U.S. Demographic Data Regression Analysis

Team Members:

Name	Roll No.
Pawan Kumar Amat	195280007
Akshi Bansal	195280009
Lovish Kandoi	195280014
Devansh Adhikari	195280019

1. Data Description:

We are considering a dataset providing some county demographic information (CDI) for 440 of the most populous counties in the United States in years 1990–92. Each line of the dataset provides information on 14 variables for a single county.

The dataset can be found **here**.

Here are the definitions for the variables considered in the population for which country demographic information (CDI) are available.

Variable		Description
id		identification number, 1–440
county		county name
state		state abbreviation
area	X1	land area (square miles)
popul	X2	estimated 1990 population
pop1834	X3	percent of 1990 CDI population aged 18–34
pop65plus	X4	percent of 1990 CDI population aged 65 years old or older
phys	Y	number of professionally active nonfederal physicians during 1990
beds	X5	total number of beds, cribs and bassinets during 1990
crimes	X6	total number of serious crimes in 1990 (including murder, rape, robbery, aggravated assault, burglary, larceny-theft, motor vehicle theft)
higrads	X7	percent of adults (25 yrs old or older) who

		completed at least 12 years of school
bachelors	X8	percent of adults (25 yrs old or older) with bachelor's degree
poors	X9	Percent of 1990 CDI population with income below poverty level
unemployed	X10	percent of 1990 CDI labor force which is unemployed
percapitaincome	X11	per capita income of 1990 CDI population (dollars)
totalincome	X12	total personal income of 1990 CDI population (in millions of dollars)
region	X13	Geographic region classification used by the U.S. Bureau of the Census, where 1 = Northeast, 2 = Midwest, 3 = South, 4 = West

The goal is to model the number of physicians (y) per 1000 inhabitants, using the other demographic variables.

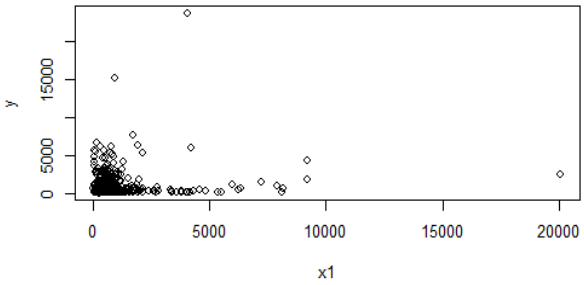
There are total 13 covariates in which:

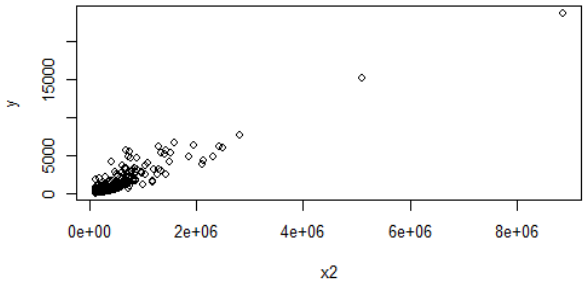
- 12 variables(X1, X2… X12) are numerical.
- 1 variable, i.e., X13 is categorical, having 4 categories.

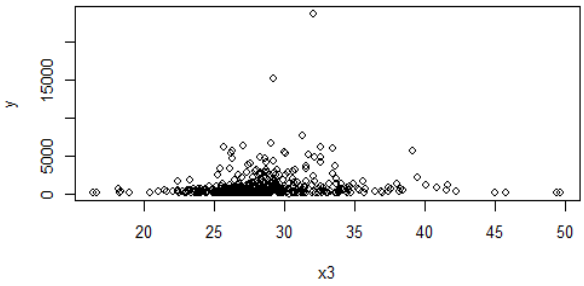
District and State names are not taken into consideration.

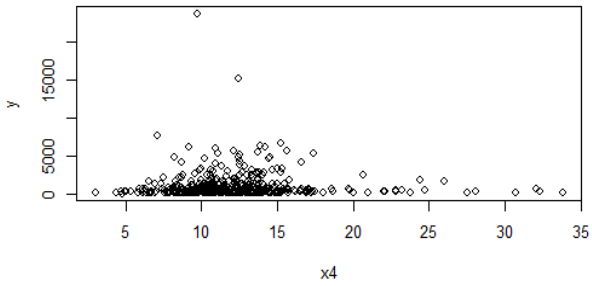
Variable X13 has been converted to 4 dummy variables, namely X13_1, X13_2, X13_3 & X13_4 for categories 1, 2, 3, & 4 respectively. These 1,2,3 & 4 values behaved as weights assigned to different regions. As the assigned weights are random this will hamper the interpretation of the coefficient of the X13 variable. To deal with this the variable is split into 4 binary variables. Each one of them indicating the presence/ non presence of a region.

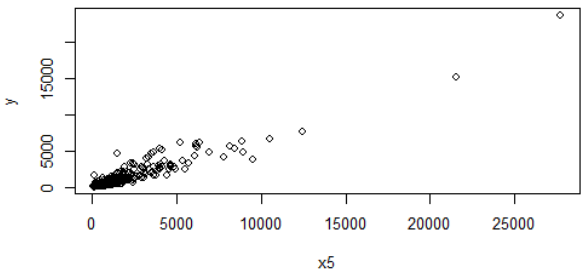
1.1 Scatter plots of Predictor variables vs. Response variable

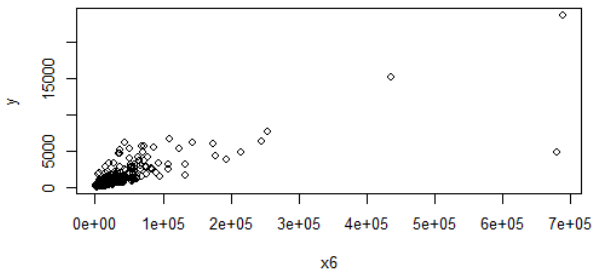


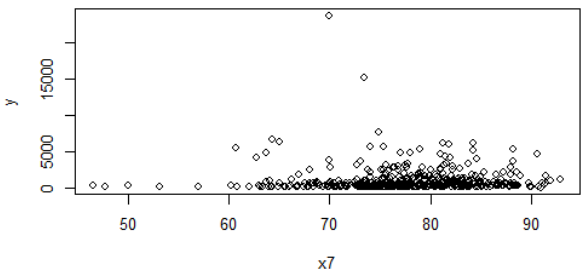


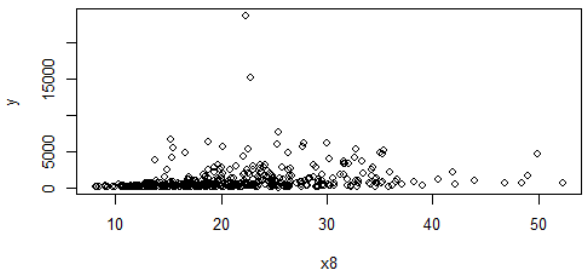


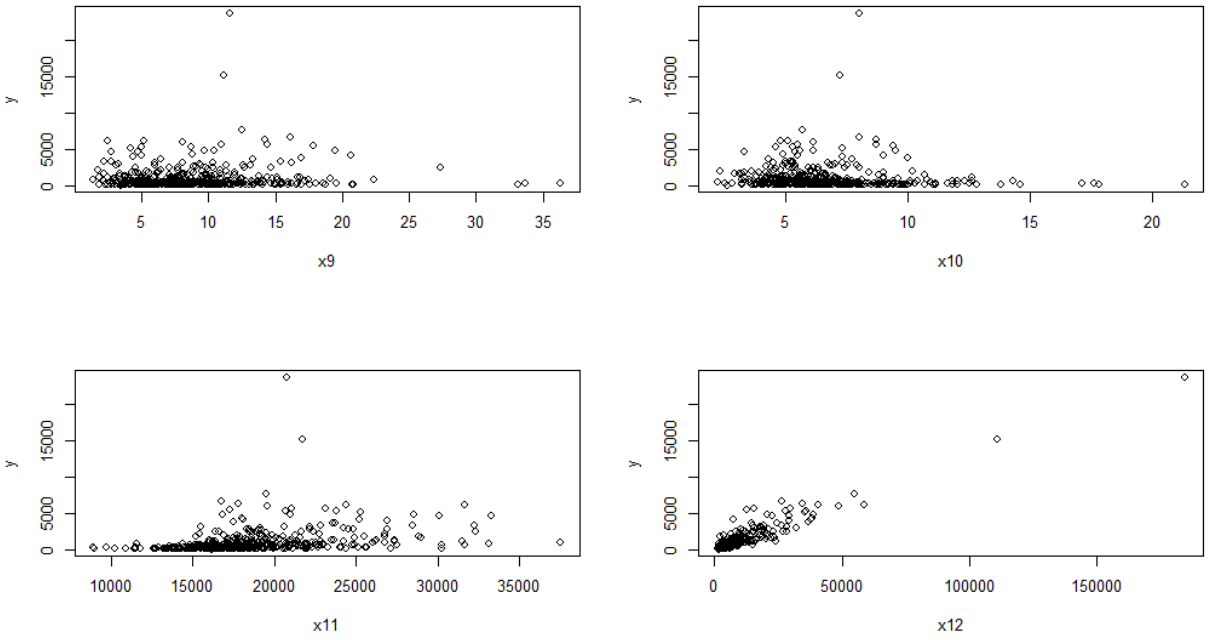




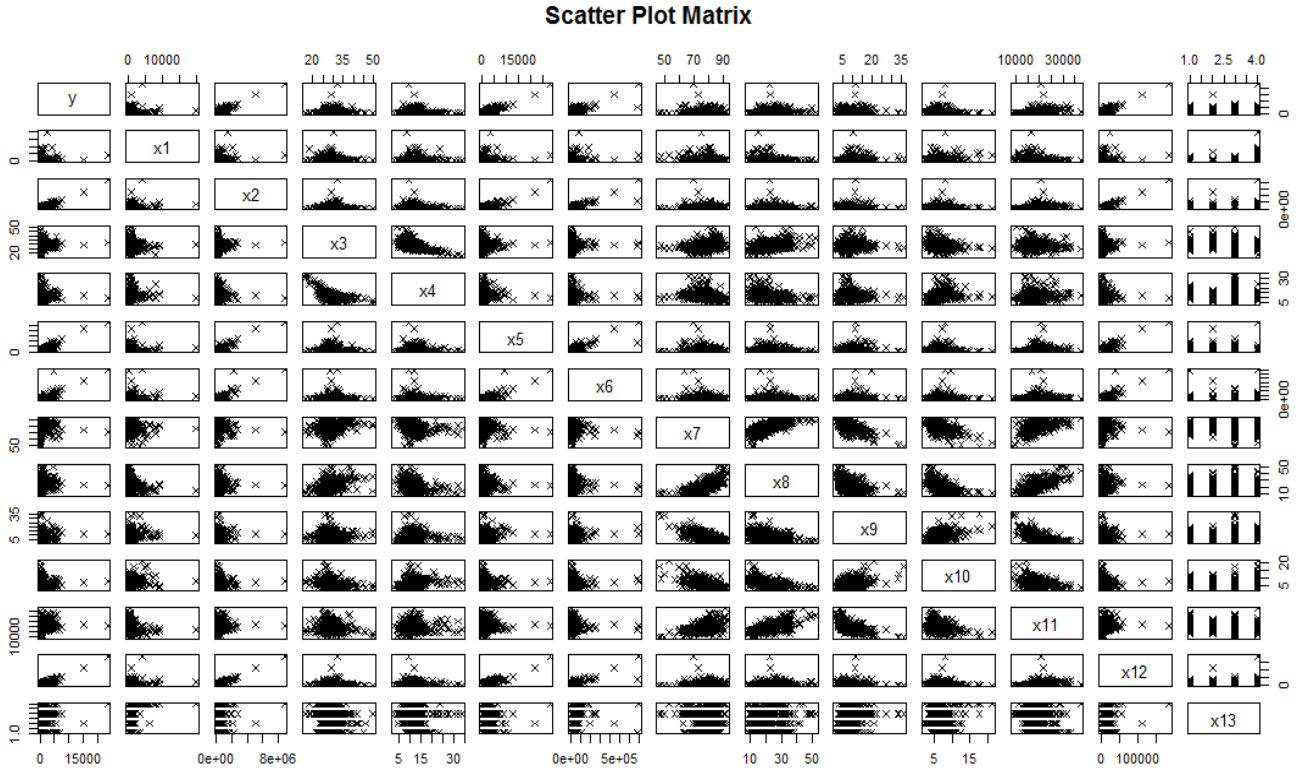








1.2 Pair wise scatter diagram of the response and the predictor variables:



1.3 Interpretation:

The main observations from the above scatter plots:

- Y is highly linearly correlated with variables X2, X5, X6 and X12.
- Y is moderately correlated with other variables.
- X2, X5, X6 and X12 highly correlated among themselves.
- X7, X8 and X9 are also highly correlated.
- Rest other variables are moderately correlated to non-correlated with each other.

As the covariates are highly correlated with each other, it seems that there is presence of multicollinearity.

2. Baseline Model and key Assumptions:

Using the above interpretations, the following model is proposed:

$$Y \sim X1+X2+X3+X4+X5+X6+X7+X8+X9+X10+X11+X12+X13\_1 +X13\_2 +X13\_3 \quad \text{-- eqn. (1)}$$

(The variable X13_4 has been removed before model fitting as if this variable is not removed there will a linear relationship in the predictors ~ X13_1+X13_2+X13_3+X13_4=1)

Multiple linear regression analysis makes following key assumptions:

- There must be a linear relationship between the outcome variable and the independent variables.
- Multivariate Normality– It assumes that the residuals are normally distributed. This assumption can be tested using Shapiro-Wilk normality test.
- No Multicollinearity- It assumes that the independent variables are not highly correlated with each other. This assumption will be later tested using Variance Inflation Factor (VIF) values.
- Independent errors: This means that residuals should be uncorrelated. This assumption can be tested using Durbin-Watson Test.
- Homoskedasticity–This assumption states that the variance of error terms are similar across the values of the independent variables. This assumption can be diagnosed using Studentized Breusch-Pagan Test.

Figure 2.1: Baseline model summary

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x4 + x5 + x6 + x7 + x8 +
    x9 + x10 + x11 + x12 + x13_1 + x13_2 + x13_3, data = dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-1757.38  -125.57     3.56   115.76  2042.89

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.196e+02  5.626e+02   1.457 0.145894
x1          -5.165e-03  1.391e-02  -0.371 0.710635
x2          -1.853e-03  3.412e-04  -5.431 9.47e-08 ***
x3           7.908e+00  6.439e+00   1.228 0.220082
x4           2.251e+00  5.850e+00   0.385 0.700636
x5           5.038e-01  2.527e-02  19.935  < 2e-16 ***
x6          -1.232e-03  7.324e-04  -1.682 0.093266 .
x7          -1.245e+01  5.313e+00  -2.343 0.019610 *
x8           2.440e+01  5.742e+00   4.249 2.64e-05 ***
x9           8.160e-02  7.473e+00   0.011 0.991293
x10         -6.676e+00  1.035e+01  -0.645 0.519394
x11         -2.681e-02  1.068e-02  -2.509 0.012464 *
x12          1.401e-01  1.367e-02  10.246  < 2e-16 ***
x13_1       -1.613e+02  6.650e+01  -2.426 0.015692 *
x13_2       -2.037e+02  6.298e+01  -3.234 0.001316 **
x13_3       -2.178e+02  6.176e+01  -3.526 0.000468 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 354.3 on 424 degrees of freedom
Multiple R-squared:  0.9622,    Adjusted R-squared:  0.9608
F-statistic: 718.7 on 15 and 424 DF,  p-value: < 2.2e-16
```

Interpretation:

The R squared value is very high, i.e., 0.9622 and the p-value is very small (2.2e-16), which denotes that the model is significant.

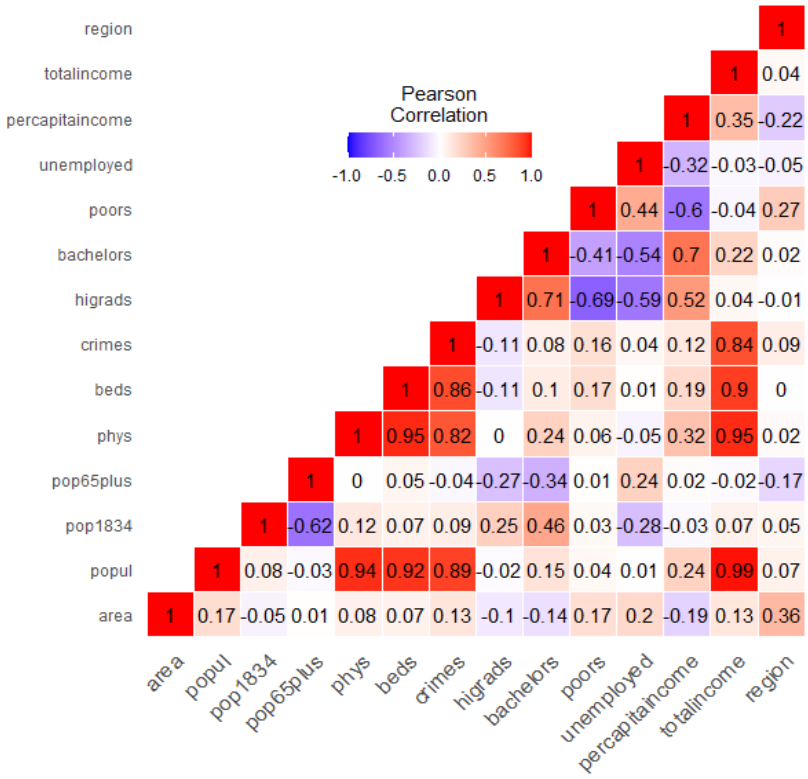
2.1 Testing the assumptions:

Taking into consideration the level of significance to be 5%, the following tests have been performed:

2.1.1 Presence of Multicollinearity

The variance inflation factor (VIF) identifies correlation between independent variables and the strength of that correlation. VIFs greater than 10 represent critical levels of multicollinearity where the coefficients are poorly estimated, and the p-values are questionable.

Figure 2.2: Correlation matrix (lower triangular) heatmap




```
vif(regressor)
      x1          x2          x3          x4          x5          x6          x7          x8
1.626646  147.538952    2.547717    1.908229   11.703812    6.362719    4.858415    6.757539
      x9         x10         x11         x12        x13_1       x13_2       x13_3
4.235749    2.049140    6.579838  108.500465    2.779751    2.575588    3.023550
```

Interpretation

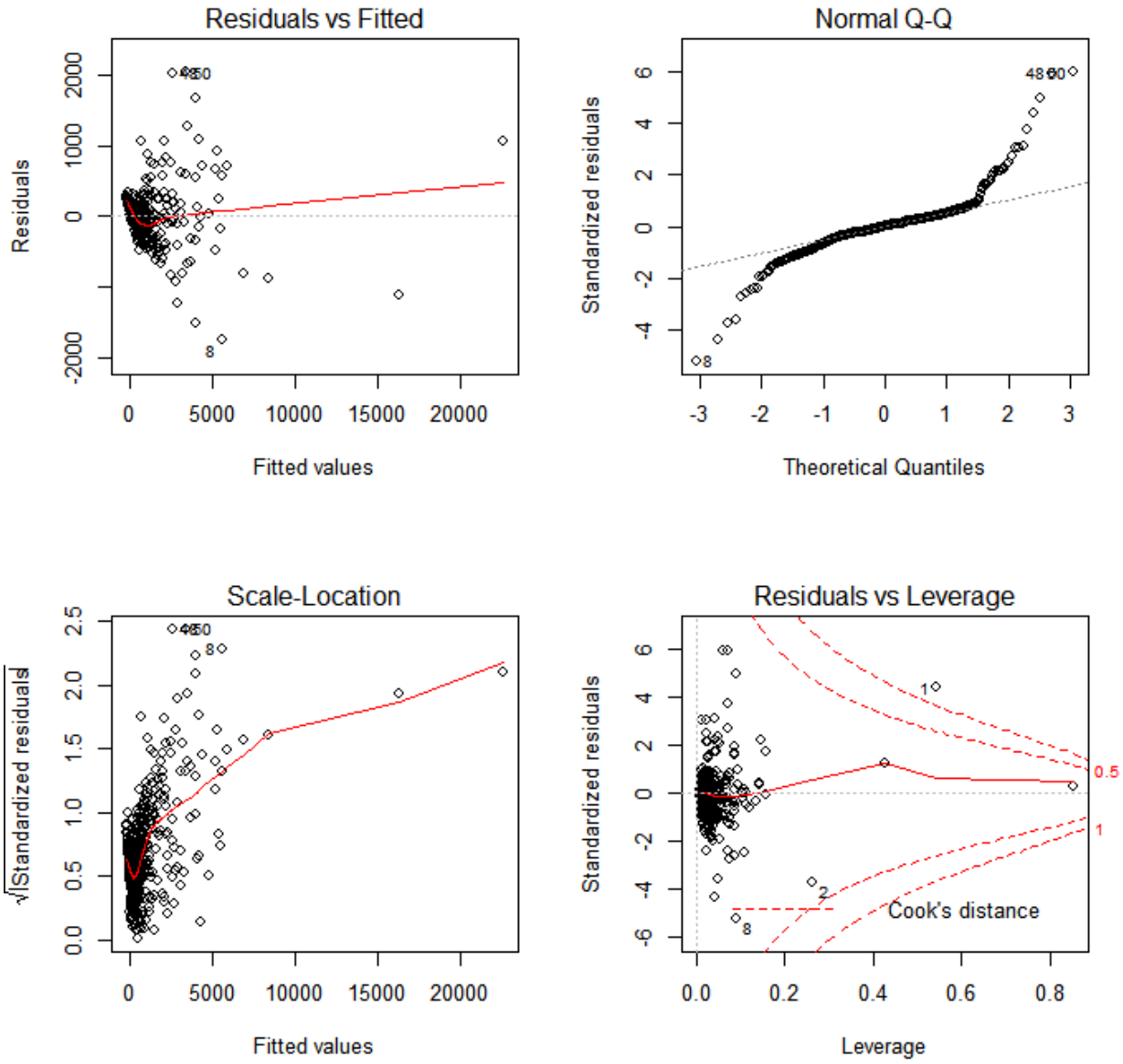
As the VIF values of X2, X5 and X12 are > 10, there exists multicollinearity.

2.1.2 Test for Heteroskedasticity

The Breusch-Pagan test fits a linear regression model to the residuals of a linear regression model (by default the same explanatory variables are taken as in the main regression model) and rejects if too much of the variance is explained by the additional explanatory variables.

Under *H_0* the test statistic of the Breusch-Pagan test follows a chi-squared distribution with parameter (the number of regressors without the constant in the model) degrees of freedom.

Figure 2.3: Baseline Model plots




```
        studentized Breusch-Pagan test

data:  regressor
BP = 112.17, df = 15, p-value < 2.2e-16
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 477.5502, Df = 1, p = < 2.22e-16
```

Interpretation

Since the p-values for Breusch-Pagan and Non-constant Variance Score tests are negligible, there is heteroskedasticity in the model as the null hypothesis is rejected.

2.2.3 Test for Independence of errors

The function Durbin-Watson Test verifies if the residuals from a linear model are correlated or not under the null hypothesis (H0) that there is no correlation among residuals, i.e., they are independent.

```
> regressor = lm(formula = y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13_1+x13_2+x13_3, data = dataset)
> library("lmtest")
> dwtest(regressor)

        Durbin-Watson test

data:  regressor
DW = 2.198, p-value = 0.9797
alternative hypothesis: true autocorrelation is greater than 0
```

Interpretation

Since the p-value is very large, so null hypothesis is accepted, and hence it can be concluded that the errors are independent.

2.2.4 Test for Normality of residuals

If the Shapiro-Wilk Normality test is significant, the distribution is non-normal.

```
        Shapiro-Wilk normality test

data:  errors
W = 0.84083, p-value < 2.2e-16
```

Interpretation

The p-value for Shapiro-Wilk normality test is negligible. Hence, residuals are not normal.

3. Variable Selection

From figure 2.1 of baseline model summary, it can be seen that p-values for predictor variables X1, X3, X4, X6, X9 and X10 are greater than the 5% level of significance, & hence it can be concluded that not all the predictor variables present in the baseline model (equation (1)) are significant.

So, a variable selection algorithm is performed to get a reasonable set of significant predictors.

3. 1 Stepwise AIC Backward Regression

This algorithm is used in building regression model from a set of candidate predictor variables by removing predictors based on Akaike's information criterion (AIC), in a stepwise manner until there is no variable left to remove any more.

This is done using **ols_step_backward_aic( )** function in R available in **olsrr** library.

```
> # Fitting of Linear Regression model:-
> regressor = lm(formula = y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13_1+x13_2+x13_3, data = dataset)
> library("olsrr")
> k=ols_step_backward_aic(regressor)
> k


                        Backward Elimination Summary
-------------------------------------------------------------------------------
Variable        AIC           RSS             Sum Sq         R-Sq      Adj. R-Sq
-------------------------------------------------------------------------------
Full Model    6432.019    53214956.943    1352991342.055    0.96216      0.96082
x9            6430.019    53214971.908    1352991327.089    0.96216      0.96091
x1            6428.163    53232318.892    1352973980.106    0.96214      0.96099
x4            6426.296    53248469.502    1352957829.496    0.96213      0.96107
x10           6424.769    53305786.035    1352900512.963    0.96209      0.96112
x3            6424.312    53493005.782    1352713293.216    0.96196      0.96107
-------------------------------------------------------------------------------
```

From the above algorithm, the variables that are needed to be removed are X9, X1, X4, X10, X3 in a stepwise manner.

Hence, the updated model is:

Y ~ X2+X5+X6+X7+X8+X11+X12+X13_1 +X13_2 +X13_3 -- eqn. (2)

Performing **Stepwise AIC Backward Regression** on the updated model for verification:

```
> regressor = lm(formula = y~x2+x5+x6+x7+x8+x11+x12+x13_1+x13_2+x13_3, data = dataset)
> summary(regressor)

Call:
lm(formula = y ~ x2 + x5 + x6 + x7 + x8 + x11 + x12 + x13_1 +
    x13_2 + x13_3, data = dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-1761.61  -127.02    -2.27   120.29  2054.64

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.436e+02  2.797e+02   3.373 0.000810 ***
x2          -1.924e-03  3.152e-04  -6.105 2.31e-09 ***
x5           5.094e-01  2.140e-02  23.801  < 2e-16 ***
x6          -1.185e-03  7.157e-04  -1.655 0.098563 .
x7          -1.137e+01  3.996e+00  -2.845 0.004650 **
x8           2.847e+01  4.062e+00   7.008 9.40e-12 ***
x11         -3.281e-02  8.077e-03  -4.062 5.80e-05 ***
x12          1.426e-01  1.308e-02  10.903  < 2e-16 ***
x13_1       -1.361e+02  5.765e+01  -2.360 0.018716 *
x13_2       -1.788e+02  5.617e+01  -3.183 0.001563 **
x13_3       -1.906e+02  5.346e+01  -3.565 0.000404 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 353.1 on 429 degrees of freedom
Multiple R-squared:  0.962,     Adjusted R-squared:  0.9611
F-statistic:  1085 on 10 and 429 DF,  p-value: < 2.2e-16

> library("olsrr")
> k=ols_step_backward_aic(regressor)
> k
[1] "No variables have been removed from the model."
```

3.2 Checking for the presence of multicollinearity in updated model:

VIF values for the updated model (eqn. (2)) are as follows:

```
> regressor = lm(formula = y~x2+x5+x6+x7+x8+x11+x12+x13_1+x13_2+x13_3, data = dataset)
> vif(regressor)
       x2        x5        x6        x7        x8       x11       x12    x13_1    x13_2
126.736671   8.449445   6.116383   2.767298   3.403306   3.784788  99.970857   2.102818   2.062009
    x13_3
  2.280234
```

As the VIF value of X2 is > 10, there exists multicollinearity.

3.3 Fitting of an appropriate PCA regression model:

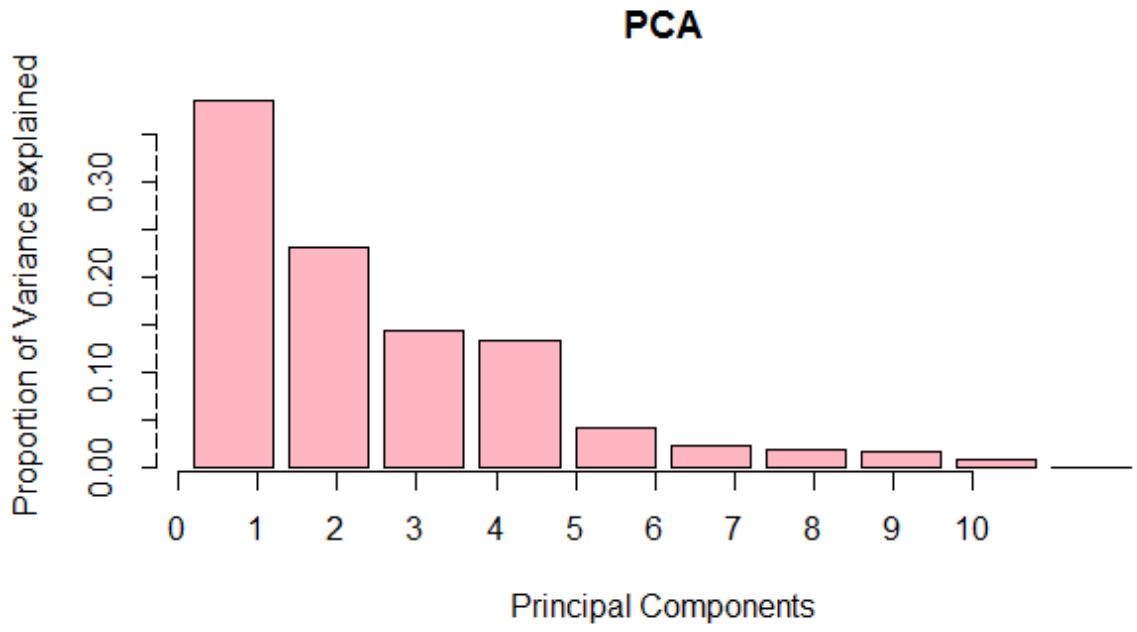
To perform Principal Component Analysis in R, the function **prcomp**() has been used which uses the singular value decomposition (SVD) (which examines the covariances / correlations between individuals).

```
> Comp=data.frame(x2,x5,x6,x7,x8,x11,x12,x13_1,x13_2,x13_3)
> myPCA <- prcomp(Comp, scale. = T, center = T)
> myPCA$rotation # loadings
              PC1         PC2          PC3          PC4          PC5         PC6          PC7
x2     -0.49477947  0.11205221 -0.028794959  0.010613835 -0.100366213  0.11017618 -0.12447377
x5     -0.47370277  0.15403395 -0.042575906  0.005248352  0.079012368 -0.21102725 -0.03648231
x6     -0.45490117  0.18198565 -0.005984711  0.031573487 -0.158789747 -0.17932172  0.22997535
x7     -0.05021359 -0.55709112 -0.034055468  0.250310323 -0.558825232 -0.05492496 -0.52385610
x8     -0.15833706 -0.52410768  0.257093608  0.212089036 -0.033627509 -0.32489736  0.63747088
x11    -0.20850650 -0.49926669  0.113373475 -0.051713785  0.673171402  0.35790684 -0.15020470
x12    -0.49762491  0.04613349 -0.014965011  0.001838444 -0.001512993  0.18339561 -0.17281027
x13_1  -0.03672211 -0.18816917  0.022740720 -0.794992730  0.038245091 -0.52086823 -0.20004883
x13_2   0.02462515 -0.05079904 -0.710233361  0.355802246  0.354278826 -0.44399091 -0.10973330
x13_3   0.05666948  0.23701475  0.641902924  0.360471253  0.254783902 -0.41525231 -0.38079726
                PC8          PC9          PC10
x2     -0.1986128565 -0.331006991 -0.744967913
x5     -0.3205922306  0.770023962  0.047077325
x6      0.8016200819 -0.009725128  0.077233362
x7      0.1153948495  0.141526291  0.015484885
x8     -0.2405326708 -0.134924666 -0.006540971
x11     0.2630459183  0.114238553 -0.075285175
x12    -0.2711536800 -0.426983308  0.656258431
x13_1   0.0007317031 -0.133772287 -0.005636325
x13_2   0.0073915387 -0.175903111 -0.010345976
x13_3   0.0351314478 -0.122337291 -0.008785411
```

Figure 3.1: Model summary of PCA regression model:

```
> comp = data.frame(x2,x5,x6,x7,x8,x11,x12,x13_1,x13_2,x13_3)
> myPCA<- prcomp(comp, scale = T, center = T)
> summary(myPCA)
Importance of components:
                          PC1    PC2    PC3    PC4    PC5    PC6     PC7     PC8
Standard deviation     1.9615 1.5225 1.1957 1.1528 0.6434 0.4796 0.43660 0.38998
Proportion of Variance 0.3847 0.2318 0.1430 0.1329 0.0414 0.0230 0.01906 0.01521
Cumulative Proportion  0.3847 0.6165 0.7595 0.8924 0.9338 0.9568 0.97586 0.99107
                           PC9    PC10
Standard deviation     0.29127 0.06664
Proportion of Variance 0.00848 0.00044
Cumulative Proportion  0.99956 1.00000
```

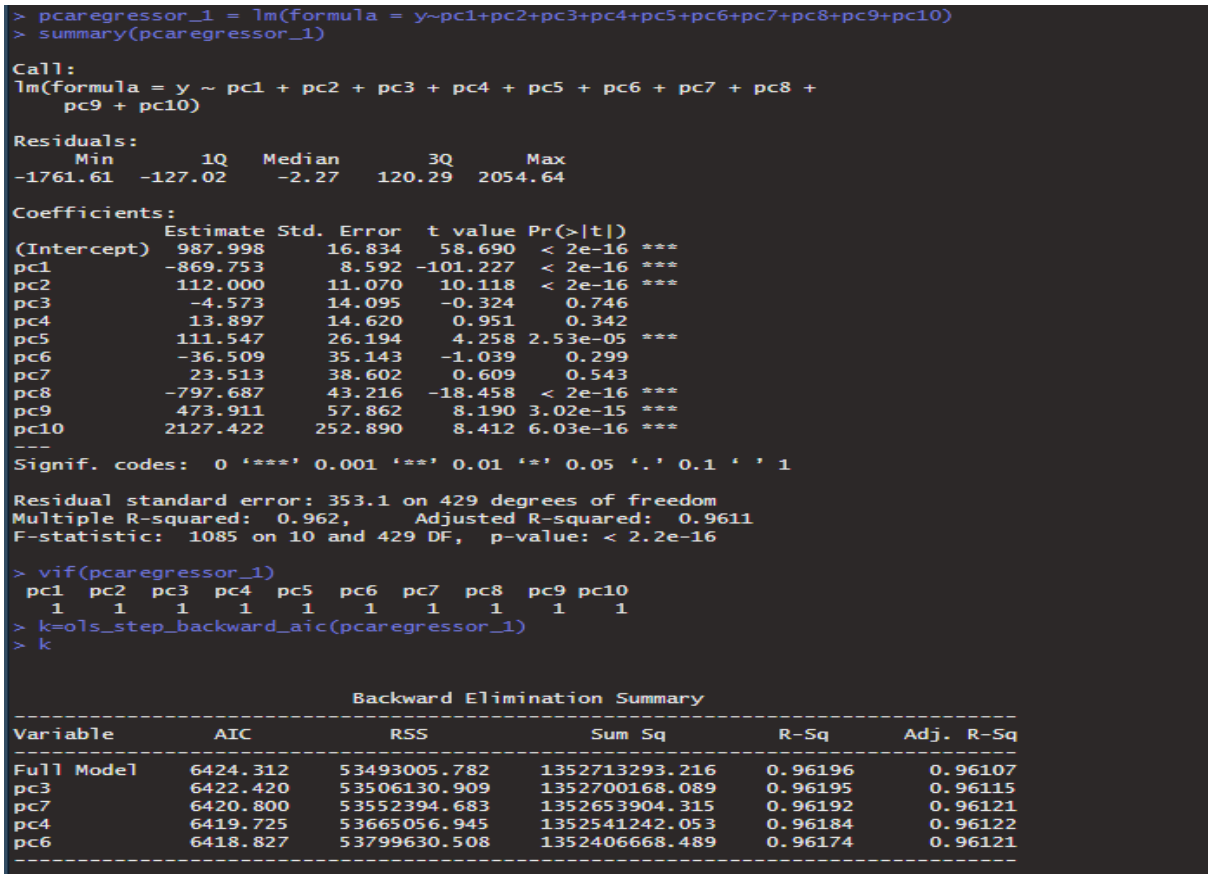
Figure 3.2 Bar Graph showing Proportion of Variances explained Vs Principal Components



Now, our updated model is:

Y ~ pc1+pc2+pc3+pc4+pc5+pc6+pc7+pc8+pc9+pc10 -- eqn. (3)

Figure 3.3 Updated model summary and outcome of backward elimination algorithm



```
> pcaregressor_1 = lm(formula = y~pc1+pc2+pc3+pc4+pc5+pc6+pc7+pc8+pc9+pc10)
> summary(pcaregressor_1)

Call:
lm(formula = y ~ pc1 + pc2 + pc3 + pc4 + pc5 + pc6 + pc7 + pc8 +
    pc9 + pc10)

Residuals:
     Min       1Q   Median       3Q      Max
-1761.61  -127.02    -2.27   120.29  2054.64

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)  987.998     16.834   58.690  < 2e-16 ***
pc1         -869.753      8.592 -101.227  < 2e-16 ***
pc2          112.000     11.070   10.118  < 2e-16 ***
pc3           -4.573     14.095   -0.324    0.746
pc4           13.897     14.620    0.951    0.342
pc5          111.547     26.194    4.258 2.53e-05 ***
pc6          -36.509     35.143   -1.039    0.299
pc7           23.513     38.602    0.609    0.543
pc8         -797.687     43.216  -18.458  < 2e-16 ***
pc9          473.911     57.862    8.190 3.02e-15 ***
pc10        2127.422    252.890    8.412 6.03e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 353.1 on 429 degrees of freedom
Multiple R-squared:  0.962,     Adjusted R-squared:  0.9611
F-statistic:  1085 on 10 and 429 DF,  p-value: < 2.2e-16

> vif(pcaregressor_1)
 pc1  pc2  pc3  pc4  pc5  pc6  pc7  pc8  pc9 pc10
   1    1    1    1    1    1    1    1    1    1
> k=ols_step_backward_aic(pcaregressor_1)
> k


                         Backward Elimination Summary
-------------------------------------------------------------------------------
Variable        AIC            RSS              Sum Sq         R-Sq      Adj. R-Sq
-------------------------------------------------------------------------------
Full Model     6424.312    53493005.782    1352713293.216    0.96196       0.96107
pc3            6422.420    53506130.909    1352700168.089    0.96195       0.96115
pc7            6420.800    53552394.683    1352653904.315    0.96192       0.96121
pc4            6419.725    53665056.945    1352541242.053    0.96184       0.96122
pc6            6418.827    53799630.508    1352406668.489    0.96174       0.96121
-------------------------------------------------------------------------------
```

It is verified from above figure that there is no multicollinearity in the model in eqn. (3), as all the VIF values are equal to 1 (i.e., less than 10).

Now, the updated model (after applying backward elimination algorithm) becomes:

$$Y \sim pc1+pc2+pc5+pc8+pc9+pc10 \quad \text{-- eqn. (4)}$$

Fig3.4: Summary of the updated model (eqn. 4)

```
> #Fitting the model with significant PCA variables
> pcaregressor_2 = lm(formula = y~pc1+pc2+pc5+pc8+pc9+pc10)
> summary(pcaregressor_2)

Call:
lm(formula = y ~ pc1 + pc2 + pc5 + pc8 + pc9 + pc10)

Residuals:
     Min       1Q   Median       3Q      Max
-1729.22  -137.39    -8.32   122.94  2072.52

Coefficients:
            Estimate Std. Error  t value Pr(>|t|)
(Intercept)  987.998     16.804   58.794  < 2e-16 ***
pc1         -869.753      8.577 -101.408  < 2e-16 ***
pc2          112.000     11.050   10.136  < 2e-16 ***
pc5          111.547     26.147    4.266 2.44e-05 ***
pc8         -797.687     43.139  -18.491  < 2e-16 ***
pc9          473.911     57.759    8.205 2.66e-15 ***
pc10        2127.422    252.440    8.427 5.28e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 352.5 on 433 degrees of freedom
Multiple R-squared:  0.9617,    Adjusted R-squared:  0.9612
F-statistic:  1814 on 6 and 433 DF,  p-value: < 2.2e-16
```

4. Verification of assumptions for updated model:

4.1 Test for Independence of errors

```
> #Testing the assumptions
> #install.packages("lmtest")
> library("lmtest")
> dwtest(pcaregressor_2)

        Durbin-Watson test

data:  pcaregressor_2
DW = 2.1912, p-value = 0.9751
alternative hypothesis: true autocorrelation is greater than 0
```

Interpretation

Since the p-value is very large, so null hypothesis of Durbin-Watson Test is accepted, and hence it can be concluded that the errors are independent.

4.2 Test for Heteroskedasticity

```
> library(lmtest)
> lmtest::bptest(pcaregressor_2)

        studentized Breusch-Pagan test

data:  pcaregressor_2
BP = 100.37, df = 6, p-value < 2.2e-16

> # NCV test (Non-constant Variance Score Test):-
> car::ncvTest(pcaregressor_2)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 466.0052, Df = 1, p = < 2.22e-16
```

Interpretation

Since the p-values for Breusch-Pagan and Non-constant Variance Score tests are negligible, there is heteroskedasticity in the model as the null hypothesis is rejected.

4.3 Test for Normality of residuals

```
> #Shapiro-Wilk test of normality :-
> errors = residuals(regressor)
> shapiro.test(errors)

        Shapiro-Wilk normality test

data:  errors
W = 0.84334, p-value < 2.2e-16
```

Interpretation

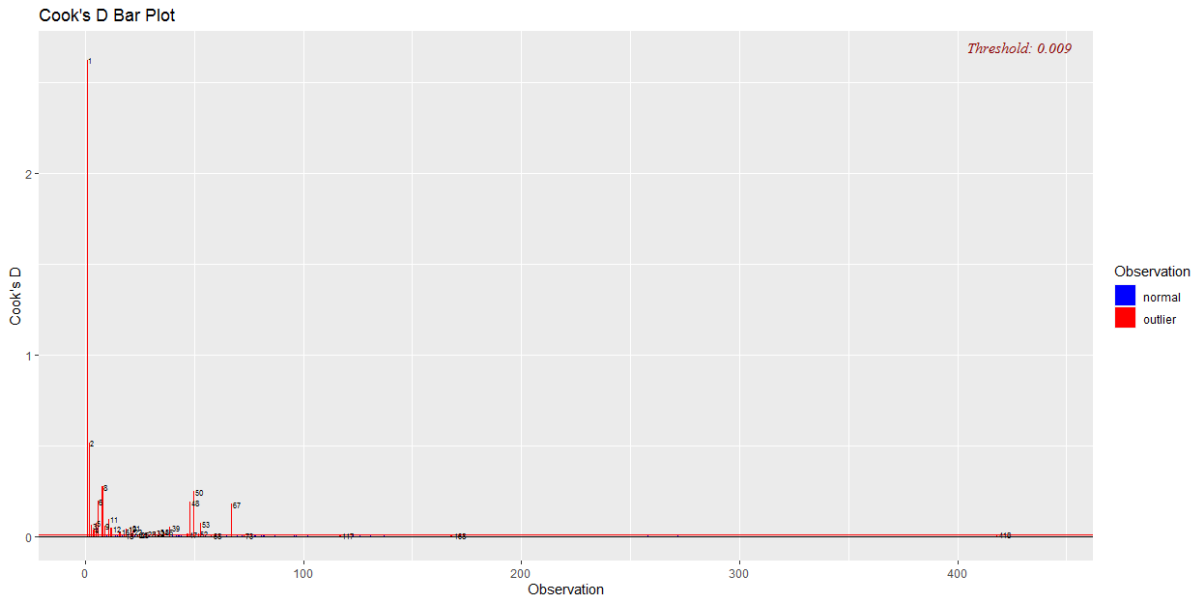
The p-value for Shapiro-Wilk normality test is negligible. Hence, residuals are not normal.

It can be concluded that the model obtained in eqn. (4) follows all the assumptions except that of homoskedasticity and normality of residuals.

5. Outlier Analysis

Cook's Distance bar plot given below shows that there are 27outliers out of 440 data points used in fitting of the model.

Fig. 5.1: Cook's Distance Bar Plot



Stable model is constructed after removing these data sets.

6. Final Model

The final fitted model is (after removing the outliers obtained above):

Y = 957.010 - 827.557*pc1 + 113.326*pc2 + 123.343*pc5 – 714.115*pc8 + 418.127*pc9 + 2080.976*pc10 -- eqn. (5)

Fig.6.1 Summary of final model:

```
Call:
lm(formula = y ~ pc1 + pc2 + pc5 + pc8 + pc9 + pc10)

Residuals:
     Min       1Q   Median       3Q      Max
-1177.62  -115.84     1.33    99.66  1148.33

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  957.010     11.946  80.110  < 2e-16 ***
pc1         -827.557     13.547 -61.089  < 2e-16 ***
pc2          113.326      9.142  12.396  < 2e-16 ***
pc5          123.343     19.075   6.466 2.89e-10 ***
pc8         -714.115     57.298 -12.463  < 2e-16 ***
pc9          481.127     47.004  10.236  < 2e-16 ***
pc10        2080.976    243.134   8.559 2.37e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 226.6 on 406 degrees of freedom
Multiple R-squared:  0.9343,    Adjusted R-squared:  0.9333
F-statistic: 961.6 on 6 and 406 DF,  p-value: < 2.2e-16
```

7. Summary of Analysis

The CDI data was taken which had 13 predictor variables (viz. X1, X2, … X13) and 1 response variable (Y). Amongst 13 variables, 1 variable was categorical and others were numerical. With the help of scatter plots, it was seen that some variables were linearly related to Y and some were non-linearly related.

Initially, multiple linear regression model was fitted with appropriate assumptions, without any transformations in the explanatory variables. It was noticed that R sq. was 96.22%. On testing the assumptions, it was found that there is multicollinearity present in the model and the errors were independent, but they weren't normally distributed and also, there was heteroskedasticity in the model. The model was found to be significant.

After this, Stepwise AIC Backward Regression variable selection algorithm was performed to get a reasonable set of significant predictors. The variables removed were X9, X1, X4, X10, X3 in a stepwise manner.

On checking for the presence of multicollinearity in the updated model, it was found that VIF value of X2 was greater than 10.

To get rid of multicollinearity, an appropriate PCA regression model was fitted. Again Stepwise AIC Backward Regression was applied to get significant principal components. After this, testing of assumptions was done on the updated model.

It was found that errors were still independent but they were not normally distributed and heteroskedasticity was still present in the model.

Cook's Distance bar plot was used to detect 27outliers out of 440 data points used in fitting of the model.
Final model has been obtained with adjusted R sq. as 93.33% on removing the outliers obtained after considering all the transformations and the deductions.

In future, suitable transformation in the dependent variable can be useful to make the residuals normal and to get rid of the heteroskedasticity in the model.