IMSE 514 – MULTIVARIATE STATISTICS
HOMEWORK 4

SURESH OOTY

Background:

13 Continuous variables and 1 binary variable are given and the data is about “Concerning house values in city suburbs”

Data Description:

Col	Heading	Description
1	CRIM	Crime rate per capita by town
2	LAZN	Proportion of residential land zoned for lots over 25000 sq. ft.
3	NRB	Proportion of non-retail business acres per town
4	CHR	CH River dummy variable (= 1 if tract bounds river; 0 otherwise)
5	NOC	Nitric oxides concentration (parts per 10 million)
6	RM	Average number of rooms per dwelling
7	AGE	Proportion of owner-occupied units built prior to 1940
8	WDIS	Weighted distances to five City employment centers
9	ARH	Index of accessibility to radial highways
10	PTAX	Full-value property-tax rate per $10000
11	PTT	Pupil-teacher ratio by town
12	B	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13	LSP	% lower status of the population
14	MEDHV	Median value of owner-occupied homes in $1000's

Exploratory Data Analysis on Variables:

The correlation was studied between the variables using R library “corrgram”. At first look, the negative & positive dimensions of the correlation matrix indicate that there are 2 group of variables in the set. The color codes indicate that these groups act opposite to each and act together within their group member.

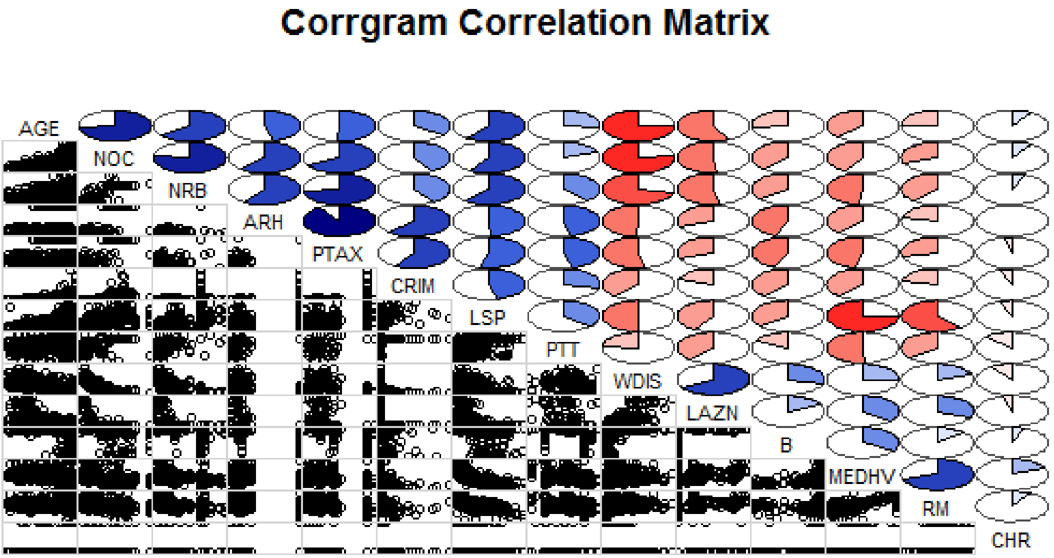


Figure 1: With all variables

The variable groups based on correlation matrix.

Table 1:

Group 1	Group 2	Neutral
AGE	WDIS	CHR (Boolean)
NOC	LAZN	
NRB	B	
ARH	MEDHV	
PTAX	RM	
CRIM		
LSP		
PTT		

Whereas, MEDHV – is the median house value which is the response variable.

The correlation matrix indicates that the house value could be influenced by LSP (% lower status of the population) and RM. However, the house value correlation with LSP indicate that the correlation is indirectly proportional. So, LSP data was chosen to transform to (100-LSP)/100 i.e., % higher status of population tLSP. After the transformation, the correlation became directly proportional. Further, a stronger correlation between ARH & PTAX was identified. ARH is an index and does have very less number of unique values; so it could be appropriate to drop this variable from consideration as a strong correlated variable PTAX is still under consideration. CHR is a Boolean value it has very weak correlation to any of the predicting variables and the response variable MEDHV. With the above consideration, the correlation matrix was reconstructed to see if any further detail could be noted.

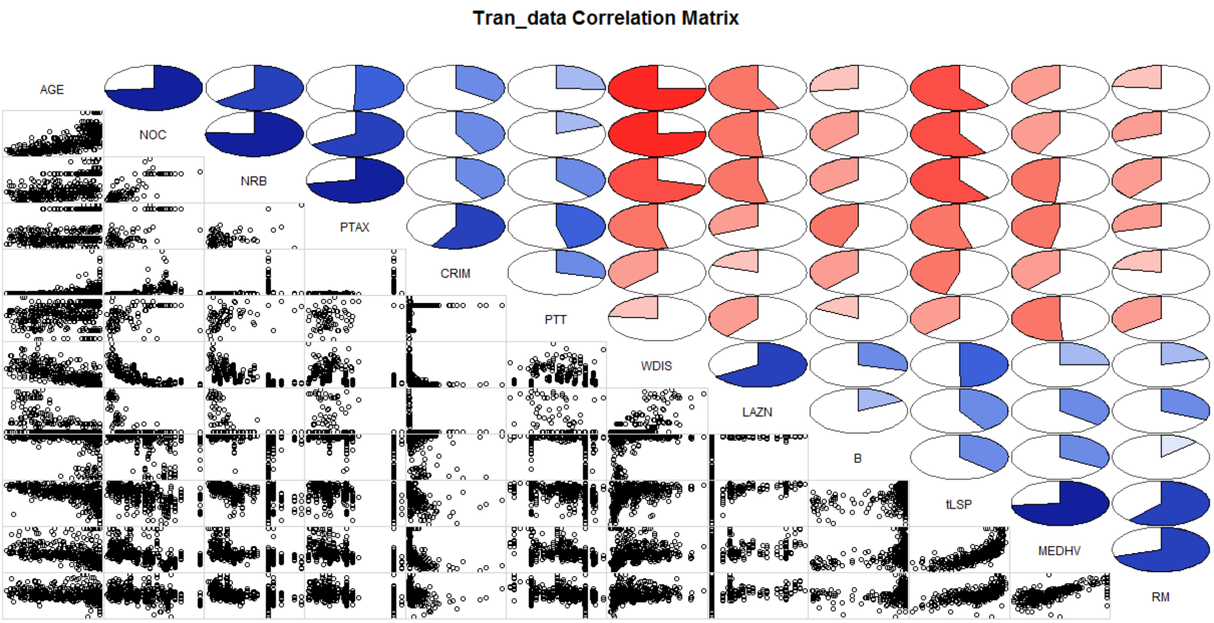


Figure 2: after transforming LSP

It could be noted from Figure 2, now the variable LSP is moved to group that is directly proportional to MEDHV. As per this new correlation data, The Table 1 could be restated as in the Table 2 below.

Table 2:

Group 1	Group 2	Neutral
AGE	WDIS	CHR (Boolean)
NOC	LAZN	
NRB	B	
ARH	MEDHV	
PTAX	RM	
CRIM	tLSP	
PTT		

A comparative study was done based on the variable CHR on these groups, to see if CHR had influenced any of the correlations.

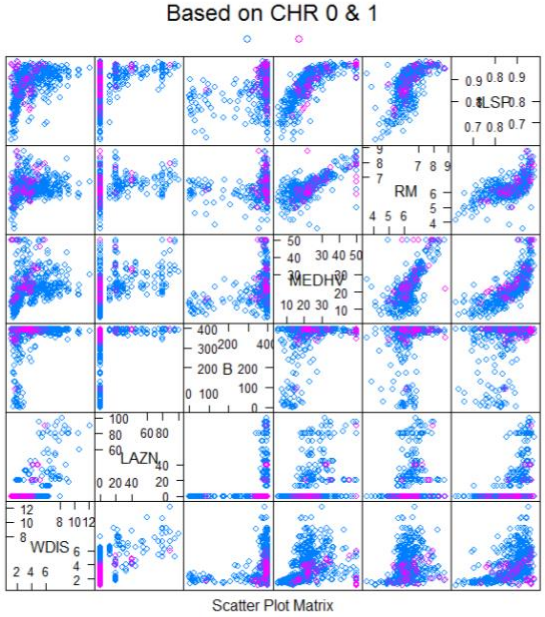


Figure 3: CHR influence on correlations in Group 2

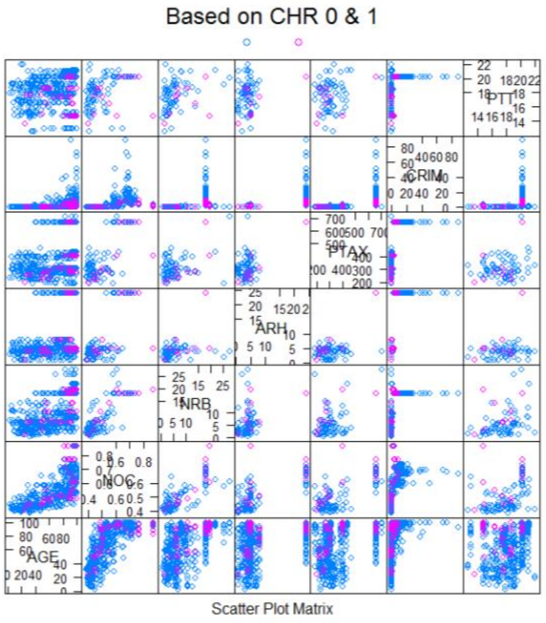


Figure 4: CHR influence on correlations in Group 1

So, it is evident that CHR need not be a significant factor in this study. So, this leads to table 3 as follows.

Table 3:

Group 1	Group 2
AGE	WDIS
NOC	LAZN
NRB	B
PTAX	MEDHV
CRIM	RM
PTT	tLSP

Further when the individual predicting variables (whose correlation is close to |0.5|) were compared against MEDHV in box plot methods, it was noted that higher CRIM did influence the house value. However, it was also noted that the variable B has no or minor impact on the CRIM variable and to the MEDHV variable.

Note: When B was transformed to Bk using the formula given (1000(Bk - 0.63)^2), No change was observed.

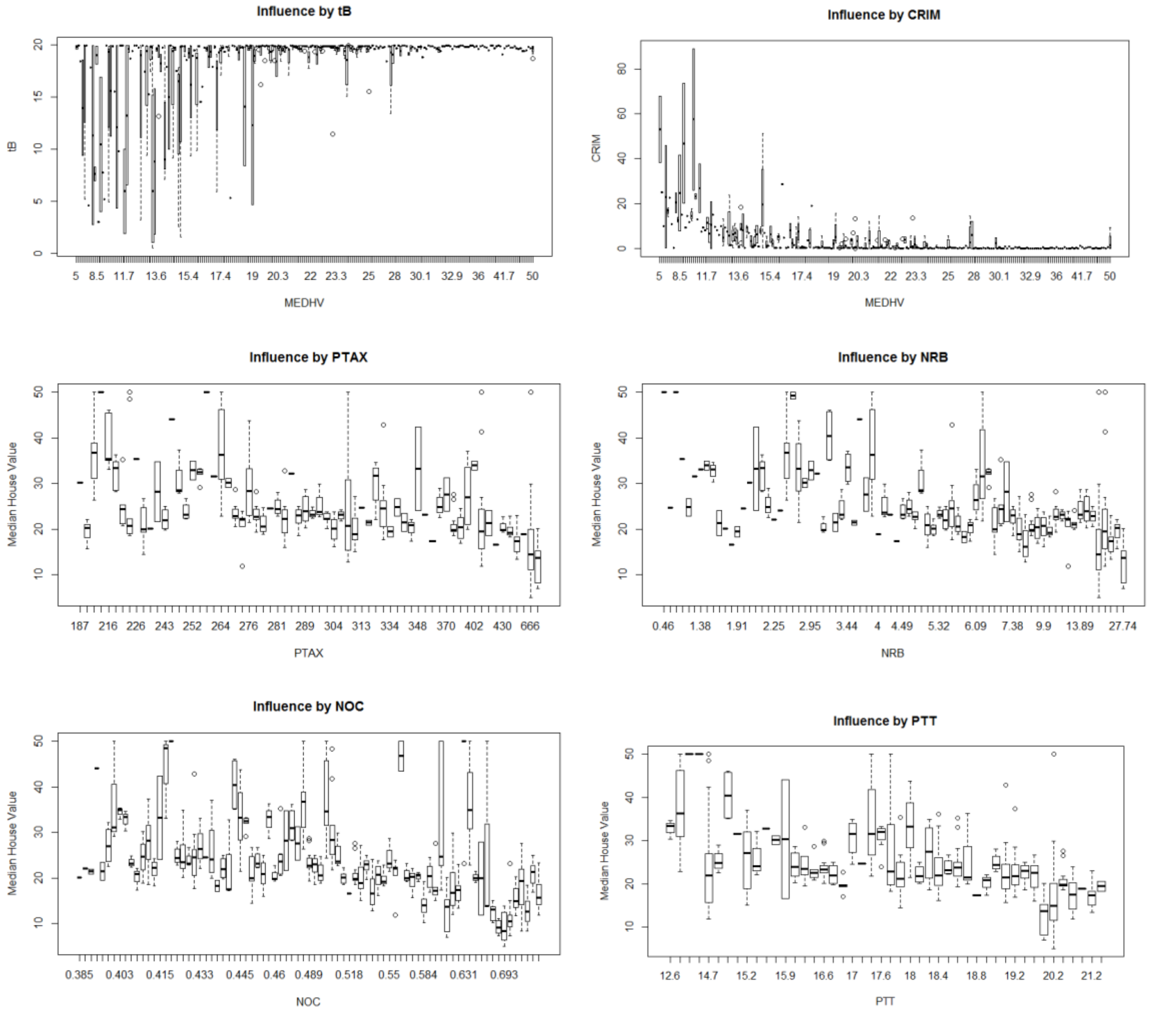


Figure 5: Individual variable (with cor value close to |0.5|) plots against MEDHV

To avoid the risk of dropping significant interactions among these variables, these variables are carried forward.

Applying PCA:

As the values are not in the similar range or unit, a Correlation method for PCA is appropriate. When the following R codes were executed, PCA supports to choose up to 3 Principal components that have significant influence on the model.

```
>pset<-cbind(CRIM,LAZN,NRB,NOC,RM,AGE,WDIS,PTAX,PTT,B,tLSP,MEDHV)
>cormat<-cor(pset)
>eigenX<-eigen(cormat)
>eigenX
```

```
$values
 [1] 5.9879999 1.5173857 1.1365994 0.8113500 0.6437476 0.5227496 0.4013911 0.2555344 0.2141050
[10] 0.1908536 0.1823074 0.1359763

$vectors
            [,1]         [,2]         [,3]
 [1,]  0.2423209  0.053025869  0.491796770      CRIM
 [2,] -0.2670150  0.163440022  0.448687075      LAZN
 [3,]  0.3491091 -0.131207420 -0.026927754      NRB
 [4,]  0.3406941 -0.274870679  0.008333434      NOC
 [5,] -0.2236836 -0.497372266  0.276340008      RM
 [6,]  0.3174623 -0.284958869 -0.164560865      AGE
 [7,] -0.3164082  0.399144533  0.150960012      WDIS
 [8,]  0.3210346 -0.021170437  0.328468644      PTAX
 [9,]  0.2129928  0.336651500 -0.085334498      PTT
[10,] -0.2002495 -0.001195848 -0.556642044      B
[11,] -0.3323154 -0.223781835  0.042645453      tLSP
[12,] -0.2866017 -0.475338924  0.054273982      MEDHV
```

The plot for the same (elbow structure)

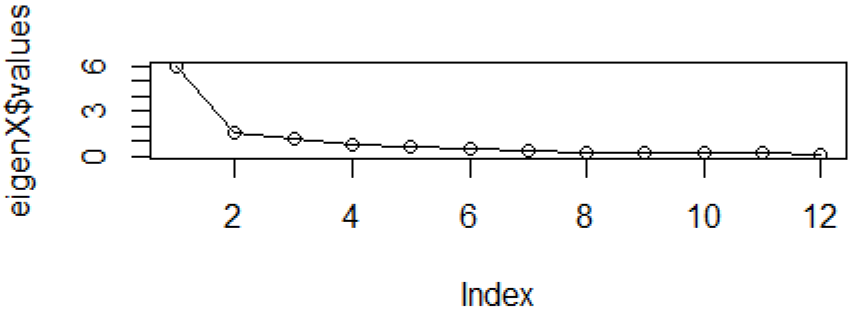


Figure 6: PCA Elbow diagram

The elbow analysis on the Eigen values show that the first 3 factors could be considered.

Model Construction using PCA:

Construction of the model with 3 PCs.

```
>newX<-X%*%eigenX$vectors
>newmodel<-lm(MEDHV~newX[,1]+newX[,2]+newX[,3])
>pressNew<-resid(newmodel)/(1-lm.influence(newmodel)$hat)
>prN<-resid(newmodel)/(1-lm.influence(newmodel)$hat)
>pressNew<-sum(prN^2)
>pressNew
[1] 16266.67
```

And the predicted Rsquare values is

```
>predR<-1-pressNew/sum(MEDHV-mean(MEDHV)^2)
>predR

[1] 0.619193
```

```
>summary(newmodel)

Call:
lm(formula = MEDHV ~ newX[, 1] + newX[, 2] + newX[, 3])

Residuals:
     Min       1Q   Median       3Q      Max 
-14.5852  -3.6818  -0.4268   3.1202  21.9641 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 30.395171   1.035122   29.36   <2e-16 ***
newX[, 1]   -0.298440   0.012318  -24.23   <2e-16 ***
newX[, 2]   -0.738003   0.036250  -20.36   <2e-16 ***
newX[, 3]    0.149799   0.008722   17.18   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.633 on 502 degrees of freedom
Multiple R-squared:  0.6271,	Adjusted R-squared:  0.6249 
F-statistic: 281.4 on 3 and 502 DF,  p-value: < 2.2e-16
```

As the PRESS value appears too high. A second predicted model was constructed using the dropped variable ARH. For which, the eigen values are as follows :

```
$values
 [1] 6.54584988 1.52266177 1.33579036 0.86400373 0.66675157 0.53745686 0.40363954 0.27750369
 [9] 0.25344519 0.21286161 0.18326459 0.13597784 0.06079336
```

Again a 3 PC selection was made to construct the predicted model (2). For which the PRESS value and Predicted Rsquare values are as follows. The PRESS showed some improvement and predicted R square value improved from 0.62 to 0.86.

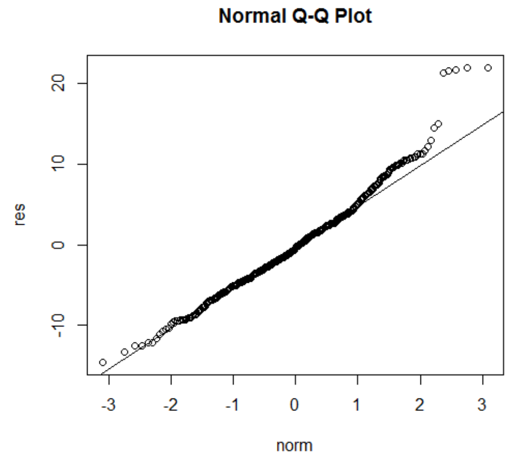
```
> pressNew_ARH
[1] 5946.249
> predR_ARH
[1] 0.8607967
> summary(newmodel_ARH)

Call:
lm(formula = MEDHV ~ newX_ARH[, 1] + newX_ARH[, 2] + newX_ARH[, 
    3])

Residuals:
    Min      1Q  Median      3Q     Max 
-9.4553 -2.3502 -0.0381  1.8340 11.5821 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   16.286581   0.536152   30.38   <2e-16 ***
newX_ARH[, 1] -0.480472   0.009651  -49.78   <2e-16 ***
newX_ARH[, 2] -1.143356   0.025700  -44.49   <2e-16 ***
newX_ARH[, 3]  0.207188   0.006162   33.62   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.407 on 502 degrees of freedom
Multiple R-squared:  0.8636, Adjusted R-squared:  0.8627 
F-statistic:  1059 on 3 and 502 DF,  p-value: < 2.2e-16
```

As a third iteration, the other dropped variable CHR was also brought back to the model to check for the PRESS value and predicted R. The PCA analysis for this iteration also suggested 3 PCs. Again, the Predicted R square and PRESS value improved.

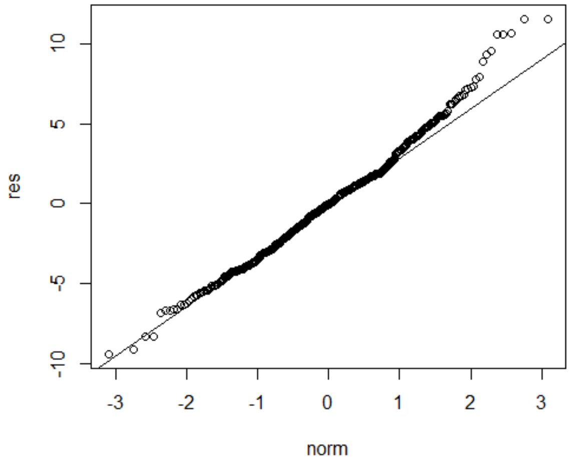
```
> pressNew_CHR
[1] 5301.326
> predR_CHR
[1] 0.8758945
> summary(newmodel_CHR)

Call:
lm(formula = MEDHV ~ newX_CHR[, 1] + newX_CHR[, 2] + newX_CHR[, 
    3])

Residuals:
    Min      1Q  Median      3Q     Max 
-8.1517 -2.2740 -0.1259  1.8246 10.6471 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   13.468722   0.548709   24.55   <2e-16 ***
newX_CHR[, 1] -0.477852   0.008990  -53.15   <2e-16 ***
newX_CHR[, 2] -1.187139   0.025126  -47.25   <2e-16 ***
newX_CHR[, 3]  0.378566   0.008323   45.48   <2e-16 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.218 on 502 degrees of freedom
Multiple R-squared:  0.8783, Adjusted R-squared:  0.8776
F-statistic:  1208 on 3 and 502 DF,  p-value: < 2.2e-16
```

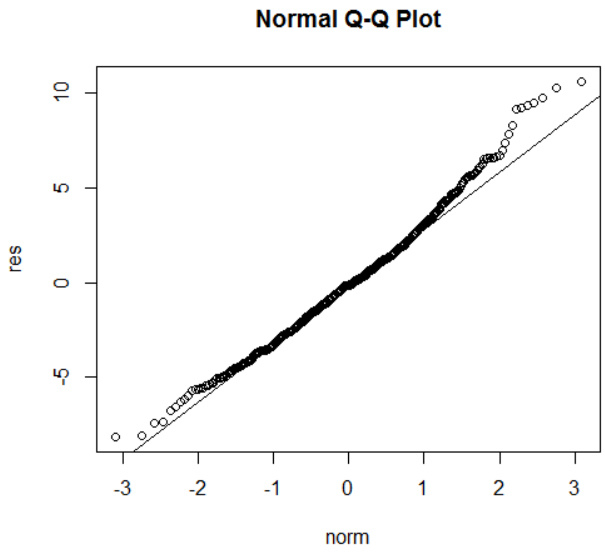
A Q-Q norm plot on these models was done for comparison.



The first model with ARH and CHR removed show some anomaly at the top right corner of the residual plot.



The second model with ARH shows some convergence at the top right corner.



This is the final model with both ARH and CHR considered for the PCA test. Though there is some outlier indication at the top right corner, this plot is comparatively better than the other two.

Model Construction using Stepwise

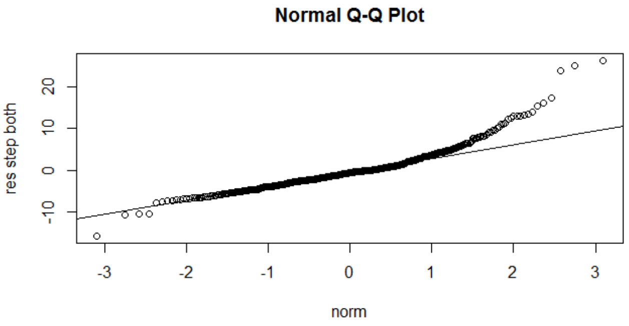
Data: The only transformed data was the tLSP, which was carried over to this iteration. None of the other variables were dropped.

During stepwise model construction, Two models were tried using “BOTH” direction and “BACKWARD” direction approaches.

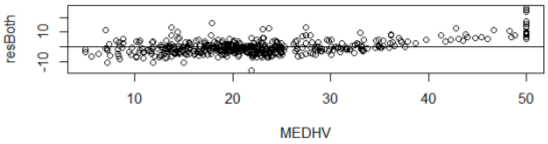
BOTH approach:

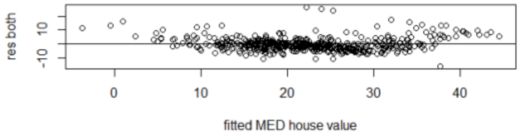
This approach retained all 12 variables except for the AGE variable.

```
Call:
lm(formula = MEDHV ~ tLSP + RM + PTT + WDIS + NOC + CHR +
B +
    LAZN + CRIM + ARH + PTAX, data = pset)

Residuals:
     Min       1Q   Median       3Q      Max
-15.5984  -2.7386  -0.5046   1.7273  26.2373

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.914201   5.868545  -2.712 0.006926 **
```



```
tLSP         52.255346   4.742436  11.019  < 2e-16 ***
RM            3.801579   0.406316   9.356  < 2e-16 ***
PTT          -0.946525   0.129066  -7.334 9.24e-13 ***
WDIS         -1.492711   0.185731  -8.037 6.84e-15 ***
NOC         -17.376023   3.535243  -4.915 1.21e-06 ***
CHR           2.718716   0.854240   3.183 0.001551 **
B             0.009291   0.002674   3.475 0.000557 ***
LAZN          0.045845   0.013523   3.390 0.000754 ***
CRIM         -0.108413   0.032779  -3.307 0.001010 **
ARH           0.299608   0.063402   4.726 3.00e-06 ***
PTAX         -0.011778   0.003372  -3.493 0.000521 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.736 on 494 degrees of freedom
Multiple R-squared:  0.7406, Adjusted R-squared:  0.7348
F-statistic: 128.2 on 11 and 494 DF,  p-value: < 2.2e-16
```





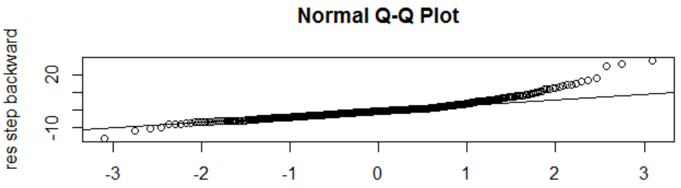
Observation on the plots:

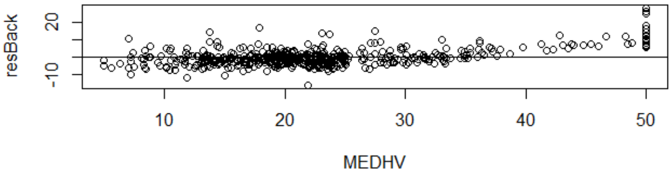
From the QQ plot and the residual plot against response, it is noted that there could be outliers when the MEDHV (house value) is larger (50)

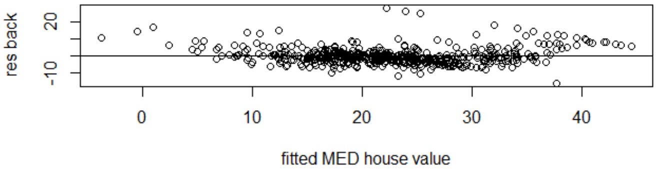
BACKWARD approach:

The backward approach was not better than BOTH approach. When compared with R-squared value and F-statistic.

```
Call:
lm(formula = MEDHV ~ CRIM + LAZN + NOC + RM + WDIS +
PTT + B +
    tLSP + factor(CHR), data = pset)

Residuals:
    Min      1Q  Median      3Q     Max
-15.803  -2.832  -0.625   1.454  27.766

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -22.992416   5.744907  -4.002 7.23e-05 **
*
CRIM          -0.061174   0.030377  -2.014 0.044567 *
LAZN           0.042032   0.013422   3.131 0.001842 **
NOC          -16.088513   3.232702  -4.977 8.93e-07 **
*
RM             4.149667   0.407685  10.179  < 2e-16 **
*
WDIS          -1.431665   0.188603  -7.591 1.59e-13 ***
PTT           -0.838640   0.117342  -7.147 3.19e-12 ***
B              0.008292   0.002688   3.084 0.002153 **
tLSP          52.500413   4.835123  10.858  < 2e-16 ***
factor(CHR)1   3.029924   0.868349   3.489 0.000527 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.
1 ‘ ’ 1
```







```
Residual standard error: 4.833 on 496 degrees of freedom
Multiple R-squared:  0.7288, Adjusted R-squared:  0.7239
F-statistic: 148.1 on 9 and 496 DF,  p-value: < 2.2e-16
```

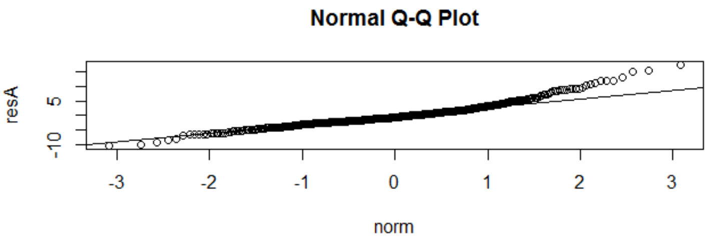
Observation on the plots:

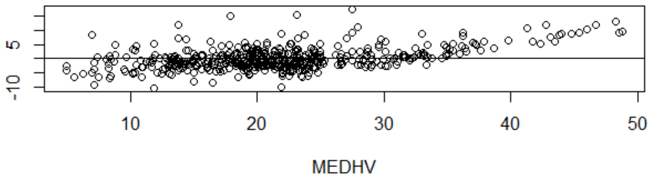
Just like the BOTH direction model it is noted that there could be outliers when the MEDHV (house value) is larger (50)

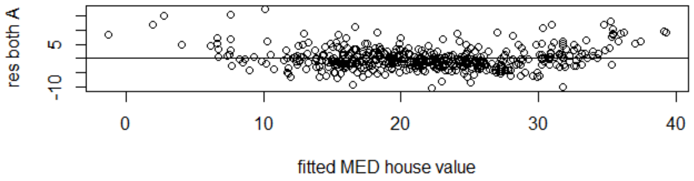
So, a rerun was carried over after dropping the values for MEDHV around 50. After dropping these outliers, on applying the stepwise “BOTH” approach the model became,

MODEL A:

```
Call:
lm(formula = hw4T$MEDHV ~ hw4T$tLSP + hw4T$RM + hw4T$PTT + hw4T$PTAX +
    hw4T$B + hw4T$WDIS + hw4T$NOC + hw4T$AGE + hw4T$LAZN + hw4T$CRIM +
    hw4T$NRB, data = hw4T)

Residuals:
    Min      1Q  Median      3Q     Max
-10.309  -2.311  -0.633   1.669  17.433

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.482541   4.873265  -1.535 0.125341
hw4T$tLSP    33.963798   4.341203   7.824 3.32e-14
***
hw4T$RM       4.067198   0.359546  11.312  < 2e-16
***
hw4T$PTT     -0.759537   0.105810  -7.178 2.71e-12 **
*
hw4T$PTAX    -0.002853   0.001899  -1.502 0.133753
hw4T$B        0.007194   0.002171   3.314 0.000991 **
*
hw4T$WDIS    -1.208875   0.164238  -7.360 8.06e-13 **
*
hw4T$NOC    -10.080874   3.084161  -3.269 0.001159 *
*
hw4T$AGE     -0.027475   0.010855  -2.531 0.011690 *
hw4T$LAZN     0.028492   0.011431   2.492 0.013024 *
hw4T$CRIM    -0.073739   0.025670  -2.873 0.004252 *
*
hw4T$NRB     -0.102194   0.049159  -2.079 0.038163 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1

Residual standard error: 3.843 on 478 degrees of freedom
Multiple R-squared:  0.7666, Adjusted R-squared:  0.7613
F-statistic: 142.7 on 11 and 478 DF,  p-value: < 2.2e-16
```



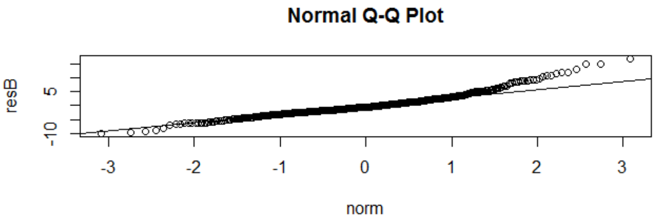


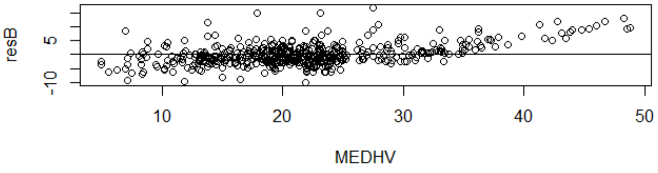


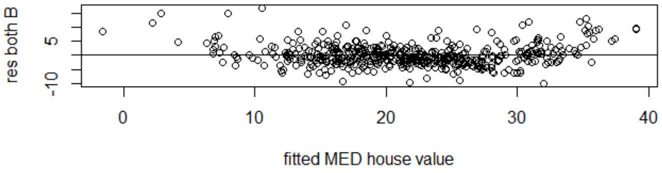
PTAX was not significant, so this variable was dropped to reconstruct the model.

MODEL B :

```
Call:
lm(formula = hw4T$MEDHV ~ hw4T$tLSP + hw4T$RM + hw4T$P
TT + hw4T$B +
    hw4T$WDIS + hw4T$NOC + hw4T$AGE + hw4T$LAZN + hw4T
$CRIM +
    hw4T$NRB, data = hw4T)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0587  -2.2841  -0.5573   1.7189  16.9603

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.681215   4.850322  -1.377 0.169008
hw4T$tLSP    34.422741   4.336110   7.939 1.46e-14 ***
hw4T$RM       4.008188   0.357861  11.200  < 2e-16 ***
hw4T$PTT     -0.816823   0.098826  -8.265 1.38e-15 ***
hw4T$B        0.007655   0.002152   3.557 0.000413 ***
hw4T$WDIS    -1.213023   0.164430  -7.377 7.19e-13 ***
hw4T$NOC    -11.501437   2.939416  -3.913 0.000104 ***
hw4T$AGE     -0.026277   0.010840  -2.424 0.015713 *
hw4T$LAZN     0.024725   0.011167   2.214 0.027296 *
hw4T$CRIM    -0.086645   0.024221  -3.577 0.000382 ***
hw4T$NRB     -0.128177   0.046075  -2.782 0.005617 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.
1 ‘ ’ 1

Residual standard error: 3.848 on 479 degrees of freedo
m
Multiple R-squared:  0.7655, Adjusted R-squared:  0.760
6
F-statistic: 156.4 on 10 and 479 DF,  p-value: < 2.2e-16
```







In this second iteration, after dropping the outliers around MEDHV = 50, the plots show no anomaly at the right top corner. However, among these two models, the model from BOTH direction was better (R-squared value 0.7613). Any trial on dropping the one of the variables did not improve the model.

Model Comparison:

Calculating PRESS values for STEPWISE constructed models.

```
> prA<-resA/(1-lm.influence(stpbth)$hat)
> pressA<-sum(prA^2)
> pressA
[1] 7611.139

R-square – 0.7613
```

```
> prB<-resB/(1-lm.influence(drpmdl)$hat)
> pressB<-sum(prB^2)
> pressB
[1] 7604.929

R-square -0.7606
```

Table of comparison based on PRESS & R-squre value.

	PRESS	R-Squared
Step-Both	7611.1	0.7613
Step-Both(dropped PTAX)	7604.9	0.7606
PCA (noARH/noCHR)	16267	0.619193
PCA-with ARH	5946.2	0.8607967
PCA-ARH_CHR	5301.3	0.8758945

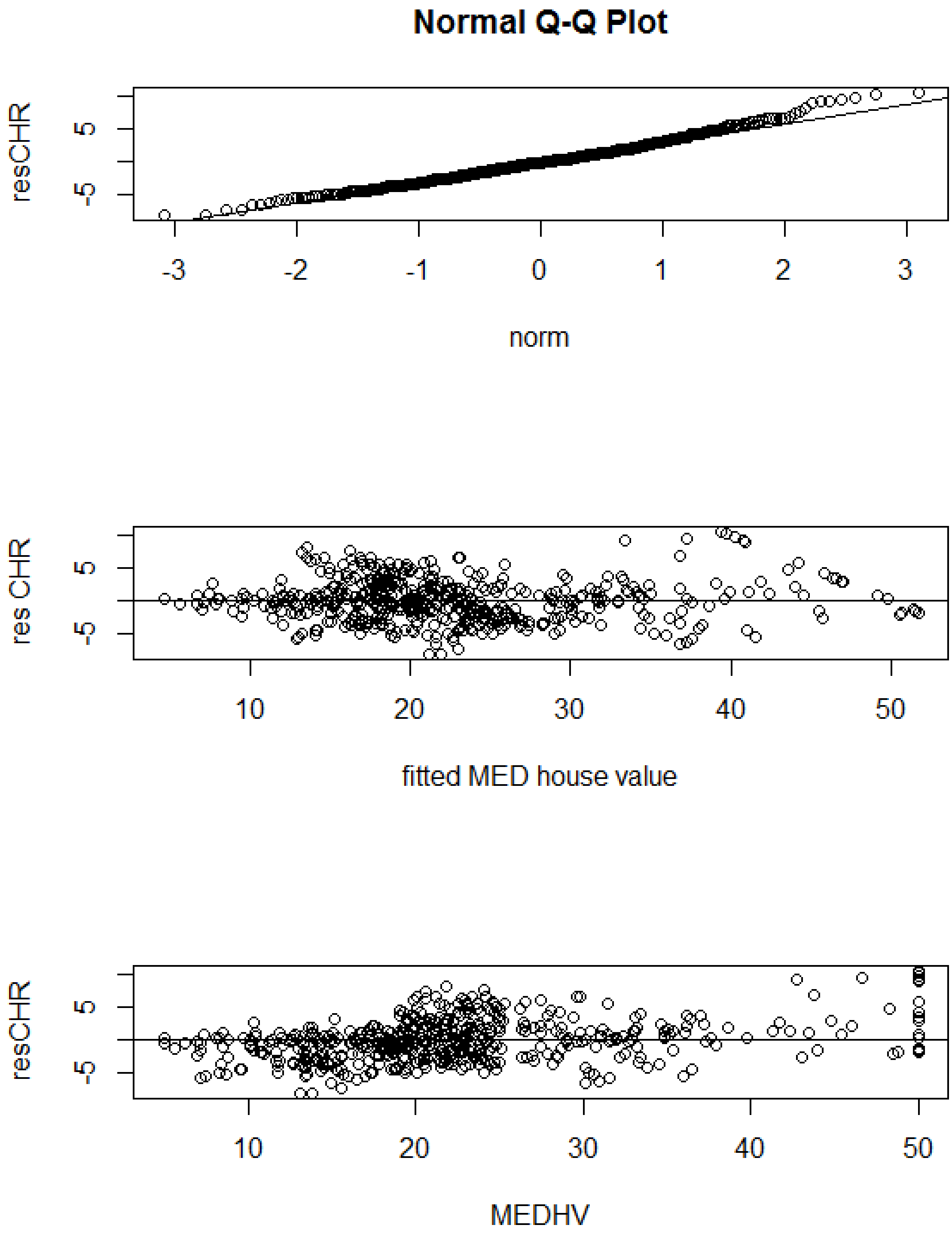
Comparing the models using ANOVA:

Note: As the Step-both & Step-both (dropped PTAX) models have been created after dropping some outliers. A direct comparison was not possible against the PCA_ARH_CHR which had more data points. Hence, a comparison was made against those models that had all the data points. Here is the summary,

```
> anova(newmodel_CHR,stpbth,test="Chisq")
Analysis of Variance Table

Model 1: MEDHV ~ newX_CHR[, 1] + newX_CHR[, 2] + newX_CHR[, 3]
Model 2: MEDHV ~ tLSP + RM + PTT + WDIS + NOC + CHR + B + LAZN + CRIM +
    ARH + PTAX
  Res.Df   RSS Df Sum of Sq Pr(>Chi)
1    502  5199
2    494 11081  8   -5882.4

> anova(newmodel_CHR,stpbk,test="Chisq")
Analysis of Variance Table

Model 1: MEDHV ~ newX_CHR[, 1] + newX_CHR[, 2] + newX_CHR[, 3]
Model 2: MEDHV ~ CRIM + LAZN + NOC + RM + WDIS + PTT + B + tLSP + factor(CHR)
  Res.Df   RSS Df Sum of Sq Pr(>Chi)
1    502  5199
2    496 11584  6   -6384.6
```

The Residual sum of squares for the PCA approach model was less and hence is a better model.

Residual plots for the chosen model:



Conclusion:

The residual plots indicate the slight anomaly at the high MEDHV values, suggest further data transformation by dropping the outliers.

Further iterations could be made by dropping data points based on outlier validation and individual box plot method to identify outliers for each variable. This could be an extensive effort that might lead to further convergence of values.