

Boston Area House Price Prediction

Vinay kumar Sharma

Rajat Agarwa

Deel

July 2020

Abstract

Here , we are working on BOSTON HOUSING data which represent the housing values in various suburbs of Boston. Our analysis starts with checking missing observation in the data(found no missing observation). Then we move to linear model setup and check overall significant of regressors. Then we standardize our regressors (not the response) moving to our main and most elementary task we split data into test and train set for further analysis and then move to linear regression assumption, Next we detect Leverage/influential points with help of Cook's Distance , DFFITS , DFBETAS , COVRATIO. Then move to deal with problems like Curvatures (correct them with suitable methods if we find any), heteroscedasticity(checking using plots of regressor vs residuals and also by Breusch Pagon test if found transformed our regressors),

Deal with non-normality(checking using Q-Q plot if found transform(Box-cox transformation) our response). Finding RMSE in each of model allows us to compare each models to get the model with best predicting power.

Key words: RMSE (Root Mean Square Error) , Breusch Pagon test , Multi-collinearity , Influential , Cook's Distance , normality , APR(Augmented Partial Residual) , CPR(Component Plus Residual).

Contents

1	Acknowledgement	5
2	Introduction	6
3	Objective	6
4	RMSE	6
5	Data Description	6
6	Data Pre-processing	7
6.1	Importing Data	7
6.2	Missing Value	8
6.3	Categorical Variable	8
7	Primary Model	8
7.1	Assumptions	8
8	Testing for overall regression and choosing which variable is significant for predicting MEDV	8
9	New model	10
9.1	Standardise our regressors	10
9.2	Splitting Data into Train and Test	11
10	Dealing with the Leverage and Influential Points .	11
10.1	Leverage Points	11
10.2	Diagnostics for Influential points	13
10.2.1	Cook's Distance	13
10.2.2	DFFITs	15
10.2.3	DFBETAS	16
10.2.4	COVRATIO	16
11	Dealing with Curvature	17
11.1	Plot of residuals vs individual regressors	18
11.2	checking on the behalf of APR and CPR	18
11.3	Transformation of regressors	20
12	Dealing with Heteroscedasticity	21
12.1	Plots of residual vs fitted response.	22
12.2	Testing of presence of Heteroscedasticity with Breusch Pagan Test	23
12.3	Now fit our model again and find RMSE.	23
13	Dealing with the Non-Normality	24
13.1	Checking Through Plot	24
13.2	Box-Cox transformation	25
13.3	Transformed response and Our new model with respect to minimum RMSE.	26
14	Conclusion	27
15	Bibliography	27

1 Acknowledgement

We would like to express our heartfelt gratitude to Dr. Minerva Mukhopadhyay for helping us in every difficulty which we face during this project. It has been a great learning experience and has also provided us with a practical insight of the theoretical knowledge gathered during the course Regression Analysis.

We also take this opportunity to thank the authors and publishers of the various books and journals we have consulted. Without those this work would not have been completed.

We would also like to thank our parents for their extensive support throughout the session. Their constant encouragement has enabled us to complete the project within the stipulated time-period.

2 Introduction

The Boston housing market is highly competitive, and we want to be the best real estate agent in the area. To compete with our peers, we decide to leverage a few basic machine learning concepts to assist us and our client with finding the best selling price for their home. Luckily, we've come across the Boston Housing dataset which contains aggregated data on various features for houses in Greater Boston communities, including the median value of homes for each of those areas. Your task is to build an optimal model based on a statistical analysis with the tools available. This model will then be used to estimate the best price homes as per demand of our clients.

3 Objective

Here, Our target variable is MEDV(Median value of owner-occupied home in 1000's dollar) . Here our interest lies in following point :

(1.)To Build a model that satisfy all usual assumption of General linear model.Is Prediction accuracy of model that we constructed is good enough ?

(2.)Which feature influence is less and more on the predicted house prices?

4 RMSE

We will use Root Mean square Error (RMSE) to check how good our model accuracy for predictions about our response. formula for RMSE is as :-

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_{test} - \hat{y}_{test})^2$$

$$RMSE = \sqrt{MSE}$$

Where y_{test} is our response for test data set and \hat{y}_{test} is the predicted value of response on test data, Calculated based on the fitted model obtained with train data.

5 Data Description

Title of the Data Set :- Boston housing dataset.

Source :- Harrison, D. and Rubinfeld, D.L. (1978) Hedonic prices and the demand for clean air. J. Environ. Economics and Management 5, 81–102.

Belsley D.A., Kuh, E. and Welsch, R.E. (1980) Regression Diagnostics. Identifying Influential Data and Sources of Collinearity. New York: Wiley.

Data Set Information :- The Boston data set has 506 rows and 14 columns. This data set contains the following columns:

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 :- nitrogen 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 mean of distances to five Boston employment centres.

RAD :- index of accessibility to radial highways.

TAX :- full-value property-tax rate per 10,000.

PTRATIO :- pupil-teacher ratio by town.

BLACK :- $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town.

LSTAT :- lower status of the population (percent).

We have these 13 Attributes as regressors(x_j). Which also have 12 Attributes as of continous formed and one is of binary valued Attribute(CHAS)

MEDV :- median value of owner-occupied homes(in 1000's dollors).This variable is taken as **response variable (Y)**.

6 Data Pre-processing

6.1 Importing Data

We will use library(MASS) for importing the library containing the Boston Housing Data setting name "data" to our boston housing data, setting Y for our response vector and X as our Design matrix and estimate our coefficients.

A rectangular box with a thin black border. Inside the box, the text "1.png" is centered in a plain, black, sans-serif font.

Figure 1: Declaring the function

6.2 Missing Value

Fortunately in given dataset it is clearly mention that there is no missing observations which make our work little bit less.

6.3 Categorical Variable

Again fortunately there is no categorical variable, actually our job is already done because in or dataset CHAS has already in indicator or binary typed so no need of doing anything for this variable.

7 Primary Model

Consider our multiple linear regression model as :

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \dots\dots\dots(1)$$

Where $\mathbf{X}_{n \times p+1} = (\mathbf{1}, x_1, x_2, \dots, x_{13})$ is the non-stochastic matrix of predictors (here $\mathbf{1}$ is the column whose all elements are 1) and $\beta = (\beta_0, \beta_1, \dots, \beta_{13})$ is the vector of regression coefficients and $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)$ is the error vector.

7.1 Assumptions

- (i) $\epsilon_i \stackrel{iid}{\sim} \mathbf{N}(\mathbf{0}, \sigma^2) \forall i$, where σ^2 is an unknown constant.
- (ii) \mathbf{X} is of full column rank.

8 Testing for overall regression and choosing which variable is significant for predicting MEDV

Let's first see

overall regression of our model -

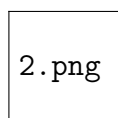


Figure 2: Estimates of coefficients

Clearly we see that we reject the hypothesis that

$$H_0 : \beta_0 = \beta_1 = \beta_2 = \dots = \beta_{13} = 0$$

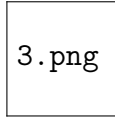


Figure 3: Test of overall regression

therefore there exist atleast one β_j which is significant for predicting $Y(\text{MEDV})$.

Performing Individual regressor test -



Figure 4: Individual regressors test result for significant

Clearly we see regressors INDUS and AGE is not significant for predicting MEDV. But it doesn't mean that these two variables are bad for predicting MEDV.

9 New model

Consider new linear regression setup with 'MEDV' as the response, and the remaining variables, except 'ZN', 'CHAS' and 'RAD', as predictors as :

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \quad \dots\dots\dots(1)$$

Where $\mathbf{X}_{n \times p+1} = (\mathbf{1}, x_1, x_2, \dots, x_{10})$ is new non-stochastic matrix of predictors with remaining regressors and a column of ones (here $\mathbf{1}$ is the column whose all elements are 1), here $p=10$ and $\beta = (\beta_0, \beta_1, \dots, \beta_{10})$ is the vector of regression coefficients and $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)$ is the error vector.

9.1 Standardise our regressors



Figure 5: new standardise design matrix

9.2 Splitting Data into Train and Test

Here use the concept of test and training of data to check our predictive power of model on the test set so lets first standardise the X matrix and then split the data into 80% to train and 20% to train and then move to outlier detection and check other assumptions.

7.png

Figure 6: splitting into test and train

Now we fitted our model from train data and then find RMSE from test data,

8.png

Figure 7: splitting into test and train

10 Dealing with the Leverage and Influential Points .

10.1 Leverage Points

Leverage points are those which have unusual x values but does not affect the model estimates much. Leverage points can be find out with the help of Hat matrix(H), where Hat matrix is given as:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

if \mathbf{h}_{ii} is ith diagonal element of Hat matrix, then traditionally ith point is considered as leverage if

$$\mathbf{h}_{ii} > \frac{2 * \mathbf{p}}{\mathbf{n}} \quad \dots\dots(a)$$

Clearly from the given threshold of $\mathbf{h}_{ii} = 0.05418719$ total 31 points are identified as leverage, we remove those now the total number of observations in train set is become 375. Further move to detection of influencial points by using *Cook's distance* , $DFFITs_i$, $DFBETAs_{ji}$, *CovRatio* .

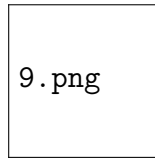


Figure 8: Codes for Leverage points

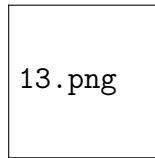


Figure 9: Codes for Leverage points

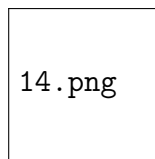


Figure 10: Leverage points graphically

10.2 Diagnostics for Influential points

Influential points are those which have both X and Y unusual. These points have considerable influence on estimates of coefficients and these pull the direction of regression line towards itself.

Our aim is to compute *Cook's distance*, *DFFITs_i*, *DFBETAs_{ji}*, *CovRatio* and identify the influential point.

10.2.1 Cook's Distance

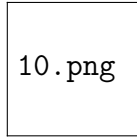
Cook's distance is one of the measures used for identification of Influential Points. Cook's distance is calculated as :

$$C_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})^T (\mathbf{X}^T \mathbf{X}) (\hat{\beta} - \hat{\beta}_{(i)})}{p \text{MSres}}$$

$$C_i = \frac{r_i^2 * h_{ii}}{1 - h_{ii}}$$

where, p is number of regressors, n is total no. of observations. $r_i = \frac{e_i}{\sqrt{\text{MSres}(1-h_{ii})}}$ is the internal studentized residual and e_i is the i th residuals.

It is clear from formula that it contains both part r_i (studentized residual) and h_{ii} (i th diagonal element of hat matrix) which measure how far observation is from data and how well the model fits the i th observation y_i .



10.png

Figure 11: Code for detection of influential points using Cook's distance

Traditionally, the points for which C_i is greater than 1 is consider influential. But there is no points shows using Cooks Distance.

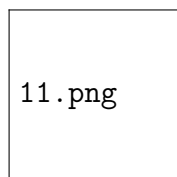


Figure 12: Cook's codes for plot with threshold=1

Here down the graphical repretation which shows there is no points outside the given traditional threshold $C=1$.

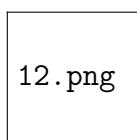


Figure 13: Cook's distance Bar plot when threshold=1

10.2.2 DFFITS

DFFITS is another method for influential point detection and it also contain both part (r_i^{*2} and h_{ii}) which measure how far observation is from data and how well the model fits the i th observation y_i . $DFFITS_i$ is calculated as :

$$\mathbf{DFFITS}_i = \sqrt{\frac{r_i^{*2} * h_{ii}}{1 - h_{ii}}}$$

$r_i^* = \frac{e_i}{\sqrt{MSres_i(1-h_{ii})}}$ is the external studentized residual. e_i is the i th residuals. $MSres_i =$

$$\frac{\sum_{j \neq i}^n e_j^2}{(n - 1 - p)}$$

h_{ii} is the i th diagonal element of matrix $H = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$

15.png

Figure 14: Code for calculation of influential points using DFFITS

16.png

Figure 15: Codes for graphycal reprentation of DFFITS using threshold=0.3425395

dffit.png

Figure 16: DFFITS graphycal reprentation of DFFITS using threshold=0.3425395

Traditionally cut-off is given by :

$$|\mathbf{DFFITS}_i| > \frac{2 * \sqrt{p}}{\sqrt{n}}$$

and if $DFFITS_i$ is greater than this cut-off that indicates that i th sample point is influential point. In this case 0.3425395 is consider as threshold.

10.2.3 DFBETAS

It is also used for influential point detection and it also contain both part (r_i^{*2} and h_{ii}) which measure how far observation is from data and how well the model fits the i th observation y_i , which is clear from formula of $DFBETAS_{ji}$ $DFBETAS_{ji}$ is calculated as:

$$DFBETAS_{ji} = \frac{r_i^* * C_j}{\sqrt{h_{jj}(1 - h_{ii})}}$$

r_i^* is the external studentized residual. h_{ii} , h_{jj} is the i th, j th diagonal element of matrix $H = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ respectively and C_{ji} is the j^{th} row of the matrix $C = ((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$

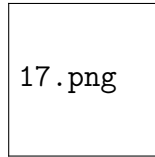


Figure 17: Code for calculation of influential points using $DFBETAS_{ji}$

Generally threshold is given as for this case

$$DFBETAS_{ji} > \frac{2}{\sqrt{n}}$$

And if $DFBETAS_{ji}$ exceeds this cut off than we say that i th observation is influential. From Fig.34 it is clear that only those points are considered to be influential points which are considered to be influential by 7 or more regressors out of 9 numerical regressors which are tested for DFBETA

10.2.4 COVRATIO

The *Cook's distance*, $DFFITs_i$, $DFBETAS_{ji}$ provides insight about the effect of observation on estimates of coefficients and fitted value y but they do not provide precision of estimation. COVRATIO also used for this purpose. We express the role of i th observation on precision of estimates in terms of $COVRATIO_i$.

We define $COVRATIO_i$ as

$$covratio_i = \frac{(MSres_i)^p}{(MSres)^p * (1 - h_{ii})}$$

$MSres_i$ is the estimate of the variance of response variable when i th data point is removed from the data. $MSres$ is the estimate of the variance of response variable using all the data. h_{ii} is the i th diagonal element of matrix $H = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$.

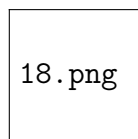


Figure 18: Code for computation of influential points using COVRATIO_i

If $COVRATIO_i > 1$ then i th observation improves the precision of estimates and if $COVRATIO_i < 1$ then i th observation degrades precision. Traditionally, Cut-off for COVRATIOS is defined as

$$|\mathbf{covratio}_i - 1| > \frac{3 * p}{n}$$

There is one influential point using this approach

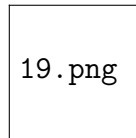


Figure 19: COVRATIO codes for graph representation

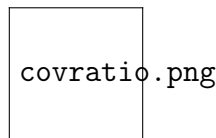


Figure 20: COVRATIO graph representation

Now taking **all of the unique influential observaion's** we get on the basis of DFFITs , DFBETA , COV RATIO.

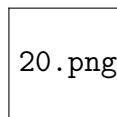


Figure 21: Codes for overall influential points

Overall there are 44 influential points. After deleting leverage and influential points we have now 331 observations are left.

After deleting leverage and influential points RMSE is 5.698279 which is less than the previous RMSE(5.9042) which implies it is better model than previous ones. That means There is lots effect of Outliers in the model w.r.t. to this train and test data.

11 Dealing with Curvature

Here we check that if any curvature or nonlinear regressors are present with respect to our response, there are some task here

1.) First we check whether the residuals show any pattern when plotted against the individual regressors. Identify at least one regressor which does not seem linearly related with $E(y)$.

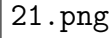
21.png

Figure 22: codes for new RMSE after deleting all influential points.

2.) Check the difference between Augmented Partial Residual (APR) and Component Plus Residual (CPR) plots to justify our claim.

3.) Choose a suitable transformation of the regressor(s) (that we tested in (2.) which is linear in $E(y)$). WE can get an idea of the appropriate transformation from the plot of partial residuals of y vs the regressor(s), or simply from the plot of errors vs the regressor(s).

4.) Verify (using correlation coefficient) if linearity has improved by using the transformed regressor(s).

5.) Change the regressor(s) chosen in part (3.) by the transformed one. Compute \hat{y}_{test} and RMSE again. Keep the change if you get a non-increasing RMSE after transformation, otherwise do not transform.

11.1 Plot of residuals vs individual regressors

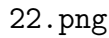
22.png

Figure 23: Codes for plot

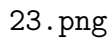
23.png

Figure 24: residual vs individual plots

Clearly from the plot of residual vs regressor plot we seem "**rm**" not to be **linearly related with $E(y)$** .

11.2 checking on the behalf of APR and CPR plotting of CPR

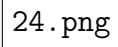
A square box containing the text "24.png".

Figure 25: Codes for plot CPR

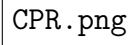
A square box containing the text "CPR.png".

Figure 26: plot CPR

plotting of APR

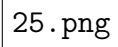
A square box containing the text "25.png".

Figure 27: codes for APR plot

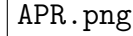


Figure 28: plot for APR

Clearly from Observing CPR and APR plot. We suggest that the **exponential transformation** is needed for the regressor 'rm'.

11.3 Transformation of regressors

So from the result we observed from the plot of CPR and APR of variable "RM", we here transform it to exponential transformation that is ,

$$g(x) = \beta_0 * \exp(\beta_1 * X_{tr_{rm}})$$

Fitting the transformed model of variable "RM"

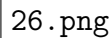


Figure 29: codes for transformation

Now we find partial coefficient between Ytrain and "rm" of Xtrain.

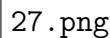


Figure 30: codes for Correlation coefficient without any transformation and with exponential transformation.

Clearly we see that without any transformation the partial correlation coefficient between response and RM is less than our transformed one so that indicate improvement in the regressor rm therefore our transformation seems good, now lets change our design matrix with replace X_{rm} with $g(x)$ and then find fitted our OLS model to find RMSE again and verify our result.

new model and finding RMSE again.

Clearly our new RMSE after transformation is decrease therefore this indicate improvement is appropriate hence we finally take this transformed model thats mean we regain our linearty and solve this curvature issue.

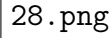
28.png

Figure 31: codes for new transformed design matrix and again for new RMSE.

12 Dealing with Heteroscedasticity

One of the assumption of our regression model is that our data should be homoscedastic. Consider the regression equation $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{10} x_{i10} + \epsilon_i$. $i=1(1)n$ In this model, if ϵ_i has constant variance say σ^2 than data is said to be homoscedastic and if variance of ϵ_i depends on i then data is said to be heteroscedastic and it is said that heteroscedasticity is present in data.

Here we have some task;

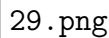
1.) check whether the residuals show any pattern when plotted against the fitted values, and then by plotting w.r.t. the individual regressors. Find the regressors with respect to which the variance (WE can approximate by e_i^2) is not constant.

2.) Considering $d'_i = \frac{ne_i^2}{\sum_{i=1}^n e_i^2}$, then find appropriate variables z_1, \dots, z_k , so that the following relation is approximately true:

$$d'_i = \alpha_0 + \alpha_1 z_{1,i} + \alpha_2 z_{2,i} + \dots + \alpha_k z_{k,i} + \epsilon'_i$$

where $\epsilon'_i \stackrel{iid}{\sim} \mathbf{N}(0, 1)$. Hence, perform a Breusch-Pagan test and verify if heteroscedasticity is present in this data.

3.) If the Breush-Pagan test rejects the homoscedasticity hypothesis, then we consider an appropriate function $\sigma_i^2 = h(Z, \alpha, \beta)$, and estimate σ_i^2 and simultaneously. [Also we may consider a linear function, i.e., $h(Z, \alpha, \beta) = \alpha_0 + \sum_{j=1}^k \alpha_j z_{j,i}$ However, We note that this choice of h may not result in positive values. We may estimate the $\hat{\beta}$ and $\hat{\alpha}$ by using this the following iterative algorithm:

29.png

4.) Once the algorithm converges, We again find the RMSE with the new estimate of β .

12.1 Plots of residual vs fitted response.

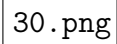
A rectangular box containing the text "30.png".

Figure 32: codes for plot of residual vs individual regressor and then with fitted response.

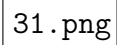
A rectangular box containing the text "31.png".

Figure 33: plot of residual vs individual regressors.

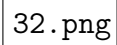
A rectangular box containing the text "32.png".

Figure 34: plot of residual vs fitted response.

Clearly from the plot of residual vs individual regressors and of fitted response ,we claim that **CRIM, INDUS, AGE, DIS, TAX, BLACK** produce **Heteroscedasticity**.

12.2 Testing of presence of Heteroscedasticity with Breusch Pagan Test

We test the hypothesis,

H_0 : Error variance is Homoscedastic.

H_1 : Error variance is Non- Homoscedastic.

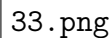


Figure 35: codes for testing the hypothesis.

Clearly from the results Chi square test statistic $Q= 16.44206$ which is greater than the critical value (12.59159) ,Hence we reject the null hypothesis that means the errors are heteroscedastic. So now lets perform itteration for converging the algorithm to find appropriate β .

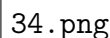


Figure 36: codes for itteration.

12.3 Now fit our model again and find RMSE.

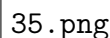


Figure 37: codes for new transformed design matrix and again for new RMSE.

Clearly we see that new RMSE(after remove hetroscedasicity with transformation) is less than the previous one so we succeed to removal of hetroscedasticity.

Now this is our new model after removal of outliers(leverage and influencial) removing curvature and now removing hetroscedasticity everything is good now,Lets move fo normality assumption.

13 Dealing with the Non-Normality

In section 7.1 we assumed $\epsilon_i \stackrel{iid}{\sim} \mathbf{N}(\mathbf{0}, \sigma^2)$ which is equivalent to assume $\mathbf{y} \stackrel{iid}{\sim} \mathbf{N}_n(X\beta, \sigma^2 I_n)$. Now as y is observed quantity, we will check only its Normality. For this we have to do some task here;

1.) Firtly we Draw the QQ-plot of the R-student residuals w.r.t. it's population distribution and Comment our results on the normality assumption based on the QQ-plot.

2.) If the plot shows departure from normality, let consider an appropriate Box-Cox transformation to normalize the data. Find the parameter λ using profile likelihood. (We may choose the appropriate λ graphically.)

3.) After the optimal λ is chosen, we recompute the R-students residual with transformed response, and draw the QQ-plot again. And from the plot observe the normality assumption on the transformed y to be correct or not.

13.1 Checking Through Plot

We will check Normality by Density curve of \mathbf{y} and by Normal Q-Q Plot. In Density Curve if the curve looks like bell shaped like normal curve , then normal assumption is appropriate. On the other hand in Normal Q-Q Plot order statistics of vector of interest are drawn in the y-axis corresponding to the theoretical order statistics from $\mathbf{N}(0, 1)$ in the x-axis. Normality assumption can be regarded as true if the points almost in a straight line.

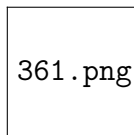
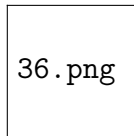


Figure 38: codes for QQ-plot of y .

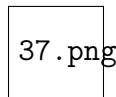


Figure 39: QQ-plot without any transformation.

Looking at the QQ-plot, We find that all the points are not in straight line. Hence the response is not normal.

13.2 Box-Cox transformation

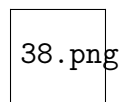


Figure 40: codes for maximised λ .

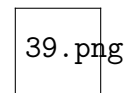


Figure 41: plot of MLE of λ .

Clearly from the graph and R output the value of λ which maximize the profile likelihood is 1. Now lets transform our response.

13.3 Transformed response and Our new model with respect to minimum RMSE.

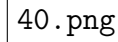
A placeholder box containing the text "40.png".

Figure 42: Codes of transformation of y.

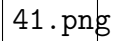
A placeholder box containing the text "41.png".

Figure 43: QQ-plot after transformation.

From the trasformed QQ-plot it is clearly seems normality assumption seems to have improved as thus obtained plot is closer to ideal situation.

Now finally calculate our RMSE with respect to this model,

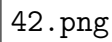
A placeholder box containing the text "42.png".

Figure 44: QQ-plot after transformation.

Finally after solving the normality issue we found our final models RMSE is 5.106228, which is minimum hence this model can be taken final for predicting the boston houses median price (MEDV).

14 Conclusion

For Boston housing data, predicting $Y(\text{MEDV})$ we have different models

15 Bibliography

1. An Introduction to Statistical Learning,with Applications in R by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani
2. Introduction to Linear Regression Analysis by Douglas C Montgomery, Elizabeth A Peck, G. Geoffrey Vining
3. The Elements of Statistical Learning:Data Mining, Inference, and Prediction by Trevor Hastie,Robert Tibshirani,Jerome Friedman
4. <http://dx.doi.org/10.1016/j.jbusres.2016.02.010> ((Moro et al., 2016) Moro, S., Rita, P., Vala, B. (2016))
5. <https://stackoverflow.com>
6. www.analyticsvidhya.com

16 Appendix

The files used in this project can be found at the following link. This include data file and R code.

https://drive.google.com/drive/folders/1_Tw7cTaX78s8HzM6nmt-PVEgtDvDLcTN?usp=sharing