Boston House Price Analysis

(A Regression Analysis ~~on the~~ Boston House Prices using R)

Sanket Agrawal (191124)
Sayan Das (191131)
Soham Ghosh (191143)
Abhiroop Chowdhury (191004)
Aritra Majumdar (191025)

Dealing with the Leverage Points and Outliers

We are given the Boston Housing dataset. We consider 'MEDV' as the response variable and the remaining variables, except 'ZN', 'CHAS' and 'RAD', as predictors. First, we standardize the regressors.
Next, we split the dataset randomly into two parts in R setting a particular set.seed(99999) so that we get the same partition every time we run our code.
One part of 100 samples is our test set $(X_{test}, \underset{\sim}{y}_{test})$ and the other part of 406 samples is our training set $(X_{train}, \underset{\sim}{y}_{train})$.
Now, we essentially do all our analysis on the training set and we use the test set only while computing the RMSE.
Let us, for simplicity, use $(X, \underset{\sim}{y})$ to denote our training set.
Thus, our model is given by : $\underset{\sim}{y} = \mathbf{X}\underset{\sim}{\beta} + \epsilon$
We calculate the OLS estimate of $\underset{\sim}{\beta}$ and obtain $\hat{\underset{\sim}{y}}_{test} = \mathbf{X}_{test}\hat{\underset{\sim}{\beta}}$
Thus, after calculation, the RMSE of $\underset{\sim}{y}_{test}$ comes out to be **4.740495**.

Now, we start with our diagnostics of leveարge points and influential points.

Leverage points: The leverage points are found using h_{ii} which is the i-th diagonal element of the hat matrix P, where $P = X(X^TX)^{-1}X^T$. Usually, any observation for which h_{ii} is greater than $2p/n$ (p is the number of number of columns of X and n is the number of sample points) is considered as a leverage point. Here, if we we use this threshold , we observe that 55 samples are levearge points. So, we plot the values of h_{ii} against the sample points and observing the plot we set a **suitable threshold** of **0.1**.

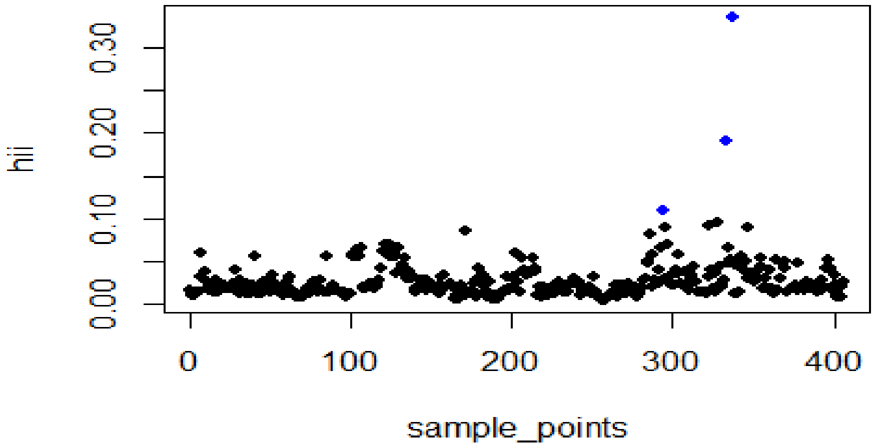


Figure 1: Leverage points

It is observed that observations corresponding to sample points 293, 333 and 337 exert significant amount of levearge on the fitted values. Hence, we drop these data points and do our analysis on the rest of our training data.

Let us name the data set $(X, \underset{\sim}{y})$ i.e. same as before as we don't want to introduce new variables.

Influential points: For detecting influential points, we will use the following measures :

i) **Cook's distance(C_i)**:

$$C_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T (X^T X)(\hat{\beta}_{(i)} - \hat{\beta})}{p\mathbf{MS_{Res}}}, i = 1(1)n$$

where $\hat{\beta}_{(i)}$ is the estimate obtained after deleting the i-th point and the rest of the symbols have usual meanings.
ii) **DFFITS**:

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{S^2_{(i)} h_{ii}}}, i = 1(1)n$$

where $\hat{y}_{(i)}$ is the fitted value after deleting the i-th point and $S^2_{(i)}$ is the external mean sum of squares with the i-th point deleted.
iii) **DFBETAS**:

$$DFBETAS_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{S^2_{(i)} C_{jj}}}, j = 1(1)p, i = 1(1)n$$

where $\hat{\beta}_{j(i)}$ is the j-th regression coefficient computed without use of the i-th point and C_{jj} is the j-th diagonal element of $(X^T X)^{-1}$.
iv) **COVRATIO**:

$$COVRATIO_i = \frac{(S^2_{(i)})^p}{\mathbf{MS^p_{Res}}} (\frac{1}{1 - h_{ii}}), i = 1(1)n$$

The traditional or emperical cut-offs for C_i, $DFFITS_i$, $DFBETAS_{j,i}$ and $COVRATIO_i$ are 1, $|2\sqrt{p/n}| = 0.33042$, $|2/\sqrt{n}| = 0.09963$ and $(1-3p/n, 1+3p/n) = (0.91812, 1.08188)$ respectively i.e. points with values more than these thresholds are considered to have influential effects.
For our analysis here, we have used the cut-offs 1, 0.4, 0.1, (0.9, 1.1) respectively. The plots of Cook's distance, DFFITS and COVRATIO using these cut-offs are given below in Figure 2.

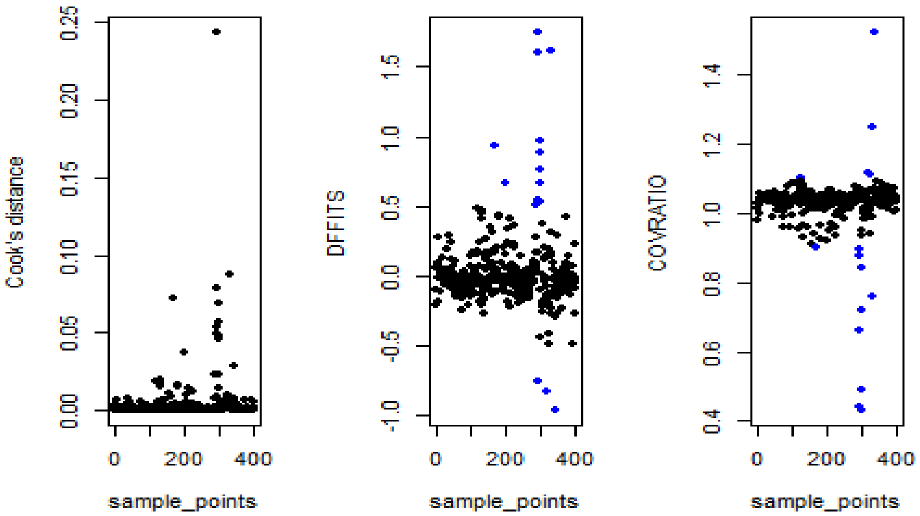


Figure 2: (left to right) Plots of Cook's distance, DFFITS and COVRATIO with the points suspected to be influential in blue

Determining the actual influential observations: Now, to find the points which have significant influence, we will use all these measures.
First,observe that the values of cook's distance are far away from the cut-off point. Hence,from Cook's distance, we cannot conclude anything. We will thus use the rest of the three measures available at our hands to find the influential points.
We take the intersection of those points which are influential according to DFFITS and COVRATIO. From these , we get 33 sample points which are suspected as influential.
Now, we look at the values of the DFBETAS for these 33 observations. We say that an observation is influential if the value of the DFBETAS for that particular point exceeds our cut-off for atleast 7 regression coefficients.
After computation, we find that the observations corresponding to the sample points 293, 295, 296, 297, 299, 302, 332 does satisfy this criterion and thus these are influential points.
The plots of Cook's distance, DFFITS and COVRATIO with the influential points are given in Figure 3.

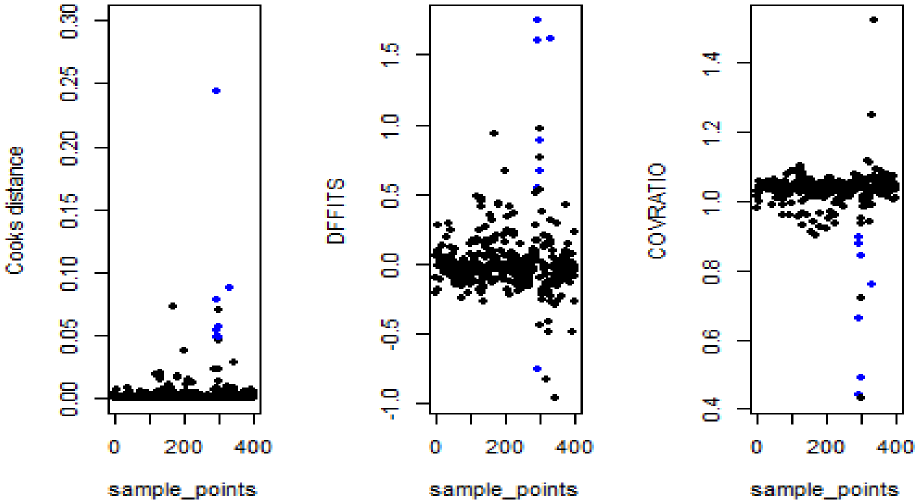


Figure 3: (left to right) Plots of Cook's distance,DFFITS and COVRATIO with the influential points in blue

After dropping the influential points, we calculate the RMSE of $\boldsymbol{y}_{test}$ again which comes out to be **4.255539**, less than the previous RMSE.
Thus, we keep this new data set with the leverage and influential points removed and continue with our analysis. Again, for simplicity, we use the notation $(X, \boldsymbol{y})$ for our new dataset hence obtaineed.

Dealing with Curvatures

We plot the residuals against each regressor and check for patterns that might indicate any deviation from the linear behaviour.

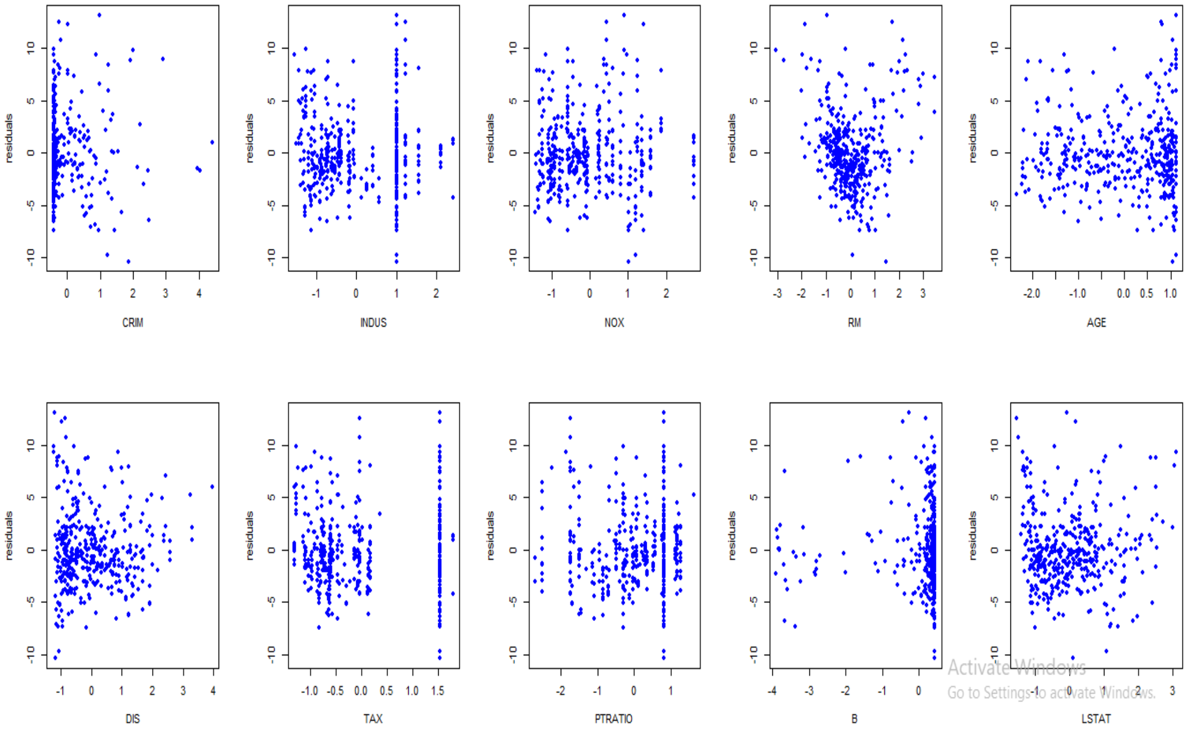
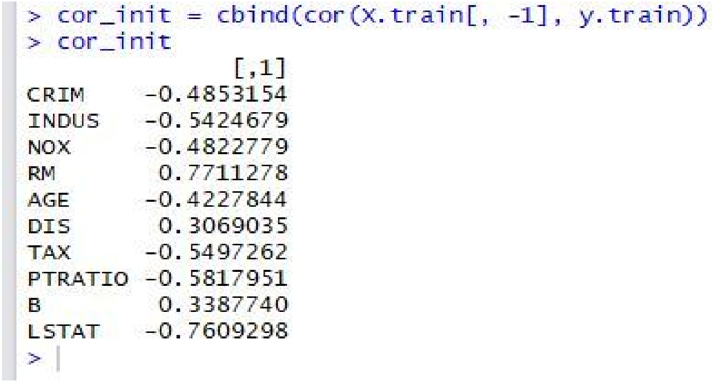


Figure 4: Residuals v/s Regressors Plot

From the above plot, it is evident that residuals are not entirely random when plotted against some regressors such as CRIM, RM, AGE, DIS, B, LSTAT and further investigation may be required. Apart from this, we also look at the following correlation table between the regressors and the response variable, which indicates a small correlation for



```
> cor_init = cbind(cor(X.train[, -1], y.train))
> cor_init
              [,1]
CRIM    -0.4853154
INDUS   -0.5424679
NOX     -0.4822779
RM       0.7711278
AGE     -0.4227844
DIS      0.3069035
TAX     -0.5497262
PTRATIO -0.5817951
B        0.3387740
LSTAT   -0.7609298
> |
```

Figure 5: Correlation table between MEDV and other regressors

the regressors CRIM, AGE, DIS and B. This aligns with our suspicion from the residuals versus regressors plot above. So, we mark these regressors alonwith RM and LSTAT as suspects for further investigation.

Component plus residual plot: The CPR plot is the scatter plot of,

$$\mathbf{e}_{y|X} + \mathbf{x}_j b_j \text{ against } \mathbf{x}_j$$

and is effective when one wants to find the non-linearity of x_j in regression model. Here, $\mathbf{e}_{y|X}$ denotes the residuals for the model,

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

and b_j is the estimate of the regression parameter corresponding to $\mathbf{x}_j$.

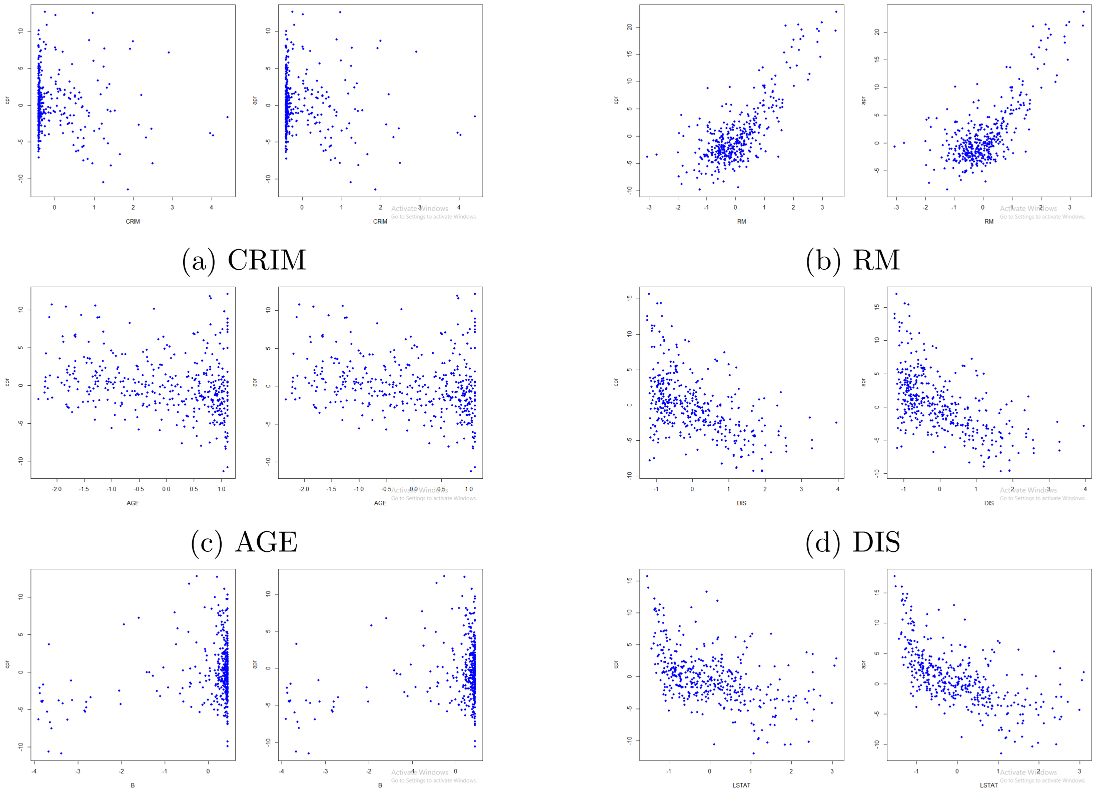
Augmented Partial residual plot: The APR plot is another display of regression diagnosis. It is the plot of

$$\mathbf{e}_{y|X^*} + \mathbf{x}_j b_j + \mathbf{x}_j^2 b_{jj} \text{ against } \mathbf{x}_j$$

where b_j and b_{jj} are the least squares estimates from the model

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{x}_j^2 b_{jj} + \epsilon; \qquad X^* = [X : \mathbf{x}_j^2]$$

The APR plot is more sensitive to non-linearity. In the absence of non-linearity the CPR plot is similar to the APR plot.
Below are the APR and CPR plots for the regressors identified as suspects.



(a) CRIM

(b) RM

(c) AGE

(d) DIS

(e) B

(f) LSTAT

Figure 6: The CPR - APR plots for the suspected regressors

At first glance, there does not appear much difference between the APR and CPR plots of any of the regressors. However, a close scrutiny reveals that there is some difference for regressors DIS, RM, LSTAT. This indicates the presence of non-linearity. Thus there is a need for a transformation. To identify which transformation to make, we use the Partial residual plots.

Partial Residual plot: Let X_{-j} be the remaining X after deleting the j-th column $\mathbf{x}_j$. Then partial Residual plot is the plot of,

$$\mathbf{e}_{y|X_{-j}} \text{ against } \mathbf{e}_{x_j|X_{-j}}$$

i.e. we have to plot the residuals obtained on regressing $\mathbf{y}$ against all the regressors except the j-th regressor versus the residuals obtained on regressing $\mathbf{x}_j$ against all other regressors. We use the following formula for calculating these residuals,

$$\mathbf{e}_{y|X_{-j}} = (I - P_{-j})\mathbf{y}$$

$$\mathbf{e}_{x_j|X_{-j}} = (I - P_{-j})\mathbf{x}_j$$

where $P_{-j} = X_{-j}(X_{-j}^T X_{-j})^{-1} X_{-j}^T$ is the projection matrix onto the column space of X_{-j}. If the relationship between $\mathbf{x}_j$ and $\mathbf{y}$ is linear, we expect the the partial residual plot to be a line passing through the origin. Below are the partial residual plots for the three regressor LSTAT.

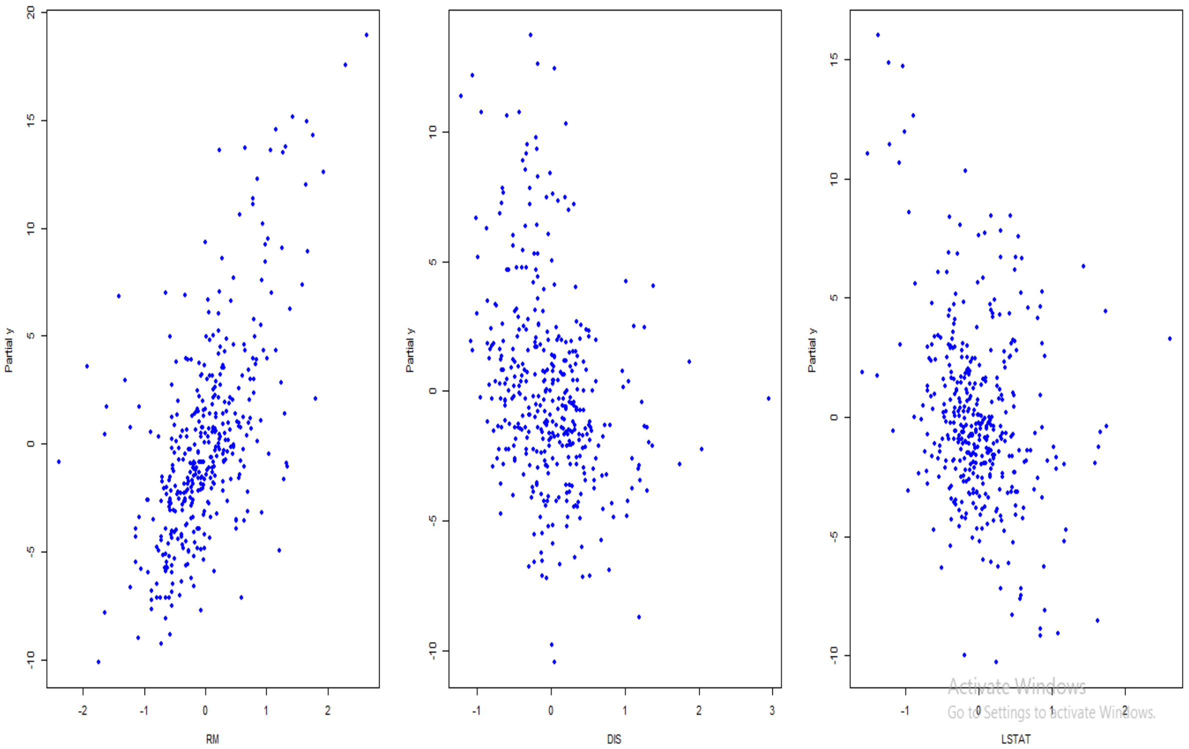
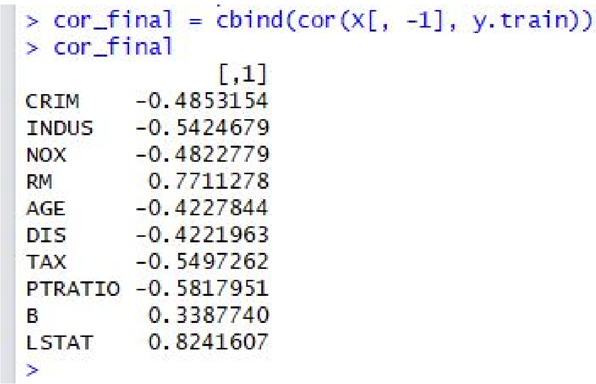


Figure 7: Partial residuals plot

The partial residual plot for RM is more less a straight line and we do not make any transformation for it. However, there seems a deviation from the straight line in the other two plots. We make the following transformation to DIS and LSTAT (chosen empirically)

$$\text{DIS} = exp(-3 * \text{DIS}) \ ; \qquad \text{LSTAT} = exp(-\text{LSTAT})$$

and check whether the correlation given in Figure 5 for both regressors has improved or not. And Figure 8 clearly indicates this.



```
> cor_final = cbind(cor(X[, -1], y.train))
> cor_final
              [,1]
CRIM    -0.4853154
INDUS   -0.5424679
NOX     -0.4822779
RM       0.7711278
AGE     -0.4227844
DIS     -0.4221963
TAX     -0.5497262
PTRATIO -0.5817951
B        0.3387740
LSTAT    0.8241607
>
```

Figure 8: Correlation table between MEDV and other regressors

Finally, we want to test whether this transformation also fits the test data better. To this end, we calculate the RMSE using the transformed regressors and it comes out to be 9.16. Thus, RMSE increases on making the transformation. Hence, we drop these transformation and proceed with the original matrix.

Dealing with Heteroscedasticity

We plot the residuals(studentized) against the fitted values and get the following Figure 11

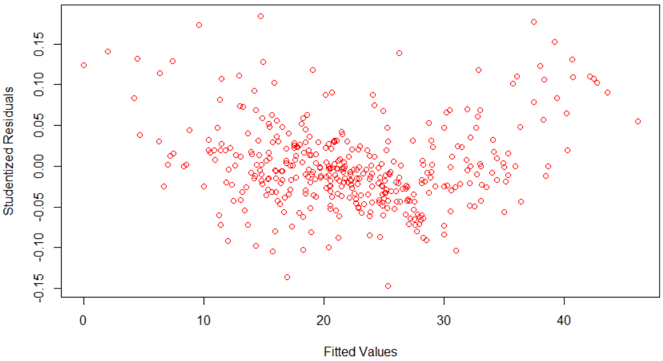


Figure 9: Studentized Residuals vs Fitted Values

Clearly, it can be seen that the graph of r_i v/s $\hat{y}_i$ does not show any constant pattern. Hence, we can roughly deduce that heteroscedasticity might be present.
Also, we plot the residuals against the individual regressors.We get the following figures.(Figure 9)

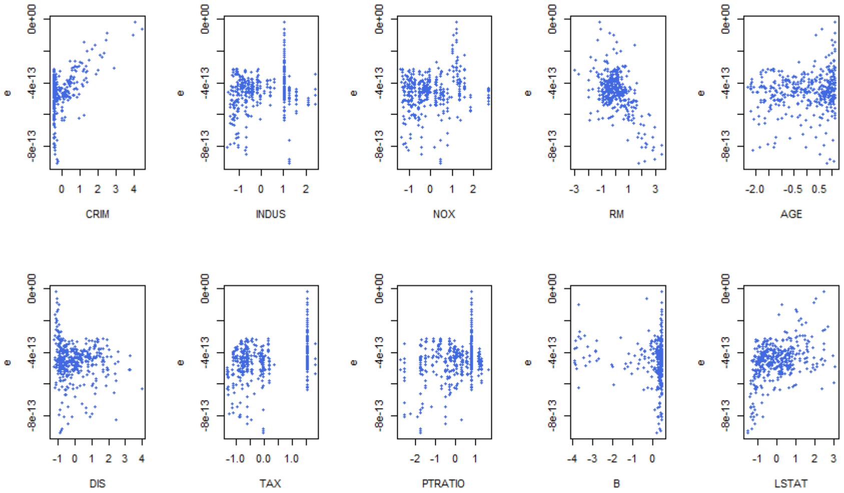


Figure 10: Residuals vs Regressors

Estimating σ_i^2 by e_i^2, we plot the squared residuals against the individual regressors. (Figure 11)

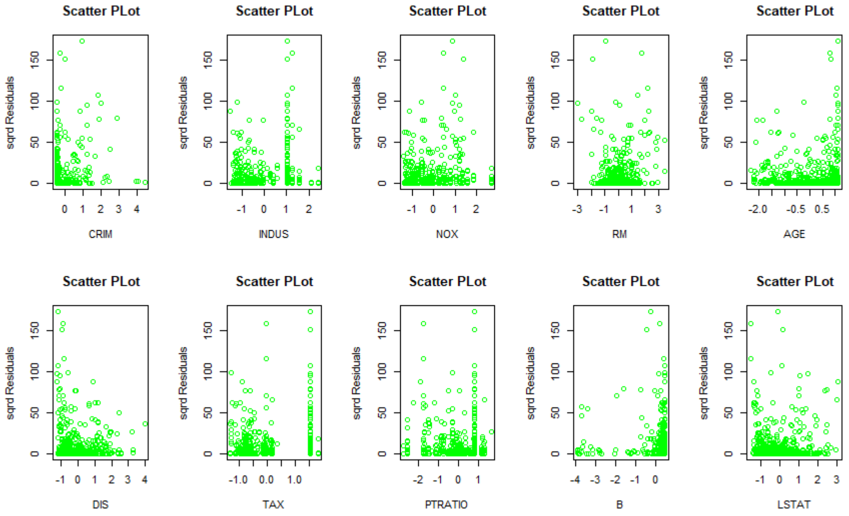


Figure 11: Squared Residuals vs Regressors

From (Figure 11), we infer that CRIM,RM,AGE,DIS,B,LSTAT show almost a non-constant pattern and are very likely the prime suspects for heteroscedasticity.

We now conduct the **Breusch-Pagan Test** for testing heteroscedasticity

among the suspected regressors.We form a new matrix $\mathbf{Z}$ whose columns are the vectors corresponding to **CRIM,RM,AGE,DIS,B,LSTAT**.

We are to test $\mathbf{H_0}$:All σ_i^2 are equal against $\mathbf{H_1}$:There exists at least one -inequality among σ_i^2's.

We calculate $d_i^* = \frac{n{e_i}^2}{\sum {e_i}^2}$

Then, we regress d_i^* upon $\mathbf{z_i}$'s and calculate the **Breusch Pagan statistic** $Q = nR^2$, where $R^2 = \frac{cov^2(d_i^*, \hat{d_i^*})}{V(d_i^*)V(\hat{d_i^*})}$

We know $Q \sim \chi_6^2$ (approx) as $n \to \infty$.
We have the observed value of Q from our data as 32.1 which is greater than $\chi^2_{6,0.05}$
So,**we may reject the null hypothesis that the errors are homoscedastic**.

From the previous part we see that the Breush-Pagan test rejects the homoscedasticity hypothesis at 5% level of significance.

We therefore consider an appropriate function $\sigma_i^2 = h(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ where we assume h to be a exponential function i.e. $\sigma_i^2 = h(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \exp(\alpha_0 + \sum_{j=1}^{n} \alpha_i z_{ji})$ and estimate σ_i^2 and $\hat{\boldsymbol{\beta}}$ simultaneously.
We apply logarithmic transformation to reduce the $h(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ to a linear model.
We take the initial value $\hat{\boldsymbol{\beta}}_{(0)}$ from the training data as $\hat{\boldsymbol{\beta}}_{(0)} = \hat{\boldsymbol{\beta}} = (\mathbf{X^T X})^{-1}\mathbf{X^T}\boldsymbol{y}_{train}$ where $\hat{\boldsymbol{\beta}}_{train}$ is the LSE of β_{train} after deleting the influential and leverage points in part (a).

$\mathbf{X}$: Design matrix of order 394 × 11
$\boldsymbol{y}_{train}$: vector of responses of order 394 × 1 obtained after removing the influential and leverage points.
We take $\hat{\boldsymbol{\alpha}}_{(0)} = (\mathbf{Z^TZ})^{-1}\mathbf{Z^T}\boldsymbol{d}$ where
$\mathbf{Z}$: matrix obtained in part (c) of order 394 × 6
$\boldsymbol{d} = (d_1, d_2, ..., d_n)$ where $d_i = \frac{\log(e_i^2)}{\frac{1}{n}\sum\limits_{i=1}^{n}\log(ei^2)}$ of order 394 × 1, $\boldsymbol{e} = (e_1, e_2, ..., e_n)$ is the vector of residuals. (Here $n = 394$)
We set $I = 1$
In the I^{th} iteration,we

1. update $\hat{\boldsymbol{\alpha}}_{(I)}$ by minimizing $S = \sum\limits_{i=1}^{n}(e_i^2 - h(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}))^2$

2. Using the updated $\hat{\boldsymbol{\alpha}}_{(I)}$ and $\hat{\boldsymbol{\beta}}_{(I-1)}$, update $\hat{\boldsymbol{\Sigma}}_{(I)} = diag(h_1, ..., h_n)$

3. Update $\hat{\beta}_{(I-1)} = (\mathbf{X^T}\hat{\boldsymbol{\Sigma}}_{(I)}^{-1}\mathbf{X})^{-1}\mathbf{X^T}\hat{\boldsymbol{\Sigma}}_{(I)}^{-1}\boldsymbol{y}$

4. $I = I + 1$

We break the loop when $(\|\hat{\boldsymbol{\beta}}_{(I)} - \hat{\boldsymbol{\beta}}_{(I-1)}\|^2 \leq 0.1$ and $\|\hat{\boldsymbol{\alpha}}_{(I)} - \hat{\boldsymbol{\alpha}}_{(I-1)}\|^2 \leq 0.1)$ or $I > 200$
After the iteration stops we collect the updated value of $\hat{\boldsymbol{\beta}}$ to calculate the new RMSE.

:Based on the this updated $\hat{\boldsymbol{\beta}}$ we get the required RMSE(updated)=**4.984998**.

RMSE has increased slightly from previous value, maybe because of the reason that we could not identify any subtle functional relationship between e^2 and z_j ,which is not evident from the plots in the naked eye.

Although the updated estimates deals with the heteroscedasticity it lowers the model performance.

Dealing with the Non-Normality

The R-Student residual is given by $r_i^* = \frac{e_i}{\sqrt{\frac{S_{(i)}^2}{n-p-1}(1-h_{ii})}}$, $i = 1(1)n$

Where, $S_{(i)}^2 = \sum_{j \neq i}(y_j - \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_{(i)})^2$
As, $e_i \sim N(0, (1-h_{ii})\sigma^2)$ and $\frac{S_{(i)}^2}{\sigma^2(n-p-1)} \sim \chi_{n-p-1}^2$ are independently distributed.
So, $r_i^* \sim t_{n-p-1}, \forall i$.
Now we draw the Q-Q plot of the R-Student residuals w.r.t. it's population distribution (which is t_{385} as n=396 and p=10).

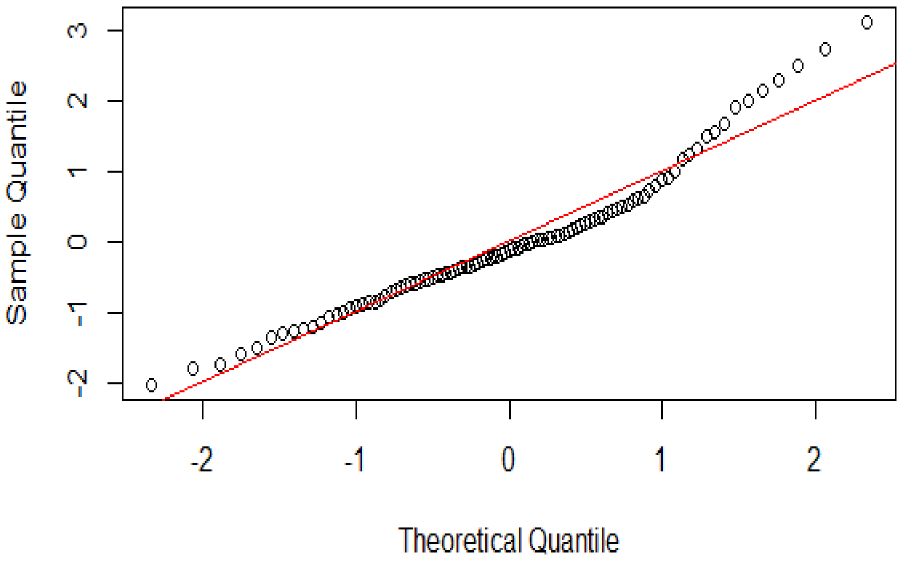


Figure 12: Q-Q plot of R-Student residuals

From the Q-Q plot(figure 12) we can conclude that the normality assumption is not correct.
Also the underlying distribution is positively skewed as it have longer right tail.

The **Box-Cox transformation** is given by, $u_\lambda = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{,if } \lambda \neq 0 \\ \log y & \text{,if } \lambda = 0 \end{cases}$

The profile log-liklihood of λ is given by,
$L(\lambda, \hat{\boldsymbol{\beta}}_{MLE}, \hat{\sigma}_{MLE}^2) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \frac{RSS_\lambda}{n} - \frac{n}{2} + n(\lambda - 1)\ln G$
where, $\hat{\boldsymbol{\beta}}_{MLE} = (X^TX)^{-1}X^T\boldsymbol{u}_\lambda$, $\hat{\sigma}_{MLE}^2 = \frac{RSS_\lambda}{n}$, $RSS\lambda = ||\boldsymbol{u}_\lambda - X\hat{\boldsymbol{\beta}}_{MLE}||^2$
and G is geometric mean of the response variable.
Now we plot the profile log-liklihood function of λ in the interval (-3,3) to find optimal value of λ.

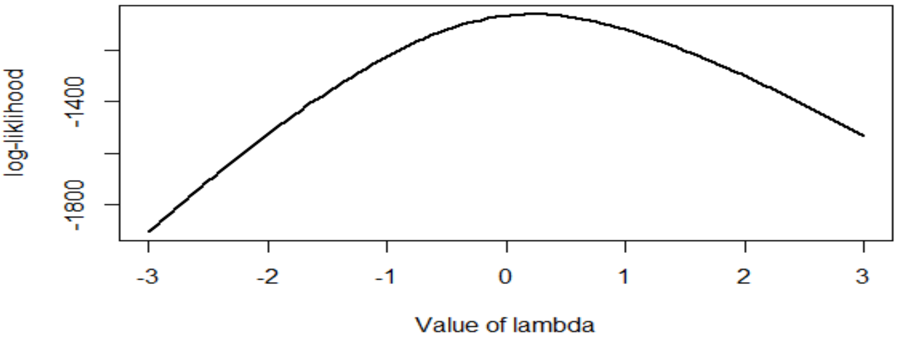


Figure 13: Profile Liklihood of λ

From the graph(figure 13) we can see that profile log-liklihood of λ is maximized approximately at $\lambda = \mathbf{0.21}$. So we choose the optimal value of $\lambda = \mathbf{0.21}$

As, $\lambda = 0.21$ using Box-Cox transformation we transform y to $\frac{y^{0.21}-1}{0.21}$. Now we draw the Q-Q plot of the R-Student residuals of the transformed variables w.r.t. it's population distribution (which is t_{385} as n=396 and p=10).

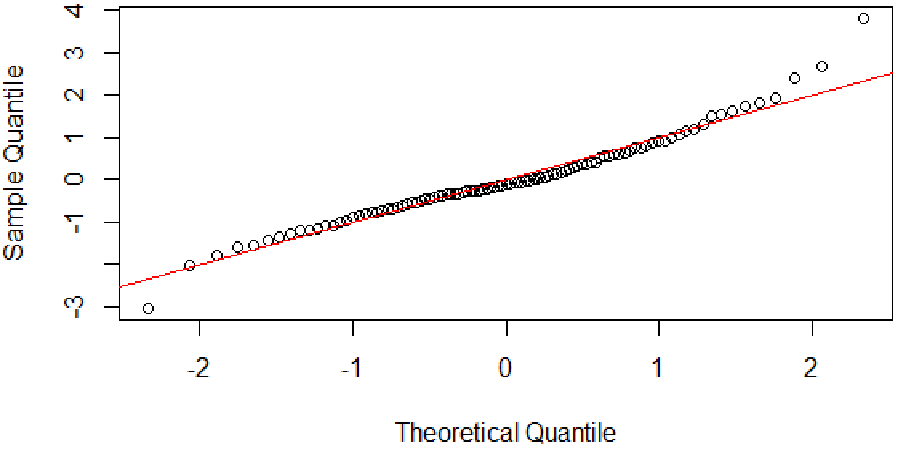


Figure 14: Q-Q plot of R-Student residuals of the transformed variable

From the Q-Q plot(figure 14) we can conclude that the normality assumption is almost correct on the transformed y. But there are very few points which have extreme R-Student values. So, the underlying distribution (although now it is symmetric) may have slightly thicker tails than it's theoretical distribution.