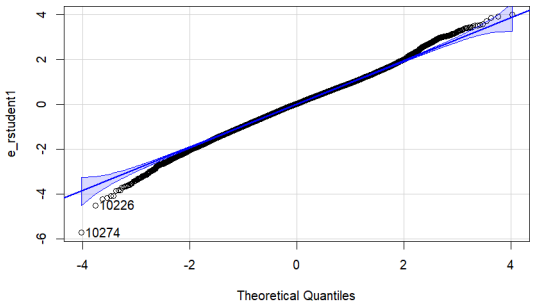

Figure 4.14: After transfomation qqplot with C.I.

4.4.2 Non-Constant Variance:

Here, after fitting the new model, fm2 we check for non constant variance . So we do **ncv test** also do **spread level plot**.

From the plot 4.11 we can see that the line is almost parallel to the horizontal so there is no non-constant variance.And also from the ncv test we get p-value which is also high

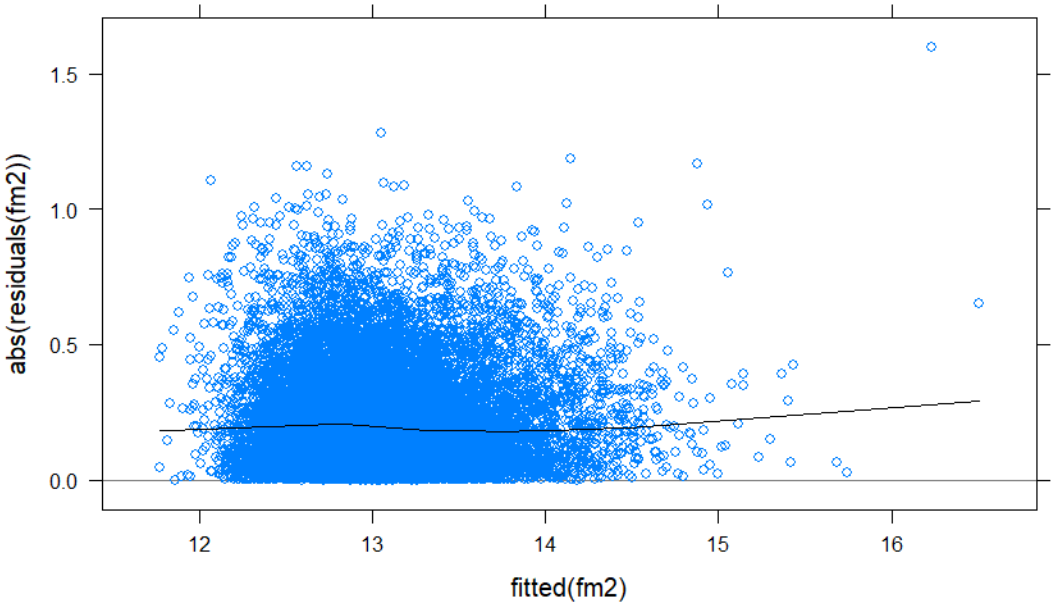

Figure 4.15: Spread Level Plot.

4.5 Predictive power of Linear Model:

Our final linear model is fm2.Esimates and p value of the are given below.

coefficients	Estimate	$Pr(>\|t\|)$
(Intercept)	3437655	2.74E-57
date2015	26058.73	1.60E-21
bedrooms	-14959.5	8.12E-17
bathrooms	29943.98	5.52E-31
sqft lot	0.1827	6.22E-05
floors	11469.44	0.000346
waterfront1	452916.6	2.30E-116
view1	90570.07	8.42E-18
view2	40038.65	3.66E-10
view3	95208.64	7.16E-27
view4	203003.9	1.48E-48
condition2	33306.1	0.38717
condition3	74423.85	0.038765
condition4	98126.88	0.006447
condition5	143442.2	7.48E-05
grade10	267016.7	0.118456
grade11	448965.6	0.008786
grade12	738279.4	1.91E-05
grade13	2290565	3.18E-31
grade3	-49811.8	0.810362
grade4	-159912	0.357165
grade5	-183994	0.281563
grade6	-156906	0.358186
grade7	-94836.6	0.578679
grade8	-11705.1	0.945367
grade9	114484.2	0.502912
sqft above	112.2671	7.80E-209
sqft basement	110.9394	1.10E-158
yr built	-1696.67	2.10E-131
zipcode9801	-79322.8	1.22E-19
zipcode9802	-104318	1.16E-67
zipcode9803	-53660.8	1.99E-21
zipcode9804	-52593	1.03E-13
zipcode9805	-74083.8	6.29E-41
zipcode9806	-97190.9	1.81E-16
zipcode9807	-73707.8	8.49E-29
zipcode9809	-218212	3.13E-88
zipcode9810	77953.49	1.07E-32
zipcode9811	87682.25	6.84E-47
zipcode9812	22201.43	0.002561
zipcode9813	-12213.6	0.127163
zipcode9814	-28303.3	0.000911
zipcode9815	-51252.4	2.83E-07
zipcode9816	-118495	8.47E-37
zipcode9817	-80384.5	6.18E-18
zipcode9818	-145752	1.02E-18
zipcode9819	-10452.2	0.230418
sqft living15	59.75191	6.14E-72
sqft lot15	-0.23656	0.00068
renovated1	47409.12	9.36E-12

The **mutliple R squared** is **0.7094** and **Predictive R squared** is **0.7093931**.

Now we use this model to predict price of the test data and compare with its actual. From this plot we can see by the model is a moderate for fitting purpose. Also

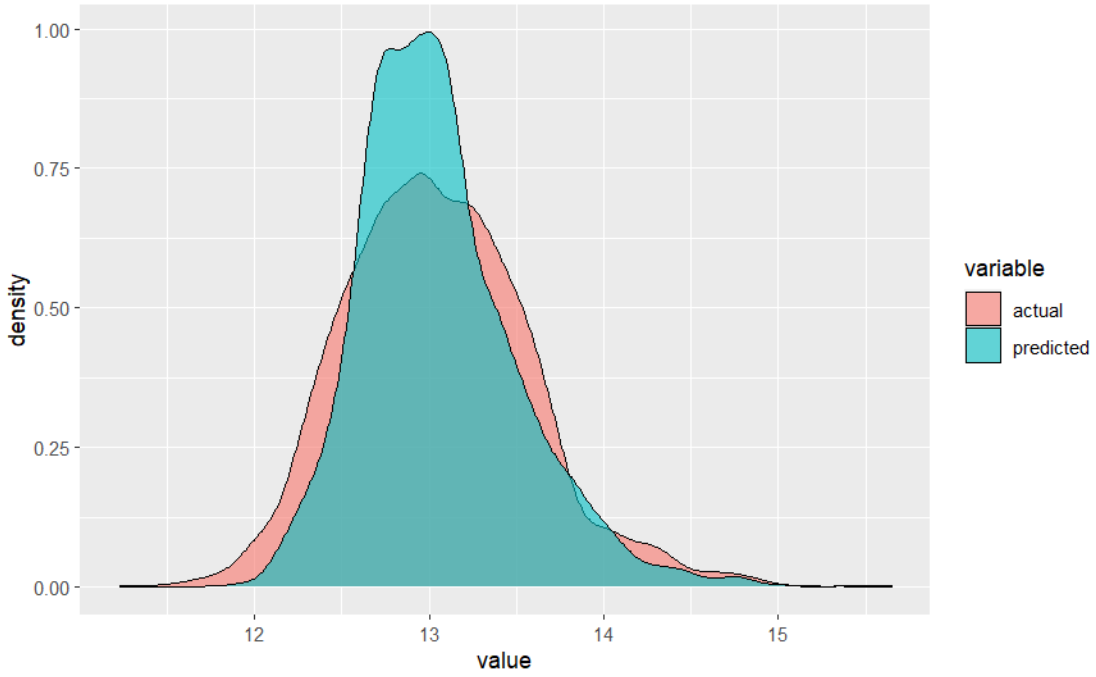

Figure 4.16: Actual vs Predicted of lm.

Correlation between Predicted and actual is **0.8462391** . So the model is good.

4.6 RLM

Since in our data there is some small amount of influential observations and the error distribution deviates from normality to some small extent so we use rlm where loss function bi.square and huber.The model for this two respectively rlm1 and rlm2.

Again, we use these two model in the test data set.Both the model seems to be very same with the previous linear model.

Also the correlation between actual and predicted are **0.8457159** and **0.8458957**

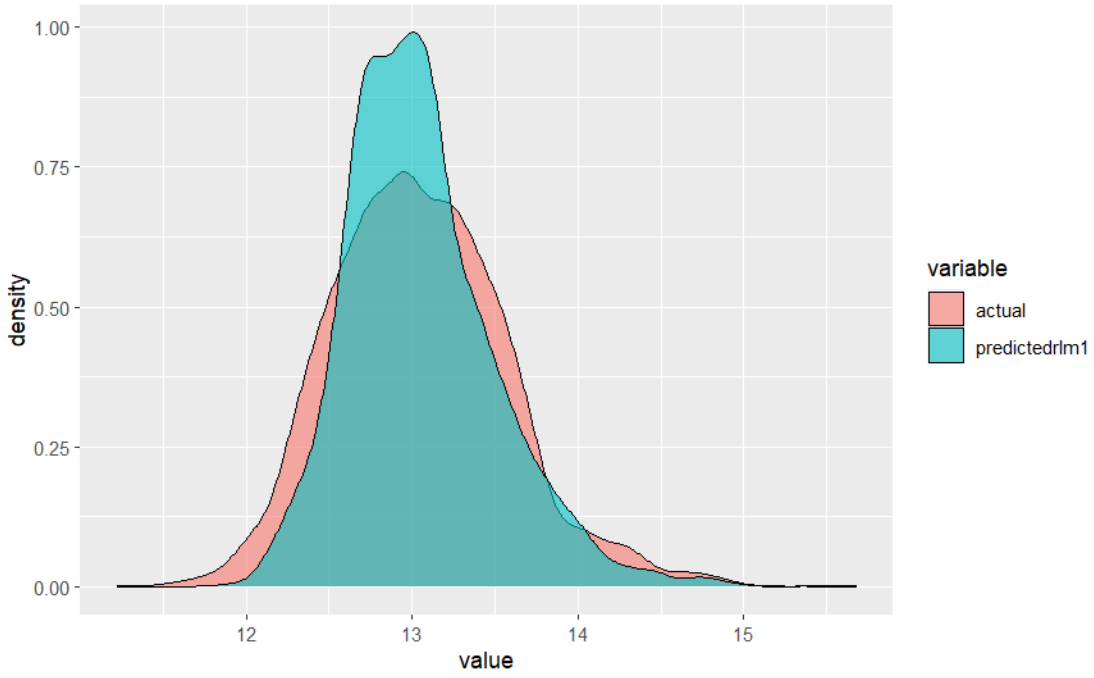

Figure 4.17: Actual vs Predicted of Bisquare.

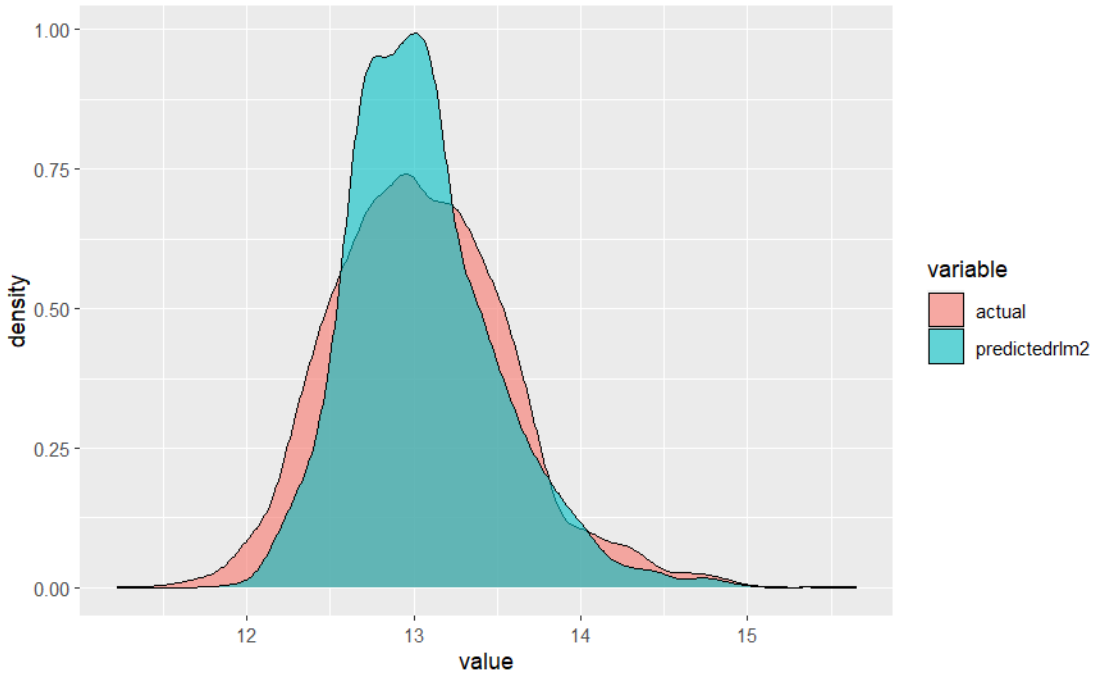

Figure 4.18: Actual vs Predicted of Huber.

4.7 LASSO

Now, we want to reduce no. of predictors in our model to get simpler model. So we apply LASSO in train data set.

From the plots we get that maximum 0.7 proportion of total variance explained in the model.By using only 17 predictors , we can explain 0.6 proportion of total variance.

Value of λ for min mse and 1se are 0.0049369372 and 0.0004394936 respectively.

From the sparse matrix we take the coefficients whose estimates are non-zero and fit a linear model with these predictors. The **multiple R square** is **0.6141** and **predictive R square** is **0.6141457**

here we do some diagonistic about linear model.From qqplot and spread level plot we get model assumpltion holds.

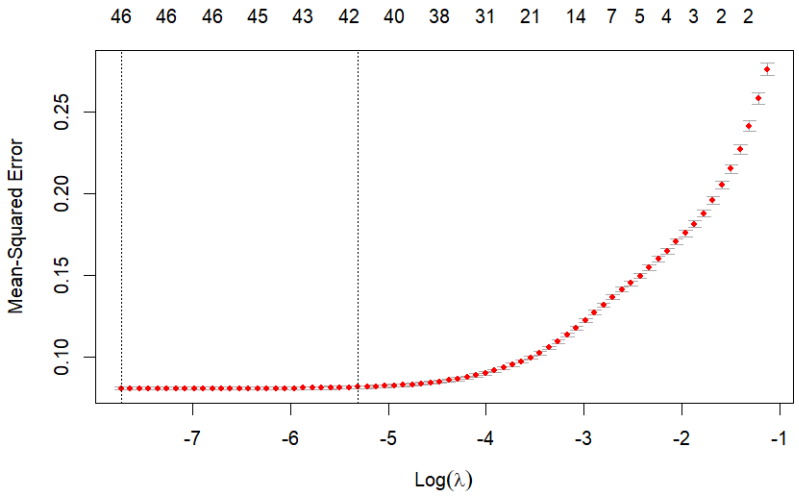

Figure 4.19: LASSO MSE.

We use the model to predict price in test data. Here correlation between predicted and actual is 0.7927181

4.8 Conclusion:

From comparing all correlation between actual and predicted between all variables multiple linear regression model is the best.But also there is no significant difference between rlm and lasso with multiple linear model.
Again if we compare all plot of actual vs predicted then also plot of multiple linear model and rlm seem to very same.
By comparing multiple R Square and predictive R square of multiple linear model and lasso we get multiple linear model is better for prediction purpose.

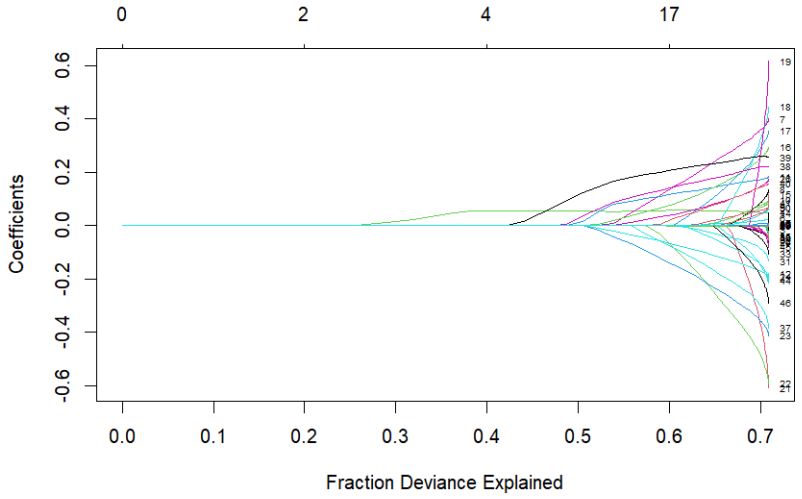

Figure 4.20: LASSO Fraction of Deviance.

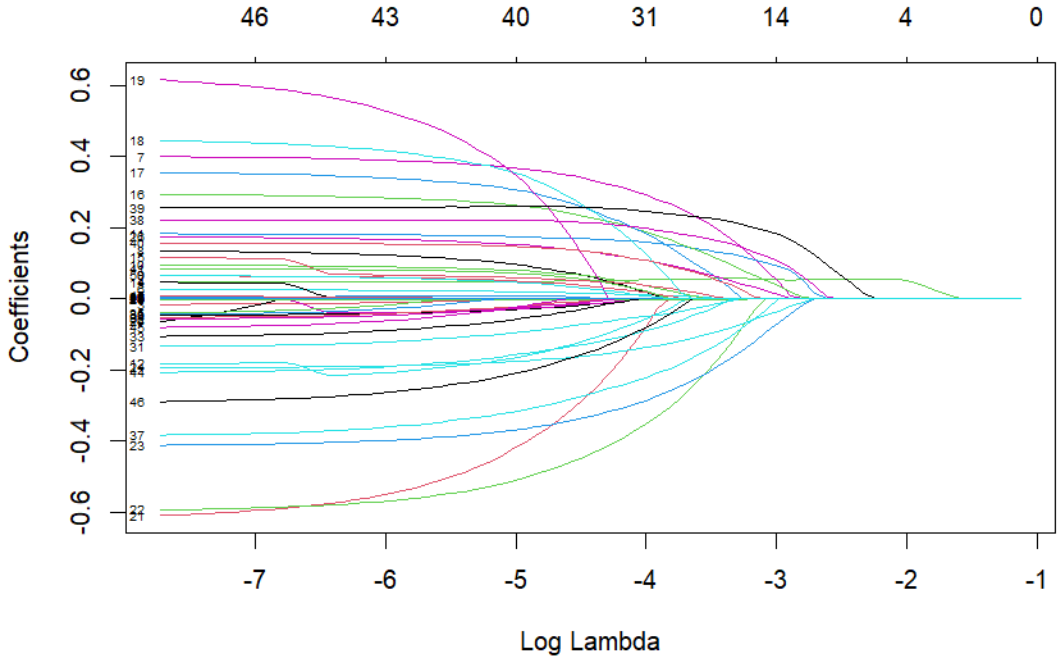

Figure 4.21: LASSO Lambda.

```
51 x 1 sparse Matrix of class "dgCMatrix"
                      s1
(Intercept)   17.869
date2014      -0.035
date2015       .
bedrooms      -0.005
bathrooms      0.049
sqft_lot       0.000
floors         0.021
waterfront1    0.377
view1          0.107
view2          0.054
view3          0.084
view4          0.177
condition2    -0.184
condition3    -0.033
condition4     .
condition5     0.062
grade10        0.272
grade11        0.320
grade12        0.380
grade13        0.424
grade3         .
grade4        -0.476
grade5        -0.536
grade6        -0.382
grade7        -0.182
grade8         .
grade9         0.159
sqft_above     0.000
sqft_basement  0.000
yr_built      -0.003
zipcode9801   -0.007
zipcode9802   -0.105
zipcode9803   -0.033
zipcode9804   -0.070
zipcode9805   -0.029
zipcode9806    .
zipcode9807    0.009
zipcode9809   -0.334
zipcode9810    0.222
zipcode9811    0.259
zipcode9812    0.151
zipcode9813    0.076
zipcode9814    .
zipcode9815    .
zipcode9816   -0.171
zipcode9817   -0.039
zipcode9818   -0.232
zipcode9819    .
sqft_living15  0.000
sqft_lot15     .
renovated1     0.051
```


Figure 4.22: Sparse Matrix of LASSO.

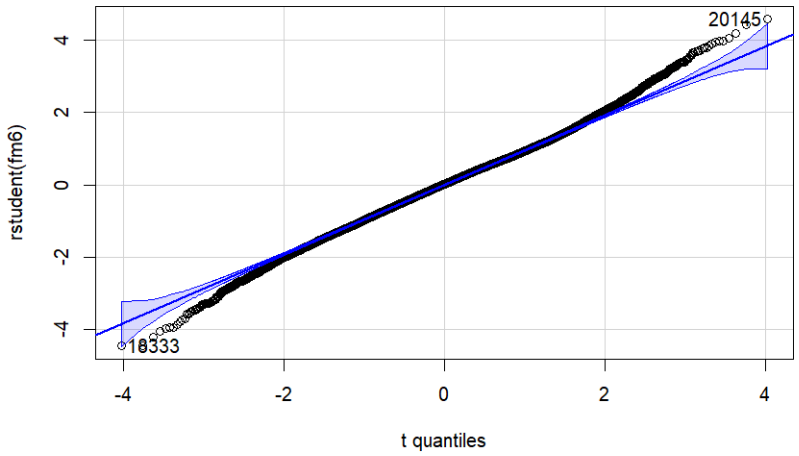

Figure 4.23: LASSO qqplot with C.I.

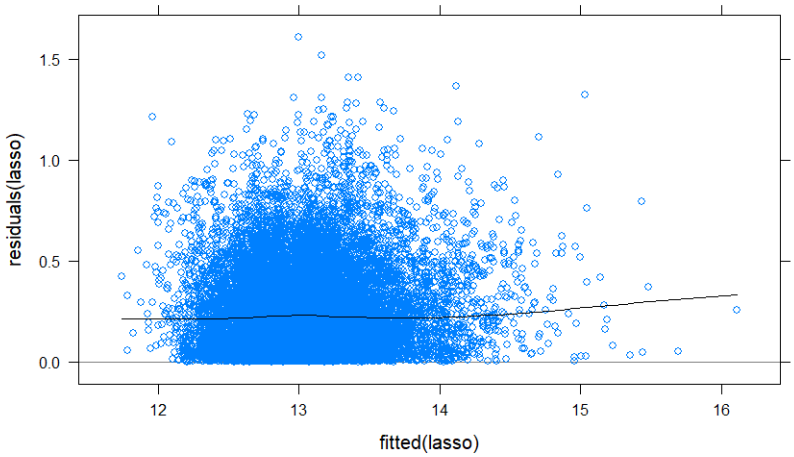

Figure 4.24: LASSO spread level plot.

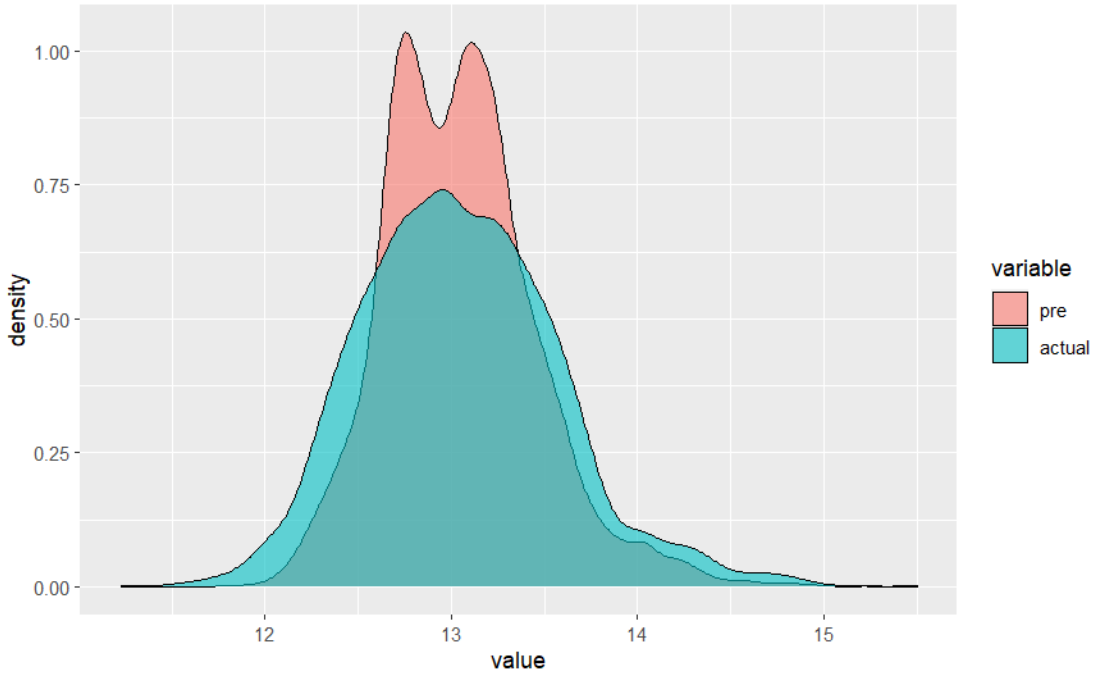

Figure 4.25: LASSO Predicted vs Actual.

Chapter 5

Acknowledgement

I would like to thank Prof. Deepayan Sarkar for his constant support and effort for helping me doing this project.
I would also like to thank my friend for helping in this project.