Regression Analysis Project Work on " Boston House Data"

Summited by:-

1) **Abhilash Shingh**
2) **Ankit Gupta**

Project Work Duration:-Feb'2020-April'2020

INTRODUCTION

In this Project work we, we worked on BOSTON HOUSING data which represent the housing values in various suburbs of Boston. First, we perform Exploratory Data Analysis (EDA), in which we checked for missing value observation, Categorical variables and made Boxplots

In the Boston housing dataset, consider the variable 'MEDV' as the response, and the remaining variables, except 'ZN', 'CHAS' and 'RAD', as predictors. Standardize the regressors (not the response). Randomly split the data into 2 parts: one part containing 100 samples (say, Xtest, ˜ ytest : test data), and the other part containing the remaining samples (say, Xtrain, ˜ ytrain : training data).

First of all, We with the leverage points and outliers. We chose threshold according to data so too much points not get dropped. Now,compute Cook's distance, DFFIT, DFBETA and COVRATIO. Hence, find the influential points and delete those. Plot Cook's distance, DFFIT and COVRAIO.

Now we Deal with Curvatures. First check whether the residuals show any pattern when plotted against the individual regressors. Check the difference between Augmented Partial Residual (APR) and Component Plus Residual (CPR) plots. Then Choose a suitable transformation for removing pattern.

After that deal with heteroscedasticity and Normality which are assumption for linear regression model. For checking Heteroscedasticity Breush-Pagan test and check Non-Normality using QQ-plot. For removing Non-Normality apply Box-Cox transformation. And then finally apply Shapiro test for normality checking.

We compute the RMSE of model at every step. And at We fit a model on final data and compute test error.

Data Structure:

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(xj). 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 10000s dollors).This variable is taken as response variable (Y).

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.

```
> library(MASS)  # library including the Boston datasets
> data<-as.matrix(Boston)
> #Structure of dataset
> str(data)
 num [1:506, 1:14] 0.00632 0.02731 0.02729 0.03237 0.06905 ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:506] "1" "2" "3" "4" ...
  ..$ : chr [1:14] "crim" "zn" "indus" "chas" ...
> #Dividing the columns of data in Y(as response) and X(as predictor)
> Y=data[,ncol(data)]    #Response
> #Number of obserbvation's in respose
> length(Y)
[1] 506
> X=data[,-ncol(data)]   #Predictor Matrix
> #Predictor matrix including intercept term
> X=cbind(rep(1,nrow(X)),X)
> #dimension of predictor matrix
> dim(X)
[1] 506  14
> colnames(X)[1]="intercept"
> #Names of each predictor variable
> colnames(X)
 [1] "intercept" "crim"      "zn"        "indus"     "chas"      "nox"       "rm"
 [8] "age"       "dis"       "rad"       "tax"       "ptratio"   "black"     "lstat"
> #So, we are removing regressor 'zn' , 'chas' , 'rad'
> j=which(colnames(X)=="zn")
> k=which(colnames(X)=="chas")
> l=which(colnames(X)=="rad")
> #removing columns of ZN,CHAS,RAD from X(Predictor matrix)
> X=X[,-c(j,k,l)]
```

Figure 1: Importing data and removing 'zn','chas', 'rad'

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 General linear regression setup with 'MEDV' as the response, and the remaining variables, except 'ZN', 'CHAS' and 'RAD', as predictors.Our model is as :

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

Where $\mathbf{X}_{n\times p+1} = (\mathbf{1}, x_1, x_2, \cdots, x_{10})$ is the non-stochastic matrix of predictors (here $\mathbf{1}$ is the column whose all elements are 1) and $\beta = (\beta_0, \beta_1, \cdots, \beta_{10})$ is the vector of regression coefficients and $\epsilon = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)$ is the error vector.Here , we are not considering three regressor as our interest doesnot include that feature for house prices.

7.1 Assumptions

(i) $\epsilon_i \overset{iid}{\sim} \mathbf{N}(\mathbf{0}, \sigma^{\mathbf{2}})$ $\forall i$, where σ^2 is an unknown constant.
(ii) $\mathbf{X}$ is of full column rank.
(iii) There is no curvature.
(iv) There is no **Hetroscedasticity**.

7.2 Standardise our regressors

here , we are standardizing our predictor matrix so as to convert all feature in same scale .

```
> #standardising the regressor of datasets
> for(i in 2:ncol(X))
+ {
+   X[,i]=(X[,i]-mean(X[,i]))/sd(X[,i])
+ }
```

Figure 2: Standardizing Predictor matrix

7.3 Splitting Data into Train and Test

Here use the concept of Validation set approach to check our predictive power of model on the test set . Let split the data randomly into nearly 80% (406 out of 506) as train set

```
> # set.seed() will help to select the same random sample
> #whenever  we select a random #sample from data
> set.seed(52)
> #Selecting 100 random sample from total observation
> indices=sample(1:nrow(data),replace=FALSE,100)
> # Setting up  training set(406 observation) and
> #test set(100 observation) data Selected from whole data
> Y_train=(Y[-indices])
> X_train=(X[-indices,])
> #observations in train datasets
> dim(X_train);length(Y_train)
[1] 406  11
[1] 406
> Y_test=Y[indices]
> X_test=(X[indices,])
> #observation in test dataset
> dim(X_test);length(Y_test)
[1] 100  11
[1] 100
```

Figure 3: Splitting into test and train

and nearly 20% (100 out of 506) as test set and then move to outlier detection and check other assumptions.

Now we fitted our model from train data and then find RMSE from test data to check our model accuracy. But , here we have to check whether the predictor matrix is full column rank or not .If the predictor matrix is full column rank then we can further estimates our coefficient estimate using least square solution

7.4 Estimated coefficient, model fitting and RMSE.

```
> det(t(X_train)%*%X_train)
[1] 1.036169e+26
> #determinant of (t(X_train)%*%X_train) is not zero . Hence we conclude the matrix is full colum
n rank matrix
> #Least square solution or co-efficients efficent of the linear model fit
> beta_hat=solve(t(X_train)%*%X_train)%*%t(X_train)%*%Y_train
> beta_hat
                [,1]
intercept 22.3175204
crim      -0.3813723
indus     -0.7324612
nox       -1.6718780
rm         2.4131685
age       -0.2408388
dis       -2.4910960
tax        0.3467440
ptratio   -2.1871680
black      0.6728200
lstat     -3.7625047
> # estimated value of the  mean housing price of testing data
> Y_test_cap=X_test%*%beta_hat
> #RMSE Value of the testing the datasets
> RMSE=sqrt(sum((Y_test-Y_test_cap)^2))/sqrt(length(Y_test))
> RMSE
[1] 5.9042
```

Figure 4: Least Square Solution of coefficients

Here , we can see that RMSE value is 5.9042 which represent the standard deviation of residual error .Here we further check for the assumption of Linear model and Diagnosis each assumption and further check whether there is decrease in RMSE or not .If RMSE is decreasing this indicate that model accuracy has been improving each step.

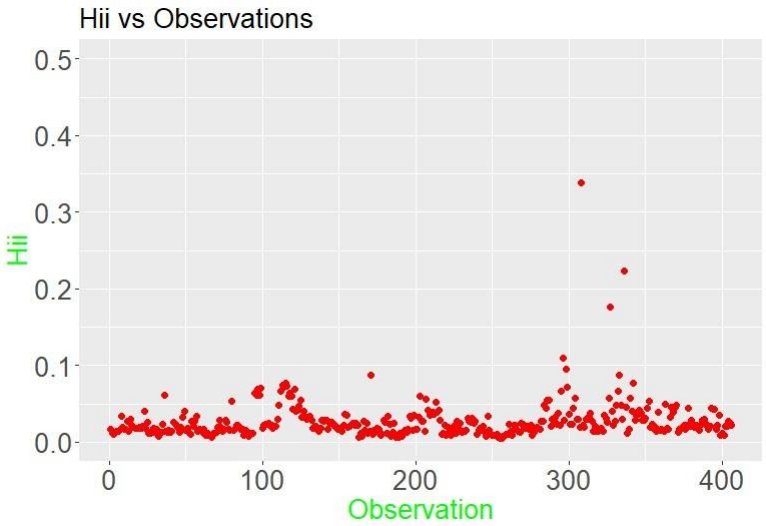
Dealing with Leverage point and influential points.

Leverage points

To find leverage point ,we usually focus on the diagonal elements h_{ii} of the H matrix,which may written as

$$hi_{ii} = \mathbf{x_i'}(X'X)^{-1}\mathbf{x_i}$$

where x_i' is the i row of X matrix.The Hat matrix diagonal is a standardized measure of the distance of the ith observation from the centre of the $\mathbf{x}$ space,thus the large hat diagonal reveal observations that are potentially influential because they are remote in $\mathbf{x}$ space from the rest of the sample. we assume that any observation for which hat diagonal is greater than $\frac{2p}{n}$+0.2 consider as leverage point.
Graph btw h_{ii} and Observations are given below



For detecting influential points we will use Cook's Distance ,DFBETA ,DFFIT and COVRATIO.

Cook's Distance:
Formula for calculating Cook's Distance

$$\mathbf{D_i} = \frac{(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}})' X'X (\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}})}{pMS_{res}}$$

Point with large values of D_i have considerable influential on the least squares estimates $\hat{\beta}$ .We usually consider points for which

$$D_i > 1$$

to be influential.

DFBETA:
The first of these is a statistics that indicates how much the regression coefficient $\hat{\beta}_j$ changes,in standard deviation units,if the$i$th observations is deleted. it is denoted by

$$\mathbf{DFBETAS_{j,i}} = \frac{\hat{\beta}_j - \hat{\beta_{j(i)}}}{\sqrt{{S_i}^2 C_{jj}}}$$

where C_{jj} is the jth diagonal element of $(X'X)^{-1}$ and $\beta_{j(i)}$ is the jth regression coefficient computed without use of ith observation.A large value of $DFBETAS_{j,i}$,indicates that observation i has considerable influence on the jth regressor coefficient. We use a cut off of $\frac{2}{\sqrt{n}}$ i.e,if $|DFBETAS_{j,i}| > \frac{2}{\sqrt{n}}$ then ith observation warrants examination.
DFFITS:
We may also investigate the deletion influence of the ith observation on the predicted or fitted value. This is done by $DFFITS_i$

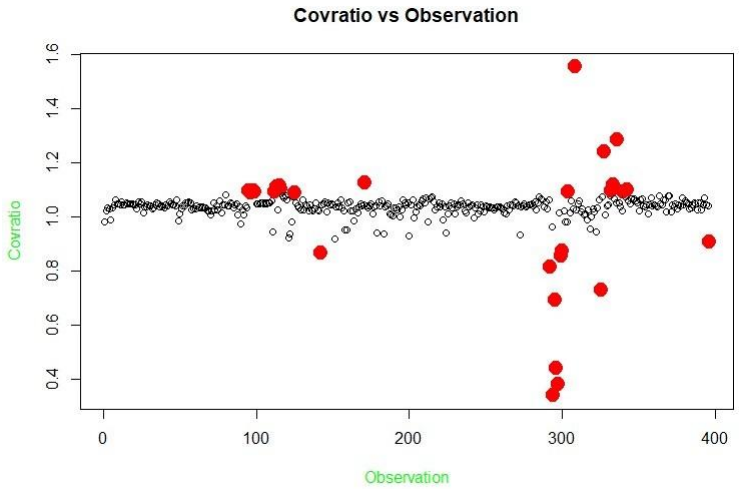
$$\mathbf{DFFITS_i} = |DFFITS_i| = \sqrt{(\frac{h_{ii}}{1-h_{ii}})} t_i$$

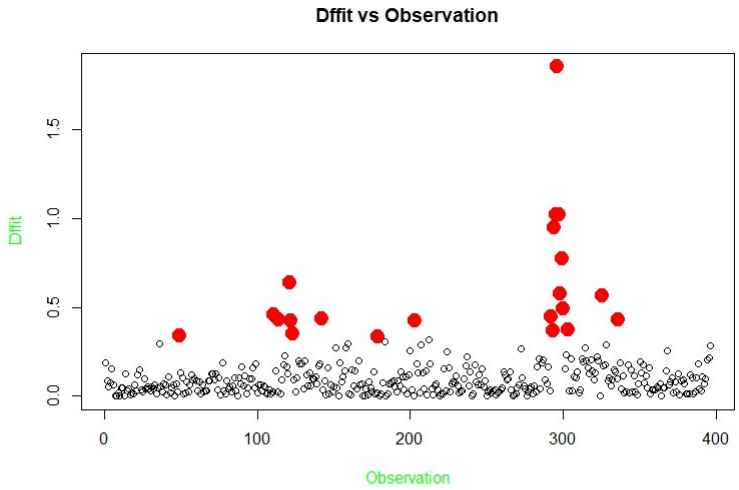
where h_{ii} is the diagonal elements of hat matrix and t_i is the R-Student.We will use the fact that any observation for which $> 2\sqrt{\frac{p}{n}}$ warrants attention,here p denotes the rank of hat matrix.
COVRATIO:
To express the role of the ith observation on the precision of estimation, we could define

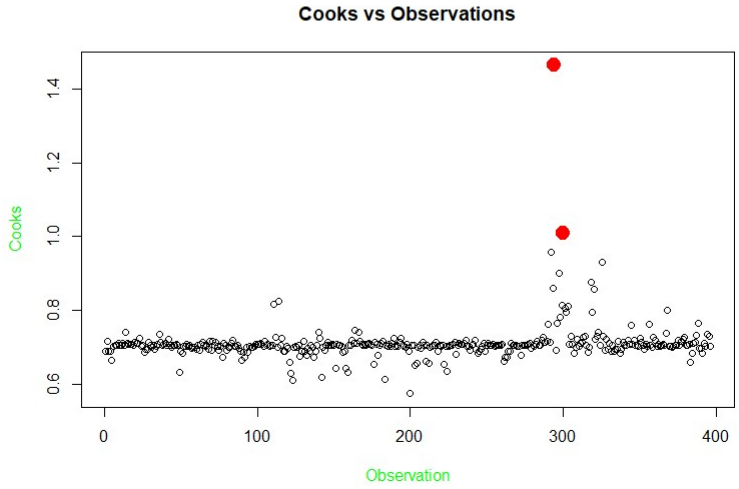
$$\mathbf{COVRATIO_i} = \frac{({S_i}^2)^p}{{MS_{Res}}^p}(\frac{1}{1-h_{ii}})$$

We will use fact that if $|COVRATIO_i - 1| > 3\frac{p}{n}$ then ith point should be considered as influential Point.

Now we plot COVRATIO , DFFITS ,Cooks Distance



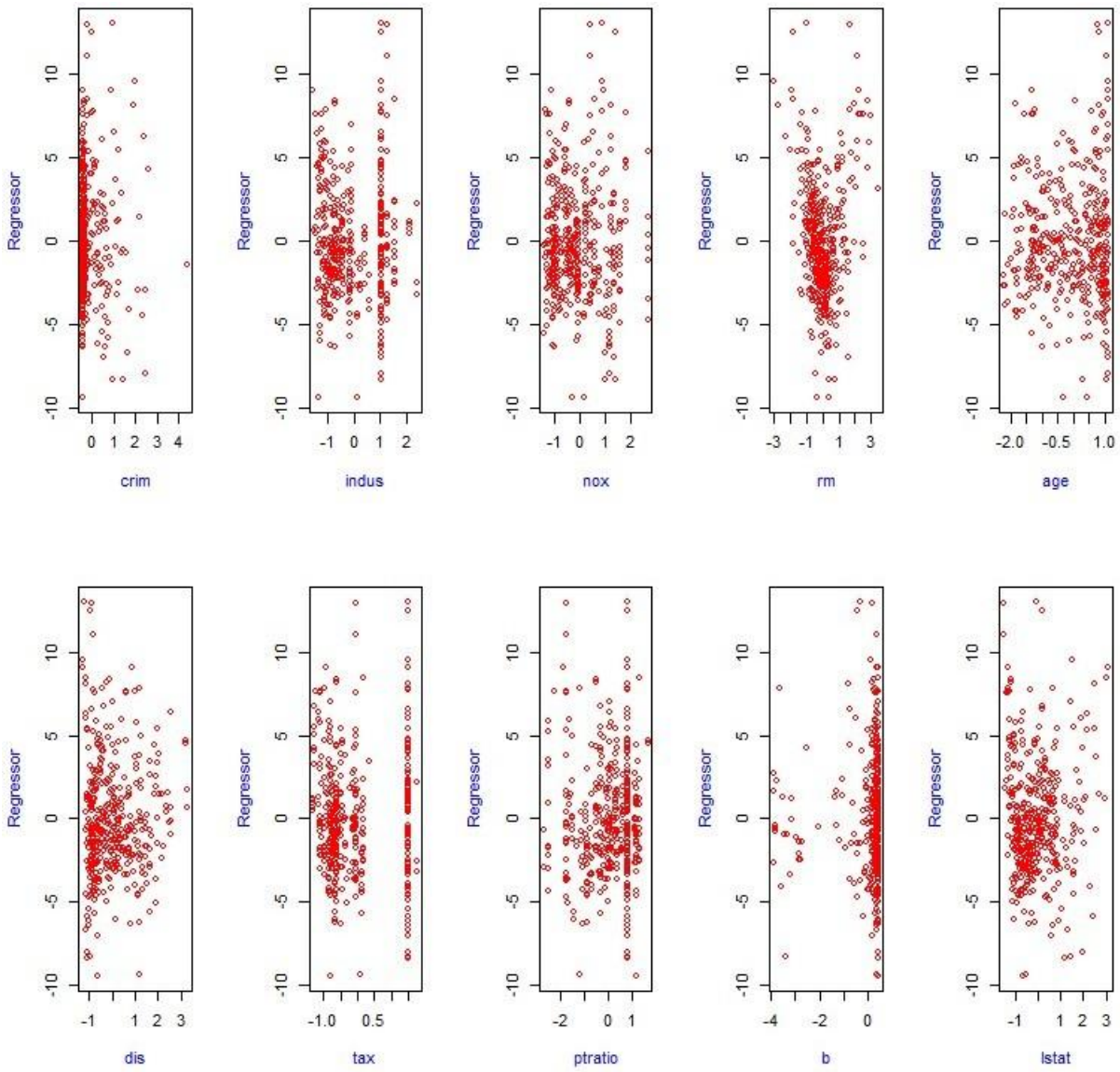




In this graph Red Points denotes the influential point.

Value of RMSE before deleting the leverage Point and influential Point is 4.549473 and after deleting the leverage point and influential point is 4.051916
Hence value of RMSE is decreases after deleting the influential point and leverage point.

Dealing with Curvatures

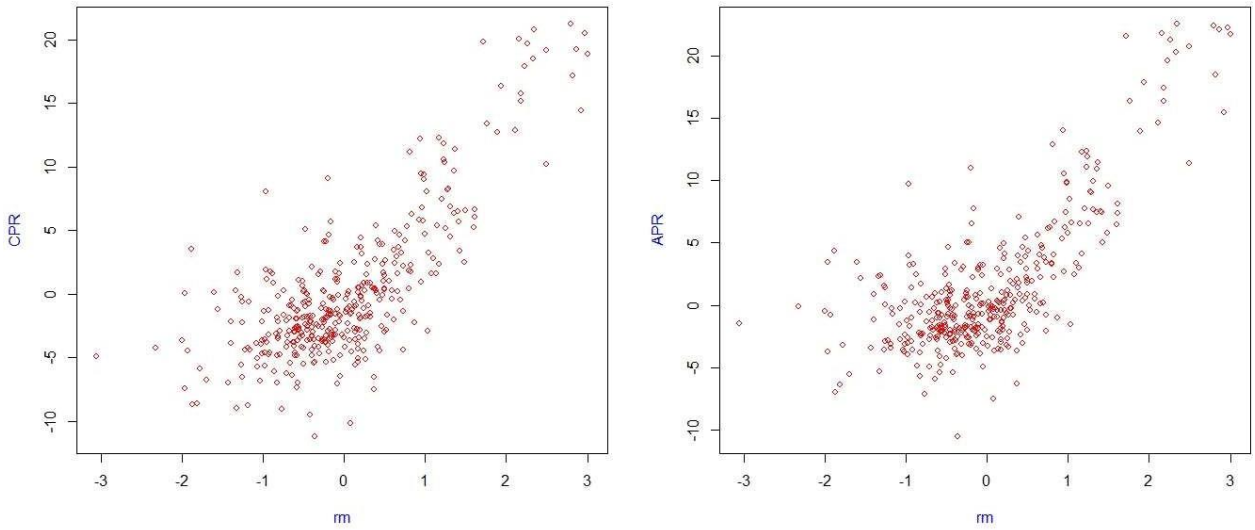


We first observe the residual vs regressor plot for each regressor.

Fig 1: Residual plots

The residual plot of rm regressor shows some U-shaped pattern.We claim that rm may not be linearly related with the response.

We now plot the CPR and APR plot of regressor(attached at the last of the pdf). There is no such major difference between APR and CPR plots of regressors ,However a close scrutiny reveals that there is some difference for regressor rm(attached below)



We now observe the CPR plot of rm to choose suitable transformation.

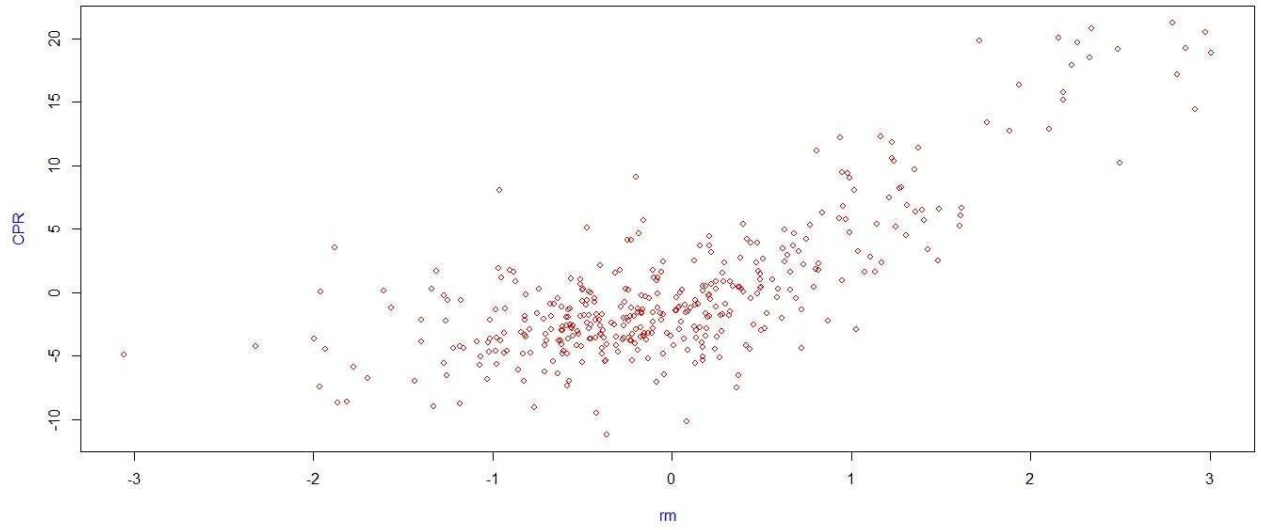


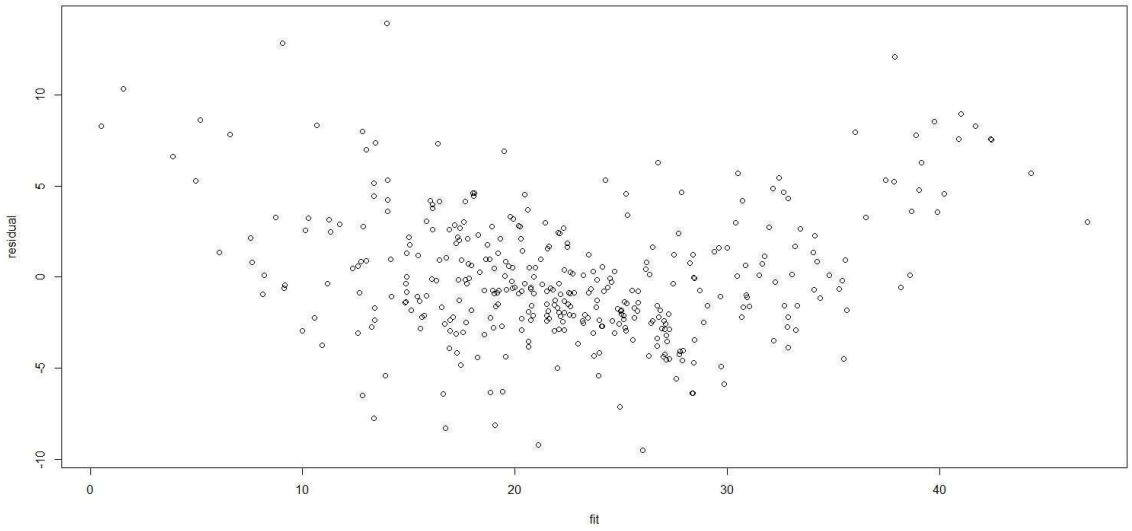
Fig 1: CPR plot of rm

Observing the graph we use the transformation $g(x) = \beta_0 e^{\beta_1 x_{rm}}$. Taking log on both sides of the model we get $log(y_{Train}) = log(\beta_0) + \beta_1 x_{Train}$, which is a linear model. We now estimate β_0 and β_1 by regressing $log(y_{Train} + c)$ on $log(\beta_0) + \beta_1 x_{rm}$. We get $\beta_0 = 20.831$ and $\beta_1 = 0.2978335$.

We now calculate correlation coefficient between y_{Train}, x_{rm} and $y_{Train}, g(x)$
Here $corr(y_{Train}, x_{rm}) = 0.7922$ and $corr(y_{Train}, g(x)) = 0.82$. As the correlation creases we use this transformation on the regresor rm and fit the regression model with the transformed rm. We now calculate the RMSE again. The RMSE comes out to be 15.72489, which is higher than the RMSE obtained without transformation. As the RMSE increases we do not take this transformation and keep the rm regressor as it was.

Dealing with the heteroscedasticity

We first observe the plot below:



From the graph of residual vs fitted value we see that the graph is U-shaped. Again from the graph of residual vs regressor in 2(a) we see that the graph of regressor 'rm' shows a U-shaped pattern.

Only one regressor shows some pattern. So, we choose $d_i{}^* = \alpha_0 + \alpha_1 Z_1 + \epsilon^*$
Now we choose: Z_1=$C_0 + C_1 X + C2X^2$ where x corresponds to 'rm'

$$= \frac{cov^2\left(d_i^*, d_i^*\right)}{var\left(\hat{d}_i^*\right) var\left(d_i^*\right)}$$

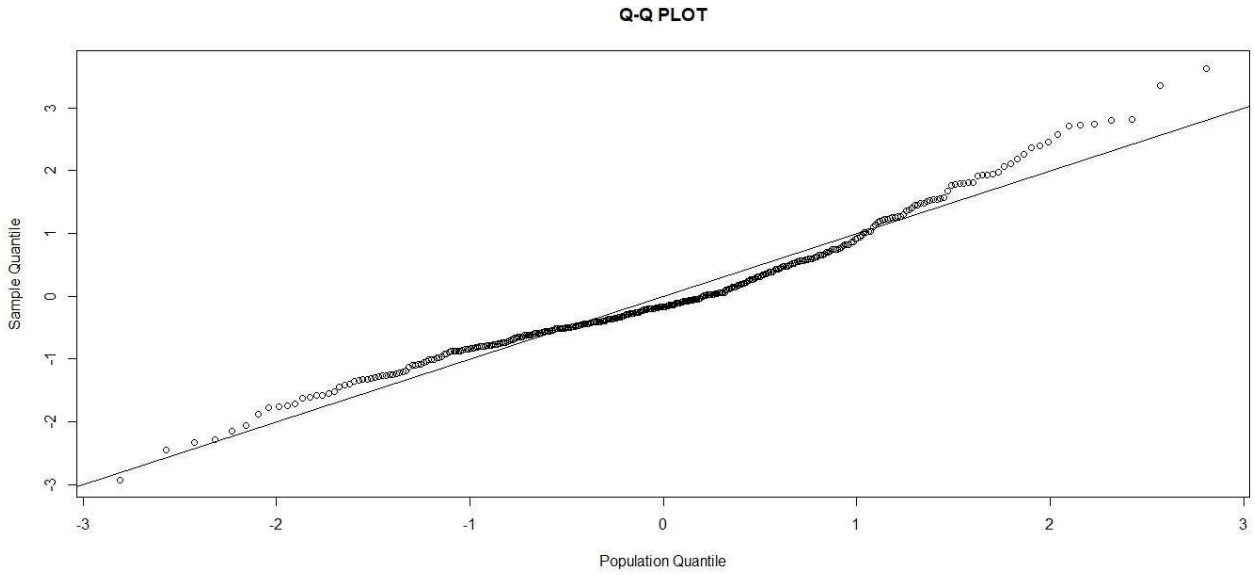
we reject H_0 at 5% level of significance if $Q_obs \geq \chi^2_{0.05,1}$. Here $Q = 66.36$ which is greater than $\chi^2_{0.05,1} = 3.84$. So we reject H_0.

Here we take: $\sigma_i^2 = \alpha_0 + \alpha_1 Z_1$
Let Σ be the dispersion matrix with diagonals as $(\sigma_1^2, ..., \sigma_n^2)$. Here we see that same as d_i^*. We find $\Sigma = diag(\sigma_1^2, ..., \sigma_n^2)$. So, $\hat{\beta} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$. So be changed and hence d_i^*. So, $\hat{\alpha}$ will be changed. We will continue updating $\hat{\alpha}$ and till there is significant difference between two consecutive values of $\hat{\alpha}$ and $\hat{\beta}$.

We now calculate RMSE with the updated β after the process converges. It is found that the RMSE comes out to be 5.013884, which is greater than the previous one. So, we suspect that our chosen $h(a, \alpha, \beta)$ is not appropriate.

Checking For normality and Box-Cox Transformation

Now we draw QQ-plot of the R-student residuals against it's population distribution. Here we consider R-student residual quantile as sample quantile and y-training data-quantile as population quantile.



From this graph, it is visible that distribution of y-training data is deviated from our normality assumption.

4.b To fix the problem, we do box-cox transformation to normalize the Ytrain data.

Box-Cox Transformation:

Formula of Box-Cox Transformation

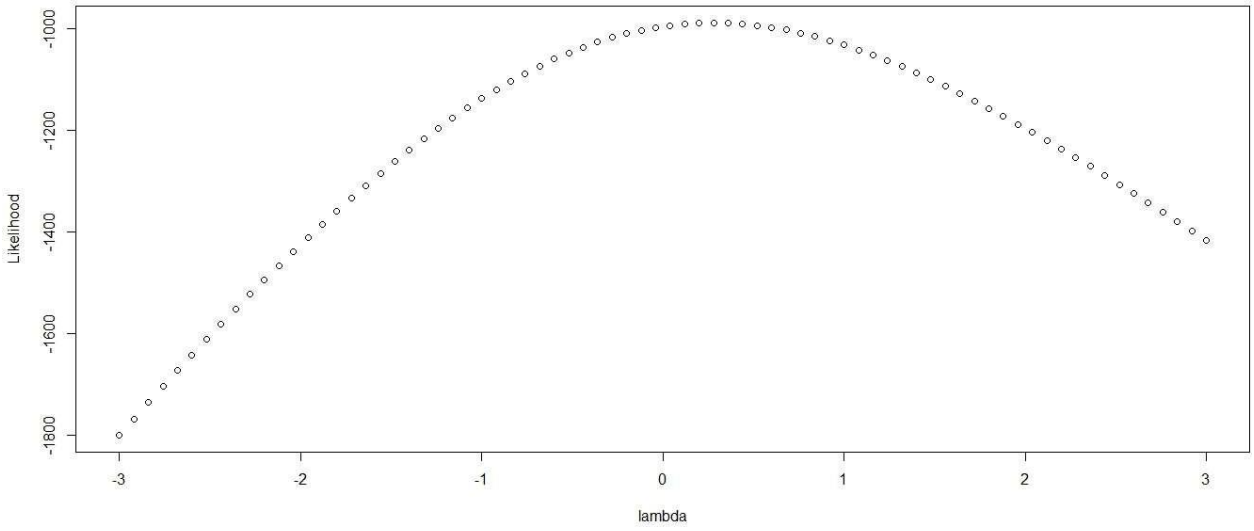
$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

To estimate the parameters of the box-cox transformation, we need the value of λ for which the expression :

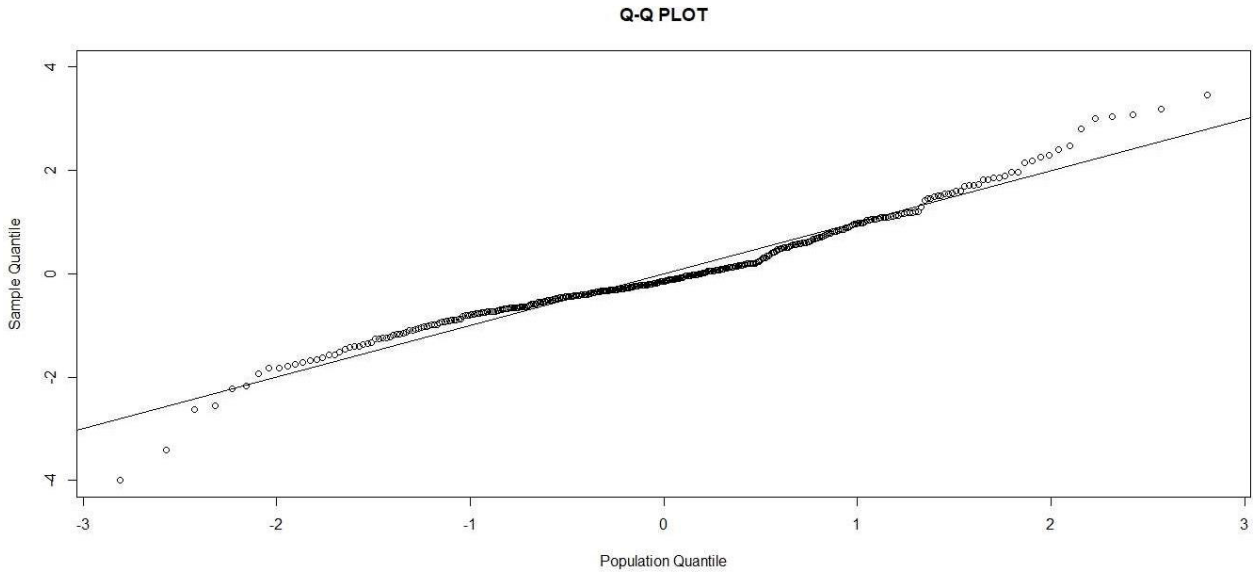
$\mathrm{L}(\lambda, \hat{\beta}^{\lambda}_{MLE}, \hat{\sigma}^{2,\lambda}_{MLE}) = -\frac{nlog2\Pi}{2} - \frac{n \log \frac{RSS_{\lambda}}{n}}{2} - \frac{n}{2} + n(\lambda - 1)\log(G)$

is maximized.

We compute the likelihood function for different λ values and plot it against λ to find the maximum value of likelihood function



Now we transform the y-training data through box-cox transformation and again fit regression model with the transformed y and recompute the R-student residuals.
Again we draw QQ-plot of the new R-student residuals against it's population distribution. We obtain he graph as follows :



We observe from the graph that, the transformed data now more close to the normality assumption.