Regression Analysis of Petrol Consumption Data

Sagar Dey, Nikhil Bhardwaj, Debarshi Chakraborty

December 20, 2021

Abstract

We have a data on petrol consumption,with several relevant explanatory variables.Our goal is to develop a model to predict the response variable with the help of the explanatory variables.The model will contain unknown parameters which we will estimate with the data at hand and then try to assess how well our model performs,if any problems arise, we try to overcome that and modify our model accordingly.We mainly try to implement our knowledge gained from the Regression Techniques course to this real life data.

Data Description:

For one year, the consumption of petrol(in millions of gallons) was measured in 48 states.
The relevant variables are:
x_1:the petrol tax(in cents per gallon)
x_2:the average income per capita(in dollars)
x_3:the number of miles of paved highway(in miles)
x_4:the proportion of the population with driver's licenses
So,here y (consumption of petrol) is our response/dependent variable and x_1, x_2, x_3, x_4 which are defined above are our explanatory variables.
For our convenience,we will use $x_1., , x_4, y$ instead of the whole names of the variables in most of the cases throughout the project.
So,after loading the dataset and storing as a dataframe, our data looks like as follows:

```
    x1   x2   x3    x4   y
1  9.0 3571 1976 0.525 541
2  9.0 4092 1250 0.572 524
3  9.0 3865 1586 0.580 561
4  7.5 4870 2351 0.529 414
5  8.0 4399  431 0.544 410
6 10.0 5342 1333 0.571 457
```

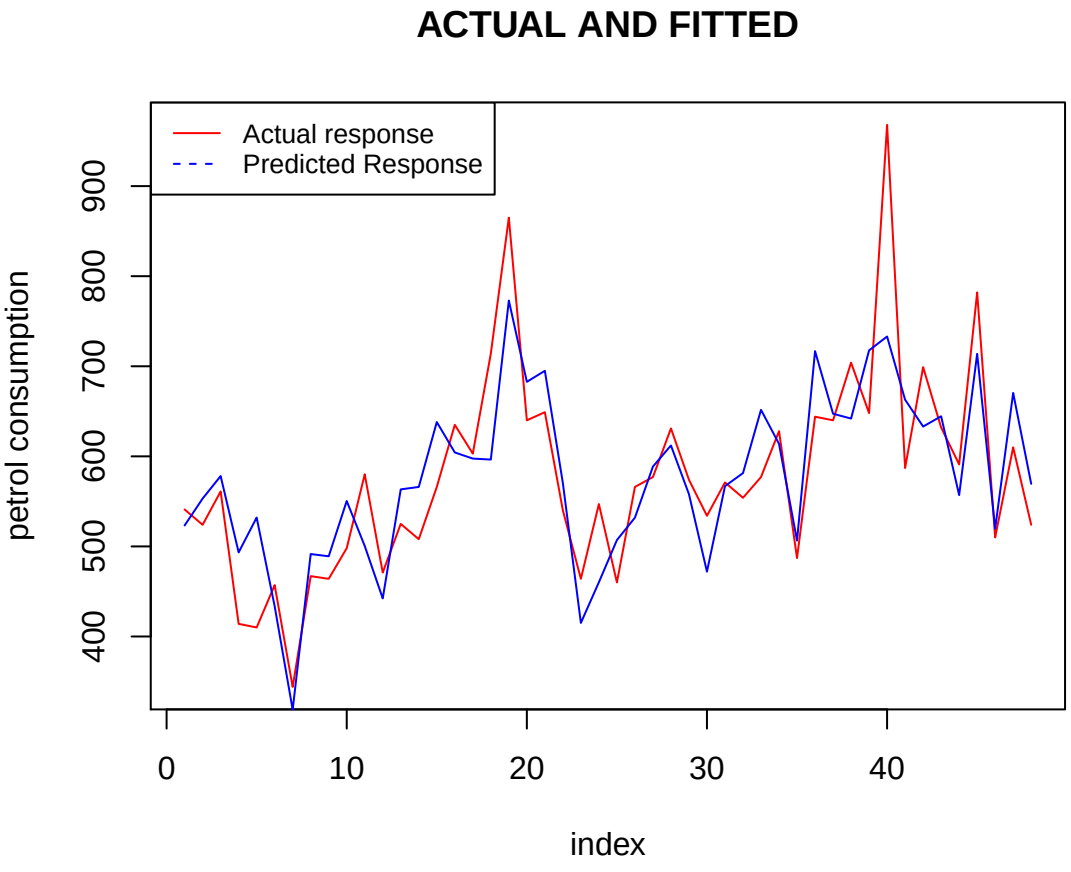
We have 48 observations and no null values,thus we directly start our work.

At first, we start with the most basic and try to fit a multiple linear regression model with all the available covariates, with an intercept term i.e. our initial model is:

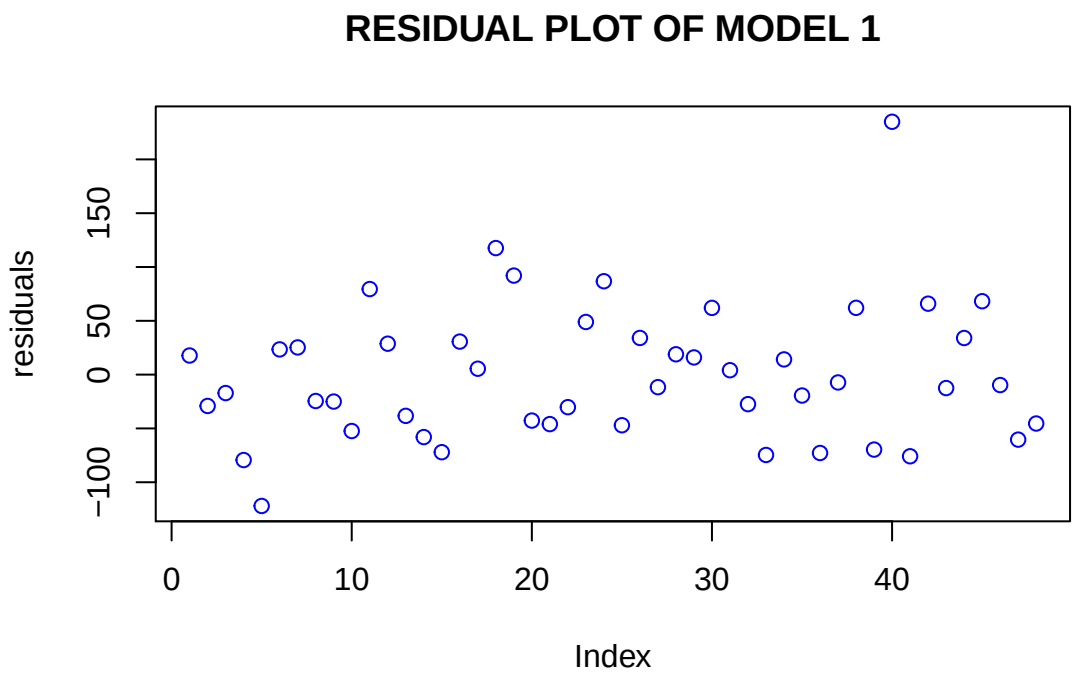
$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

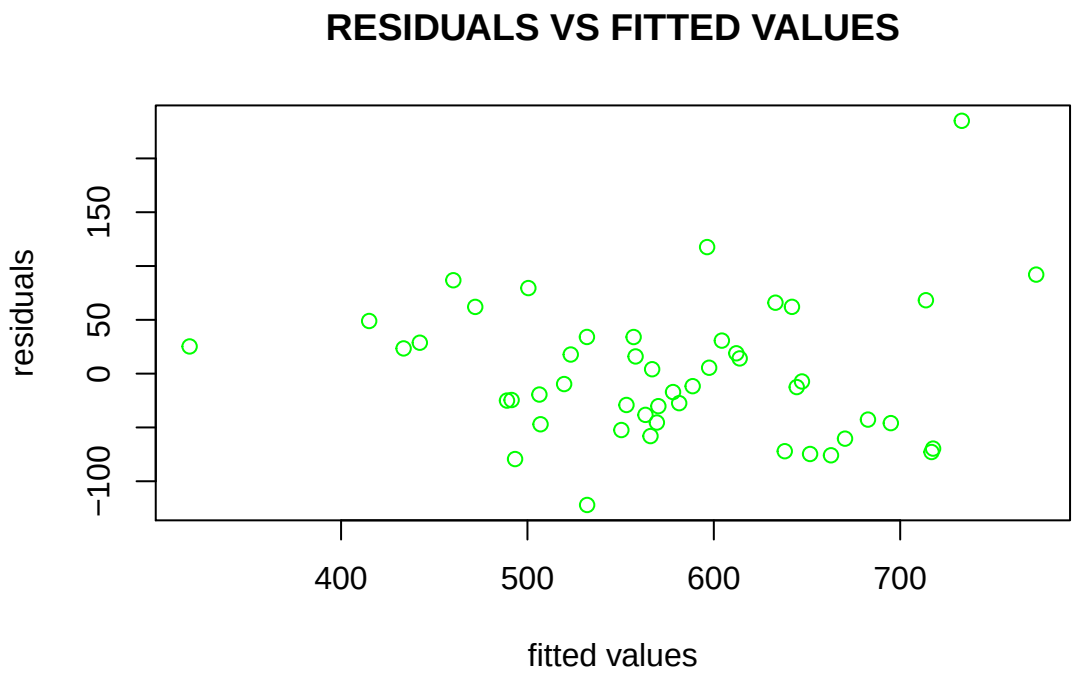
```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-122.03  -45.57  -10.66   31.53  234.95

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.773e+02  1.855e+02   2.033 0.048207 *
x1          -3.479e+01  1.297e+01  -2.682 0.010332 *
x2          -6.659e-02  1.722e-02  -3.867 0.000368 ***
x3          -2.426e-03  3.389e-03  -0.716 0.477999
x4           1.336e+03  1.923e+02   6.950 1.52e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 66.31 on 43 degrees of freedom
Multiple R-squared:  0.6787,Adjusted R-squared:  0.6488
F-statistic: 22.71 on 4 and 43 DF,  p-value: 3.907e-10
```

Now we will compare between the actual values and the fitted values.



Now,we take a look at the residual plot.





Apparently,it seems that the residuals do not exhibit any particular pattern.Also,we perform a nonparametric test-one sample Run's Test to check whether our visual inspection is justified or not.

```
Runs Test

data:  model$residuals
statistic = -0.2918, runs = 24, n1 = 24, n2 = 24, n = 48, p-value =
0.7704
alternative hypothesis: nonrandomness
```

Interpretation:

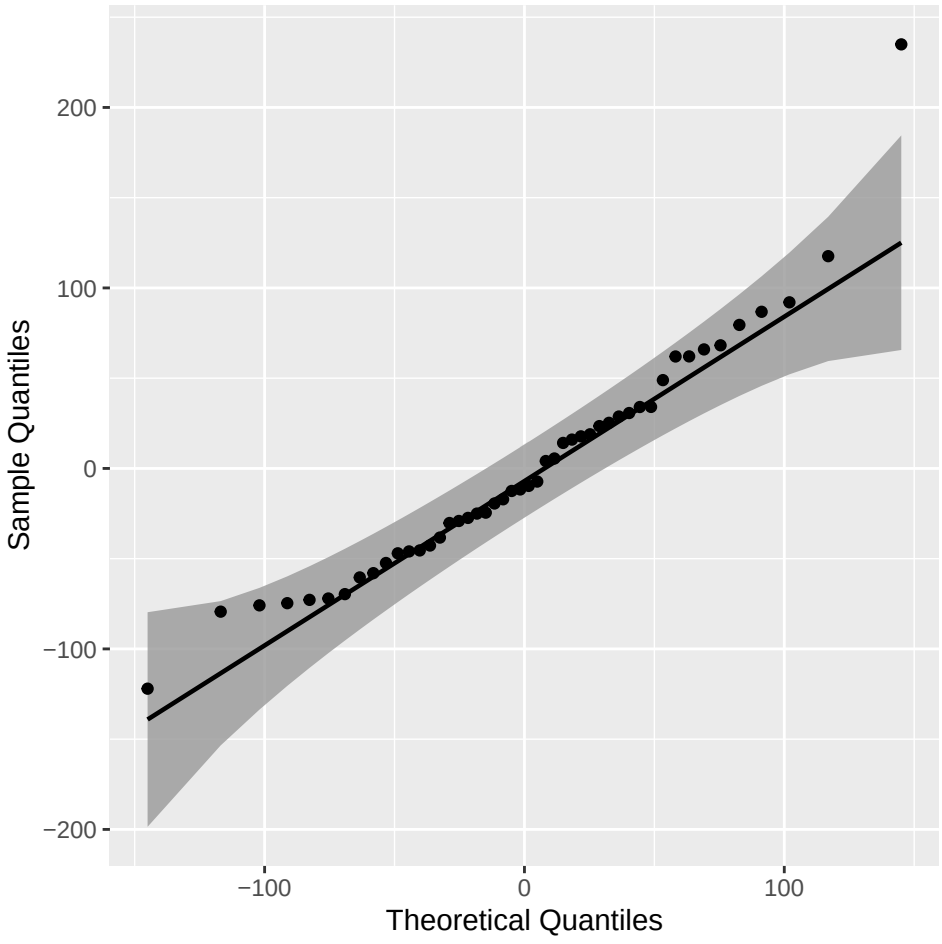
Although,the residuals can be seen to be randomly scattered since the p-value of Run's test is >0.05,still we are not sure whether our sample size is large enough or not to conclude anything. Moreover,looking at the multiple $R^2 = 0.6787$,we conclude that 67.87% of the total variability is explained by our model and from the plot of fitted values vs actual values, it is evident that our model don't agree "VERY WELL" to the actual data and hopefully we can improve.Also,we need to check whether the assumptions of simple linear regression are satisfied or not.
At first,we test whether the assumption of Homoscedasticity is satisfied or not.We will use the Breusch Pagan test for this.

p-value	0.006901944

Since,p-value for the Breusch Pagan test come out to be less than the desired level of significance $\alpha = 0.05$, hence we do not have enough evidence to conclude that the assumption of homoscedasticity holds in our model. But , the Bresuch Pagan Test is very sensitive to any violation of normality assumption, so we cannot directly conclude anything from here unless the assumption of normality holds.Thus, we proceed to check that assumption.

We will try to do some visual inspection at first,using the Quantile-Quantile Plot.

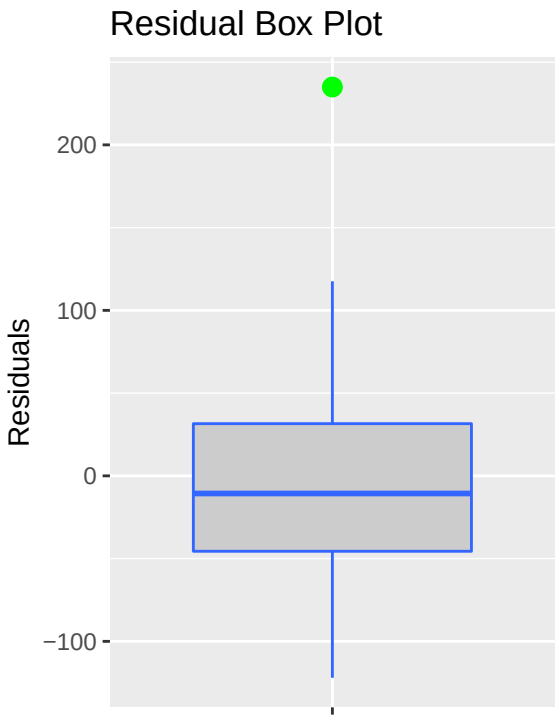
```
Loading required package:  ggplot2

Attaching package:  'qqplotr'
The following objects are masked from 'package:ggplot2':
      stat_qq_line, StatQqLine
```



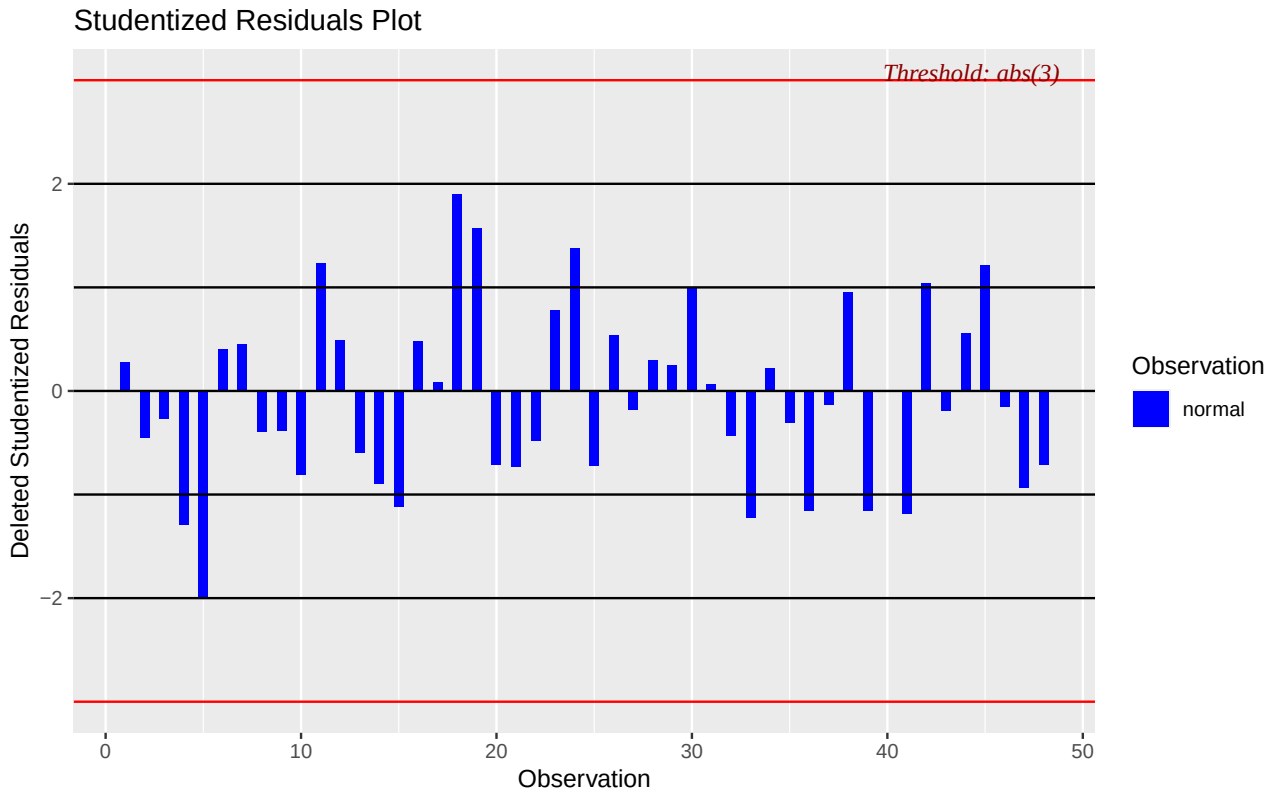
From the diagram,we don't get a very good idea whether the quantiles of the errors obtain from our data actually match we the theoritical quantiles of the normal distribution.So,we perform the Shapiro-Wilks Test for normality to verify the assumption.The p-value of Shapiro-Wilks test come out to be $0.0151 < 0.05 = \alpha$,which is our desired level of significance,Hence,we have enough evidence to conclude that the assumption of normality does not hold.Hence,we need a remedy for this.
As a remedial measure,we want to do the famous Box-Cox Transformation.But,this transformation may not work well if outliers are present.Thus,we proceed to the diagnosis of outliers,high leverage points,influential points,etc.
We are now going to encounter the following notations,terminologies and measures very often for a while:-

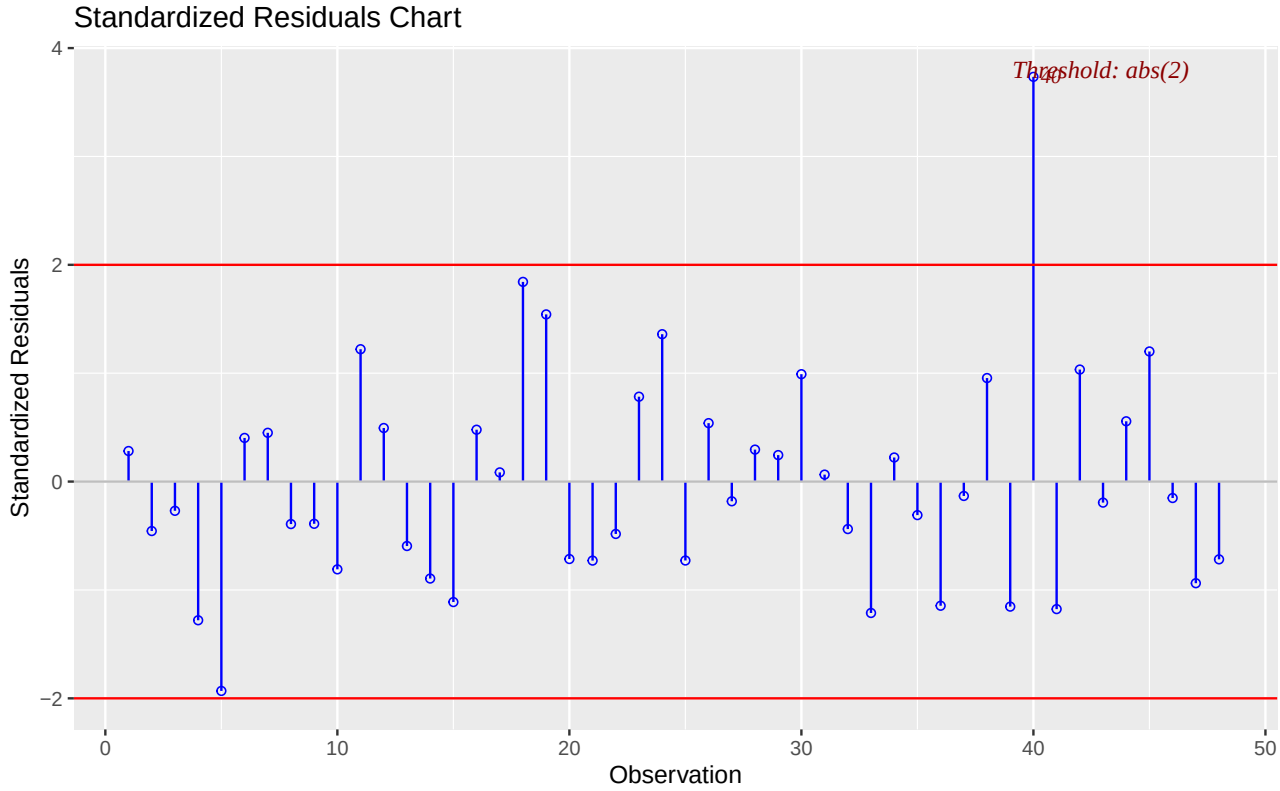
Let y_i be the actual value of the response in the i^{th} observation and $\hat{y}_i$ denote the fitted value for the same.
Let, H denote the hat matrix and h_i denote the i^{th} diagonal element of the hat matrix.Let $\hat{\beta}$ be the usual LSE and $S^2 = \sum_{i=1}^{n} \frac{e_i^2}{n-p}$ be the usual unbiased estimate of the error variance σ^2.$\hat{\beta(i)}$ and $S(i)^2$ denote the same estimates computed by omitting the i^{th} observation. Then,
Residual: $e_i = y_i - \hat{y}_i$
Internally studentized residual: $r_i = \frac{e_i}{S(1-h_i)^{\frac{1}{2}}}$ and Externally studentized residual: $t_i = \frac{e_i}{S(i)(1-h_i)^{\frac{1}{2}}}$
We will be using some more measures like DFFITS,Cook's D,etc.We will do the leave one out diagnostics only.
At first we start by plotting the raw residuals.

```
Attaching package:  'olsrr'
The following object is masked from 'package:datasets':
      rivers
```



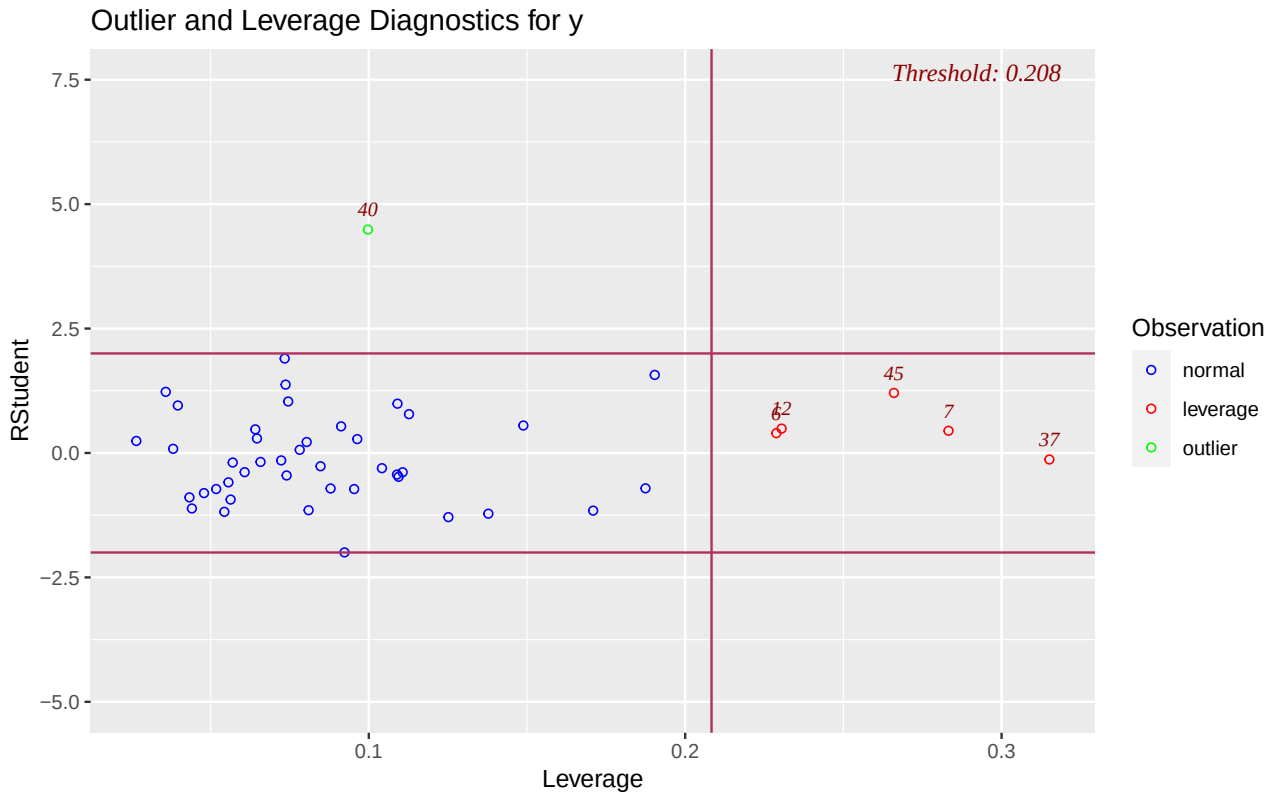
From the Boxplot above,it seems that one point is having extremely large residual,so it may be a potential outlier.Let us plot the internally and externally studentized residuals too.





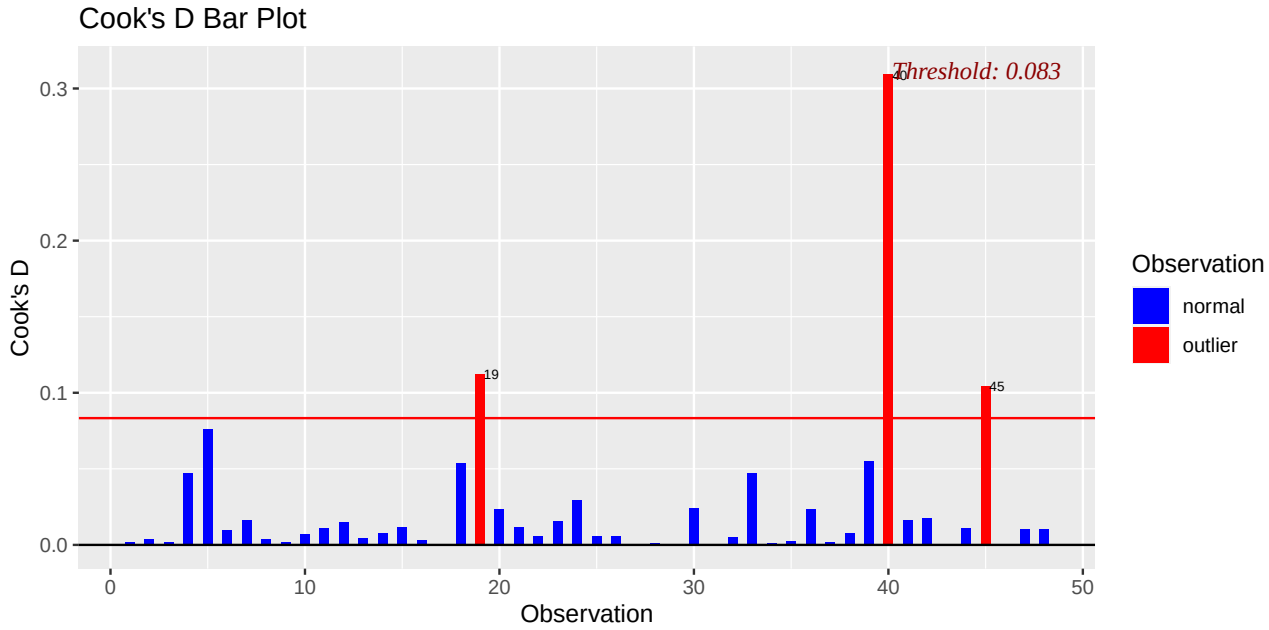
Since we know that in absence of outliers the $t_i \sim t_{n-p-1}$, a reasonable definition of "large" is a point for which $|t_i| > 2$.Although the plot of internally studentized residuals indicate one point as potential outlier, the plot of externally studentized residuals does not reveal anything.But,looking at these plots is not enough to conclude surely about outliers as this diagnostic approach fails if the point we are checking for has high leverage.

Hence, we plot the studentized residuals and hat matrix diagonals on two axes of the same plot.A reasonable definition for high leverage point is one satisfying $h_i > \frac{2p}{n}$.

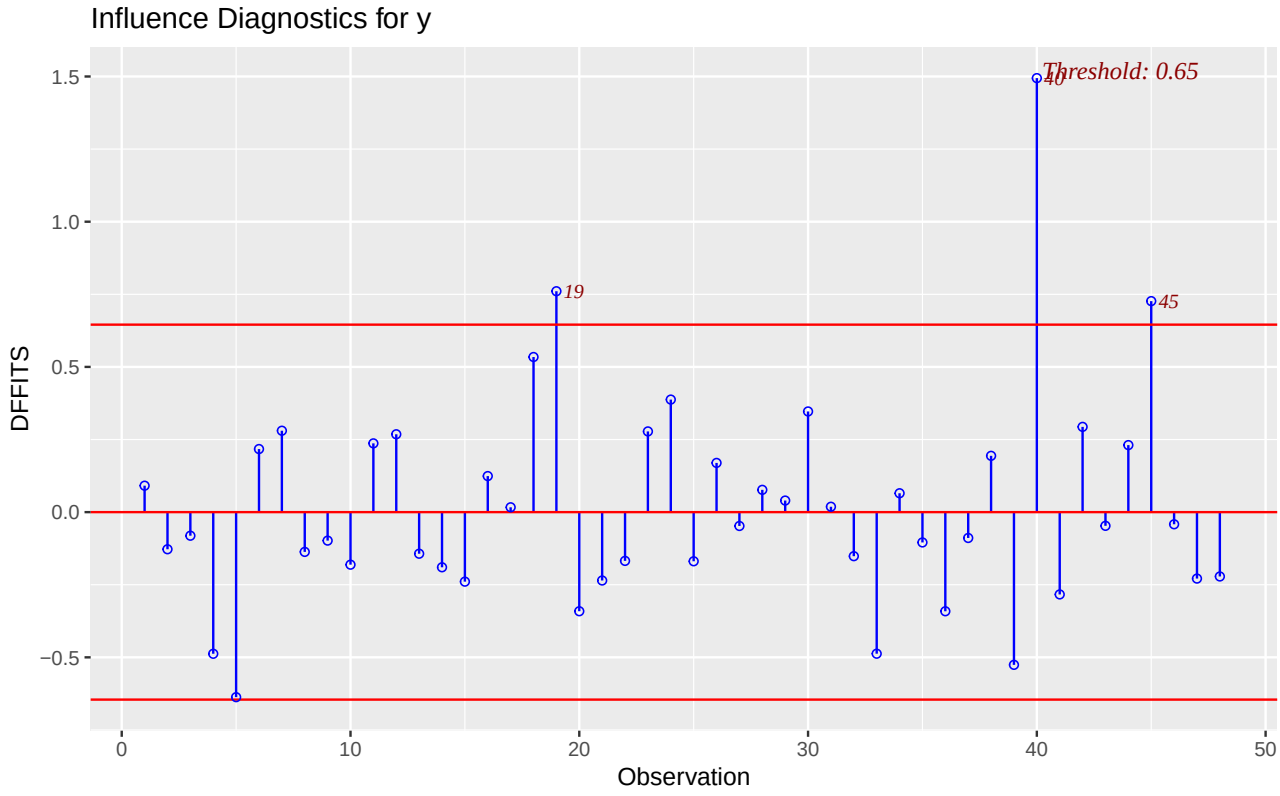


Now the above diagram gives us more clarity , we can safely conclude that the 40^{th} observation is an outlier.But,these basic visualizations are always not enough to get hold of influential points since unfortunately,the hat matrix diagonals are themselves subject to the effect of high leverage points and do not always give a reliable indication of leverage.

Hence, we use an improved measurement like Cook's D which for the i^{th} observation is defined as $D_i = \frac{r_i^2 h_i}{p(1-h_i)}$. This measure takes into account both the cases of high residuals and high leverages and thus provides some means to identify influential points.



Here we observe 3 points to be outliers.But we still check with another measure DFFITS which may capture those points which Cook's D might have missed out.It is defined for the i^{th} observation as $DFFITS_i = t_i(\frac{h_i}{1-h_i})^{\frac{1}{2}}$
The cutoff point is taken as $2\sqrt{p/n}$ here.



DFFITS identify the same points as outlier as the Cook's D had indicated.This is not surprising since the expressions of both the measures are quite similar upto a certain extent.

Conclusion:

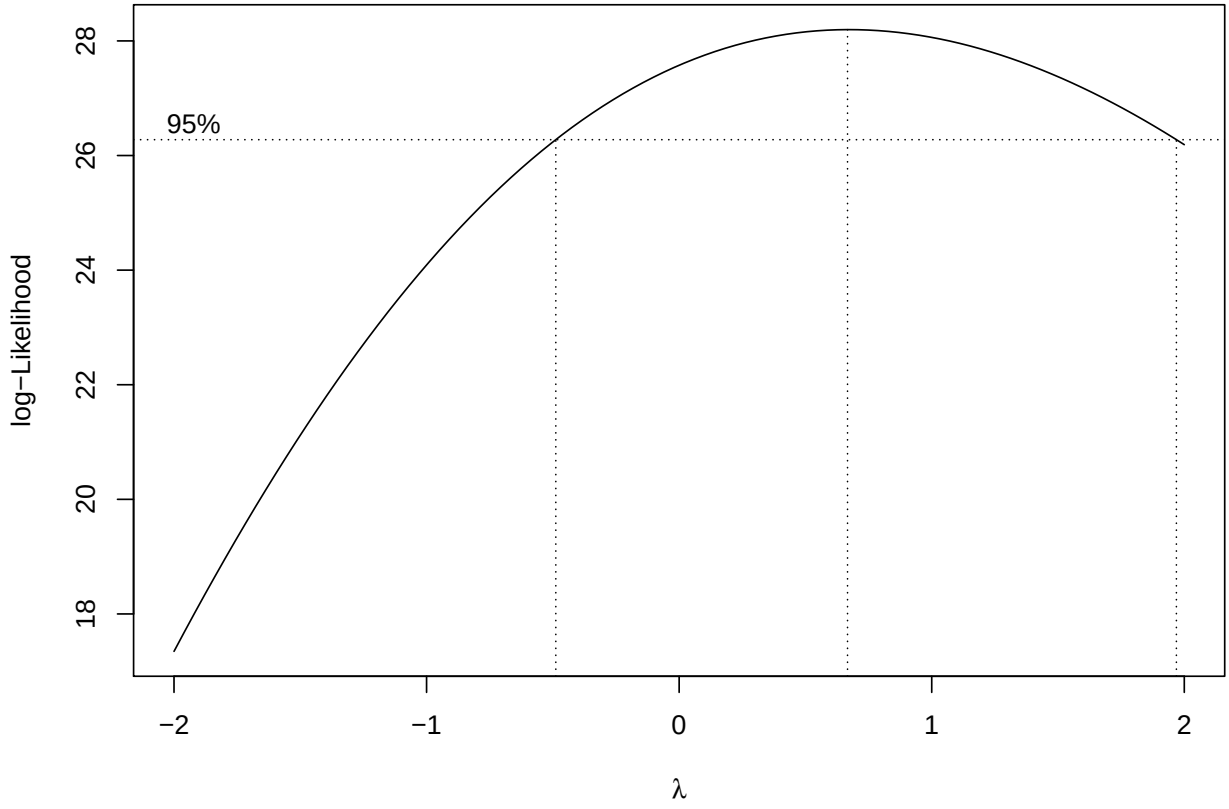
We will drop the $19^{th}, 40^{th}$ and 45^{th} observations as they are identified to be influential by both Cook's D and DFFITS.
Now , as we have eleminated outliers from our data , so we are ready to do the Box Cox transformation.
Let $Y_i^{(\lambda)}$ be the transformed response.This method assumes that there is a parameter λ such that

$$Y_i^{(\lambda)} = g(Y_i, \lambda) = x_i^T \beta + \varepsilon_i$$

where
$g(Y, \lambda) = \frac{Y^\lambda - 1}{\lambda}, \lambda \neq 0$
$g(Y, \lambda) = log(Y), \lambda = 0$

First we plot the log likelihood for a series of values of the tuning parameter λ and take that value for which the log likelihood is maximised.Then with that value of λ we transform the response.

```
Attaching package:  'MASS'
The following object is masked from 'package:olsrr':
      cement
```



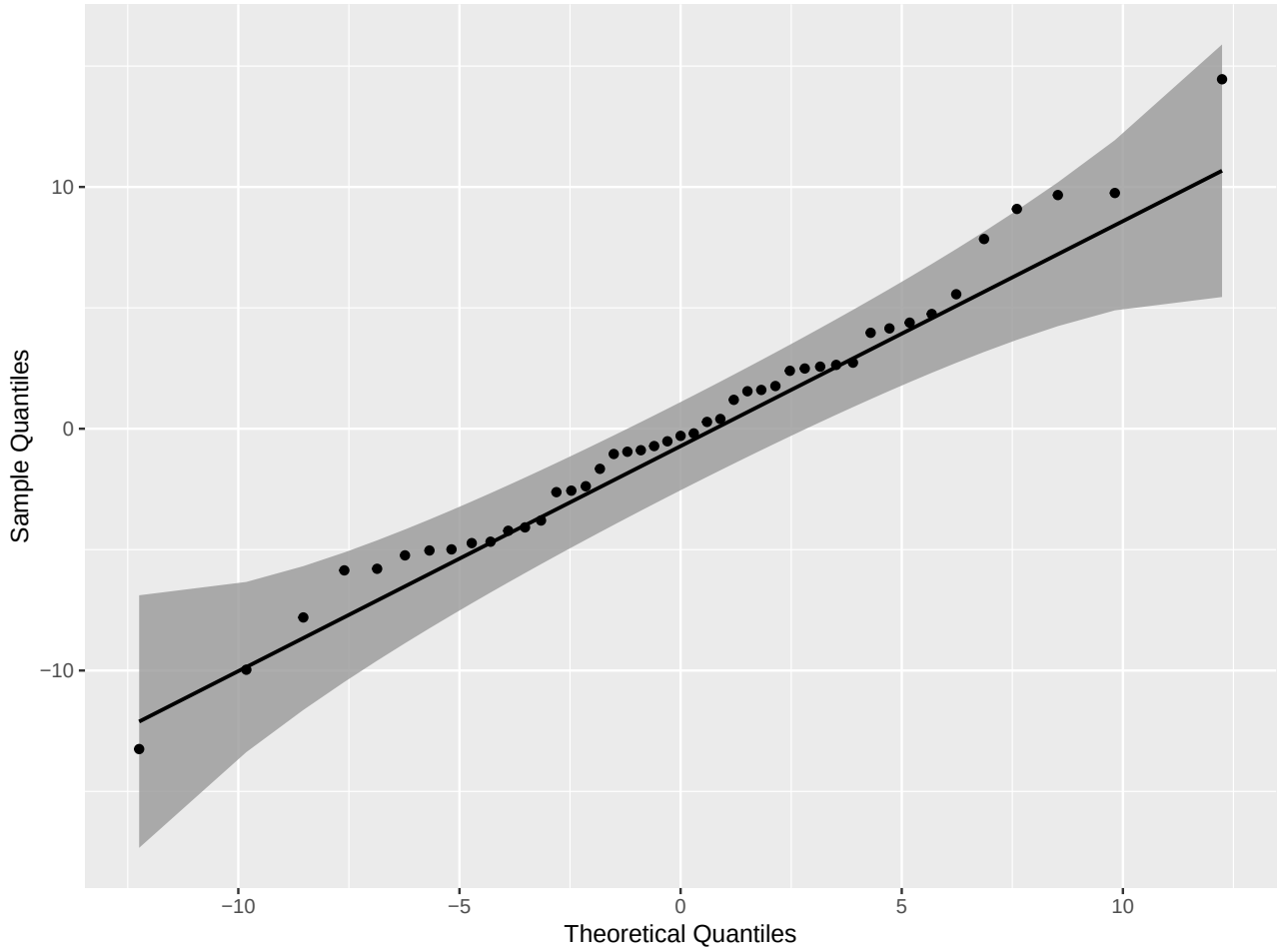

```
[1] 0.6666667
```

Now we fit the same simple linear regression model with all the 4 co variates with the transformed response $y^{(\lambda)}$.

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = df1)

Residuals:
    Min      1Q  Median      3Q     Max
-13.248  -4.080  -0.295   2.640  14.454

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.702e+01  1.666e+01   5.824 8.36e-07 ***
x1          -2.463e+00  1.203e+00  -2.048   0.0472 *
x2          -9.969e-03  1.548e-03  -6.440 1.14e-07 ***
x3           1.253e-04  3.103e-04   0.404   0.6884
x4           1.122e+02  1.879e+01   5.970 5.20e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 5.679 on 40 degrees of freedom
Multiple R-squared:  0.7087,Adjusted R-squared:  0.6795
F-statistic: 24.33 on 4 and 40 DF,  p-value: 2.941e-10
```

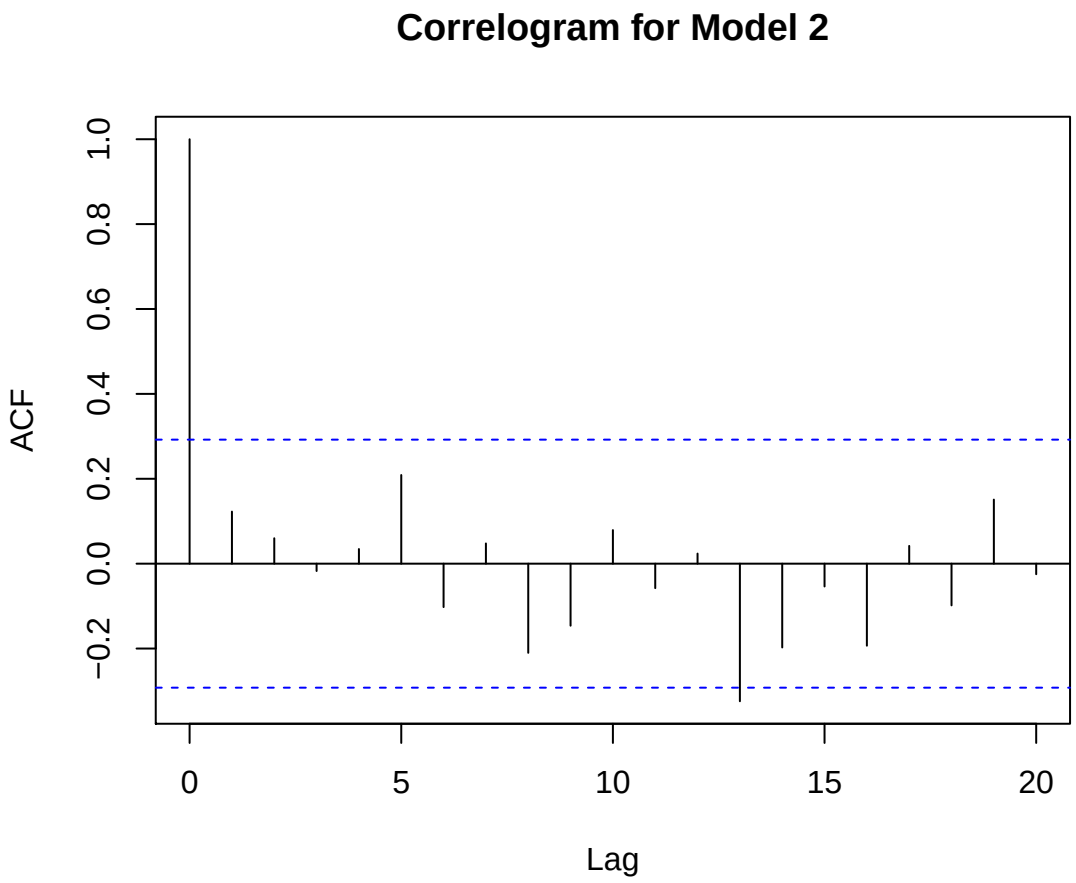
We observe that our co-efficient of determination has increased ($R^2 = 0.7087$) which is an indicator that we were successful in discarding influential points. Now,let us check again whether the assumptions of linear regression are satisfied or not.Since, we did the Box Cox Transformation to get close to Normality assumption,let us check that first.



From the Normal QQ Plot , now it seems that the distribution of the sample being tested is close to normal. Let us perform the Shapiro Wilks Test to confirm that. The p-value comes out to be $0.8021 > 0.05 = \alpha$ which is our desired level of significance.Thus,we conclude that now the assumption of normality holds.
Now,we are in a position to perform the Breusch Pagan test , the p-value comes out to be 0.86>0.05, i.e. we have no evidence to suspect heteroscedasticity.

Another important assumption of the classical linear regression model is that autocorrelation is not present in the model i.e. the error term related to any observation is not influenced by the error term related to any other observation. If present, there may be different problems in our ideal setup. So, we want to detect if autocorrelation is present in our model or not. We first plot the autocorrelogram.

```
Loading required package:  carData
```



From a visual inspection it seems that there is no evidence of autocorrelation to be present.But, we will use the Durbin Watson test to confirm our findings. The p-value for the Durbin Watson test(testing against the alternative $H_1 : \rho \neq 0$ comes out to be $0.318 > 0.05 = \alpha$ which is our desired level of significance. Thus,we conclude that autocorrelation is not present in our model.
Now, we check whether collinearity is present in our model or not, since it may lead to serious issues in the variance of estimates. There are several ways to detect multicollinearity.At first we take a look at the pairwise correlations. If any of the pairwise correlations are large , say >0.8 , we suspect collinearity to be present.

```
           x1         x2          x3          x4
x1  1.0000000 0.10399921 -0.59963864 -0.18032905
x2  0.1039992 1.00000000  0.09038648  0.03422025
x3 -0.5996386 0.09038648  1.00000000 -0.01362626
x4 -0.1803291 0.03422025 -0.01362626  1.00000000
```

No pair of explanatory variables have enough high correlation among them to suspect collinearity. But, we proceed to check with a more popular measure called Variance Inflation Factor or we may use Tolerance alternatively. High VIF and low Tolerance indicate the presence of collinearity.

```
$vif_t
  Variables Tolerance      VIF
1        x1 0.5774498 1.731752
2        x2 0.9456295 1.057497
3        x3 0.5993434 1.668493
4        x4 0.9372729 1.066925
```

We note that , the values of VIF are not at all large (not even >2.5) , equivalently the values of Tolerance are not low enough to suspect collinearity. Thus we conclude that there is no evidence of collinearity to be present in our model.

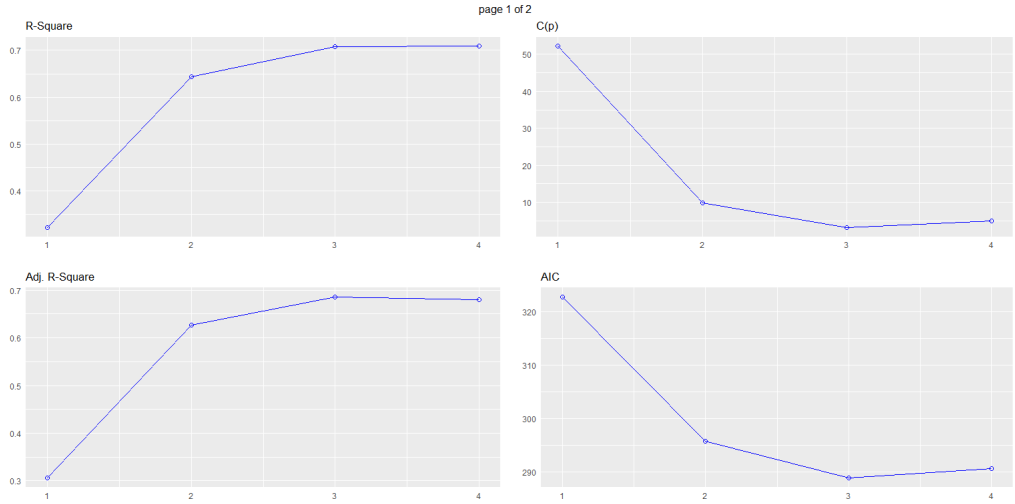
So,this model seems to be fine as all the standard assumptions hold and R^2 is also moderately good. But,let us take a close look once more at our fitted model.

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = df1)

Residuals:
    Min      1Q  Median      3Q     Max
-13.248  -4.080  -0.295   2.640  14.454

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.702e+01  1.666e+01   5.824 8.36e-07 ***
x1          -2.463e+00  1.203e+00  -2.048   0.0472 *
x2          -9.969e-03  1.548e-03  -6.440 1.14e-07 ***
x3           1.253e-04  3.103e-04   0.404   0.6884
x4           1.122e+02  1.879e+01   5.970 5.20e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.679 on 40 degrees of freedom
Multiple R-squared:  0.7087,Adjusted R-squared:  0.6795
F-statistic: 24.33 on 4 and 40 DF,  p-value: 2.941e-10
```

We can see that the p-value of the t-test for testing the significance of the explanatory variable x_3 i.e. $H_0 : \beta_3 = 0$ is $0.688 > 0.05 = \alpha$, which implies that β_3 is not significant i.e. the variable x_3 is having no significant contribution to explain the total variability.So,this phenomenon forces us to think about choosing the best subset of predictors. Now , how do we select the model? There are different procedures.We are going to use a stepwise method known as "Forward Selection" . We will get several models and choose the "Best" one according to some criterias, measures which can assess how good a model performs.So,we proceed to do that.

The result of the forward selection method is as follows:

Model	R^2	Adjusted R^2	Mallow's C_p	AIC
$y = \beta_0 + \beta_2 x_2 + \varepsilon$	0.3215	0.3058	52.175	322.7524
$y = \beta_0 + \beta_2 x_2 + \beta_4 x_4 + \varepsilon$	0.6437	0.6267	9.928	295.7752
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \varepsilon$	0.7075	0.6861	3.1632	288.8922
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$	0.7087	0.6795	5	290.7090

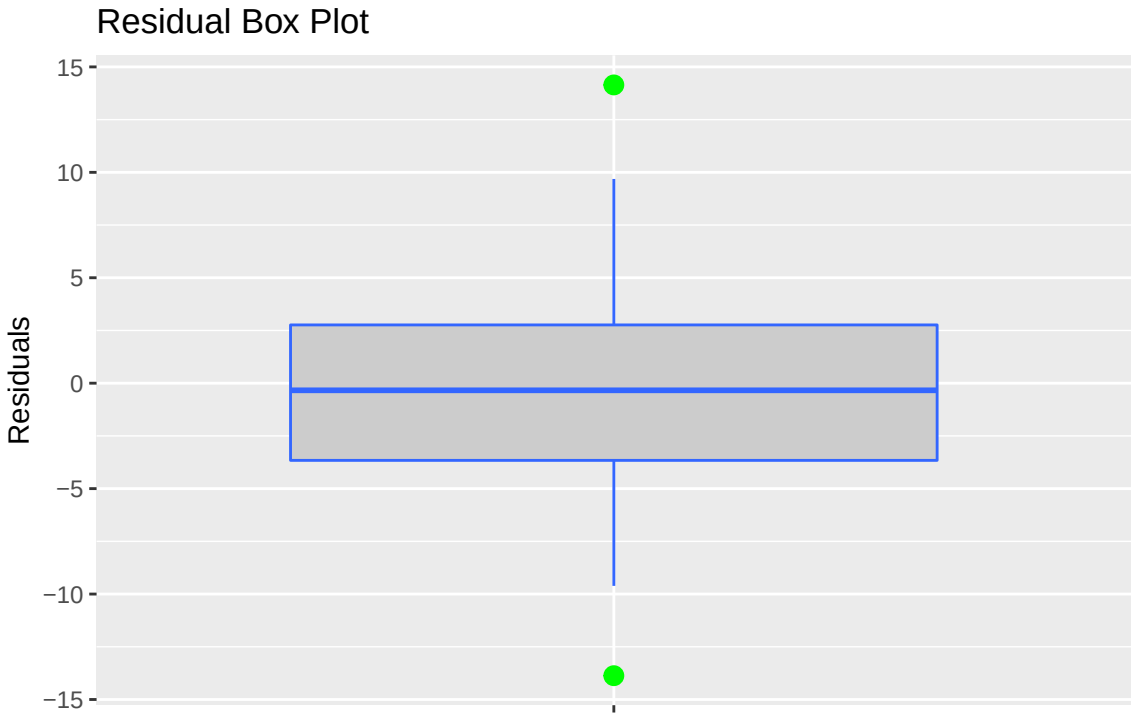


Observing the values of R^2,adjusted R^2, Mallow's C_p and AIC(Akaike Information Criterion) we choose the best subset of explanatory as x_1, x_2, x_4 and fit our final model based only on these 3 variables i.e.
$E(y^{(\lambda)}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4$

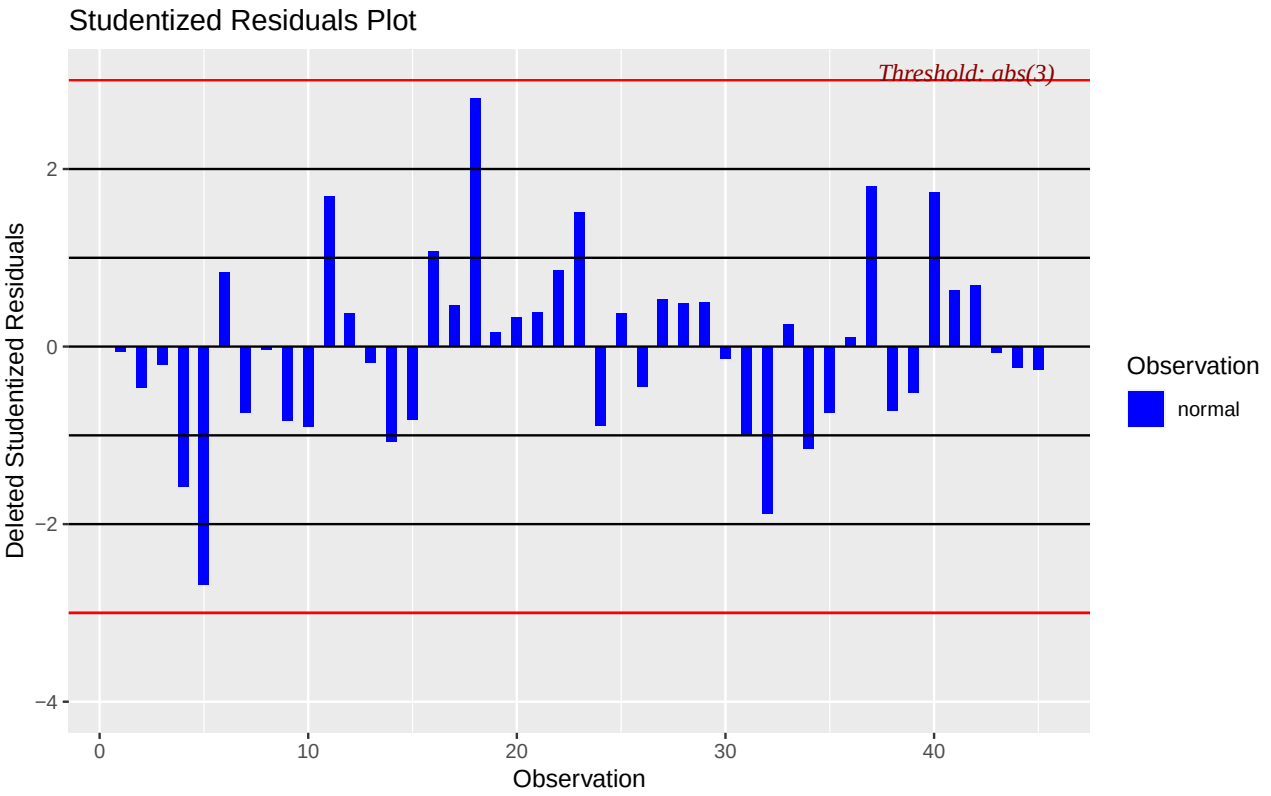
```
Call:
lm(formula = y ~ x1 + x2 + x4, data = df1)

Residuals:
     Min       1Q   Median       3Q      Max
-13.8708  -3.6606  -0.3381   2.7630  14.1451

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 100.280234  14.427378   6.951 1.93e-08 ***
x1           -2.768207   0.925443  -2.991  0.00469 **
x2           -0.009843   0.001500  -6.560 6.91e-08 ***
x4          110.878269  18.326806   6.050 3.67e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.62 on 41 degrees of freedom
Multiple R-squared:  0.7075,Adjusted R-squared:  0.6861
F-statistic: 33.06 on 3 and 41 DF,  p-value: 5.023e-11
```

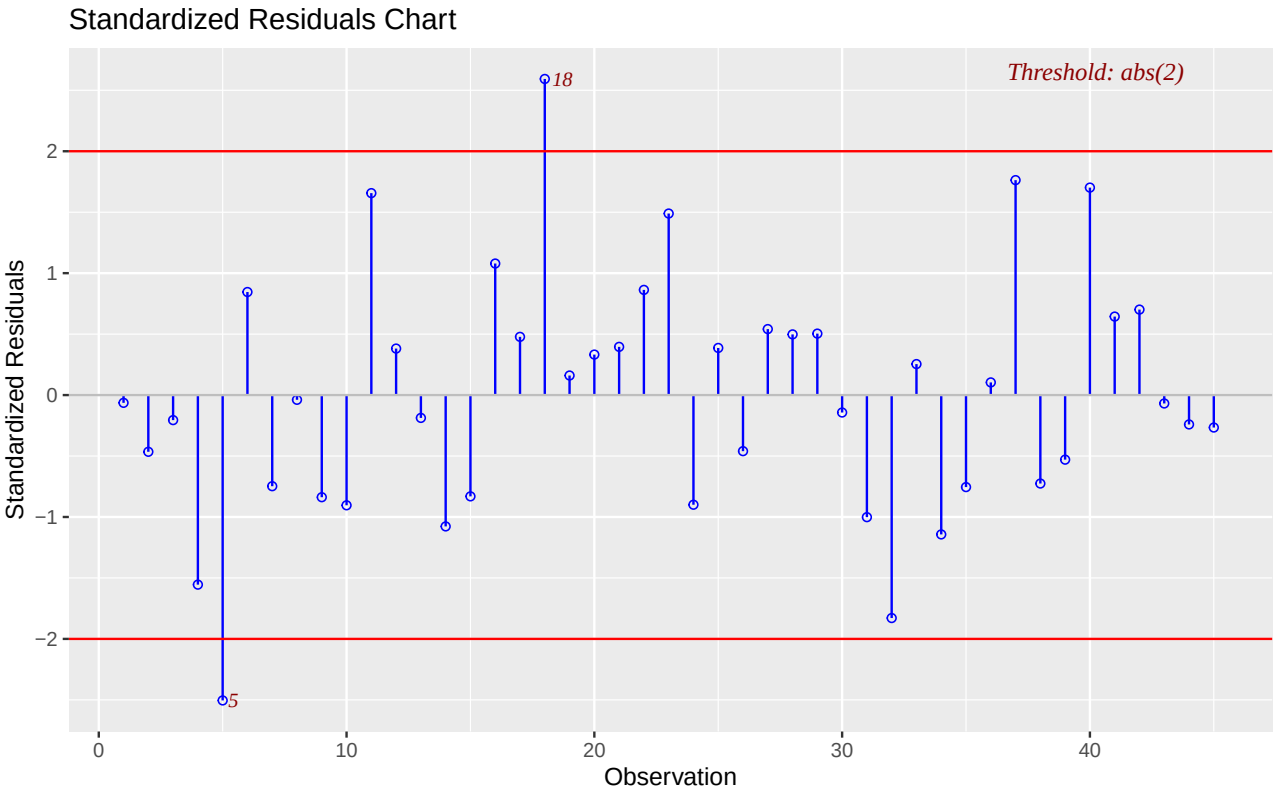
Now,since we have fitted a new model,we will again do the same diagnostics done earlier and check whether the standard assumptions hold.

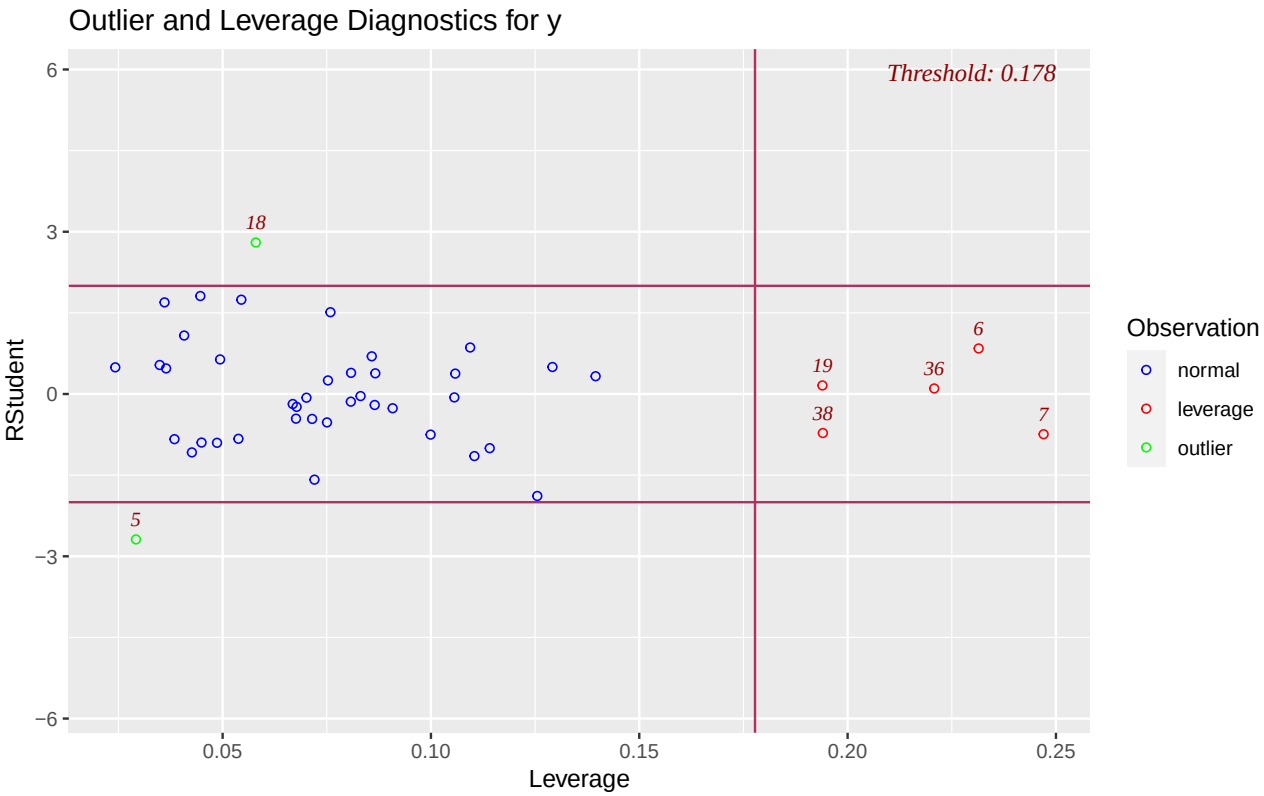
At first, we do the outlier,leverage diagnostics again.We will use the same plots and measures as earlier. Let us start with the raw residual plot.

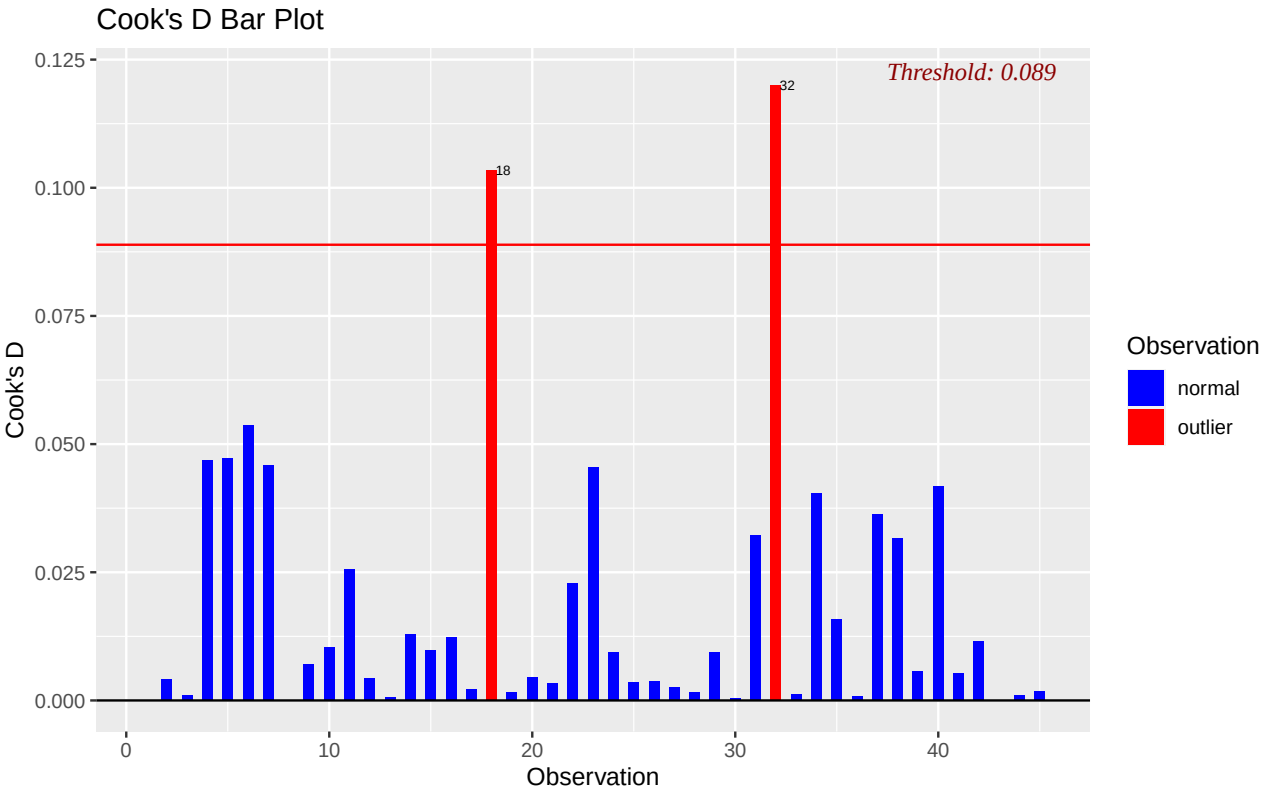


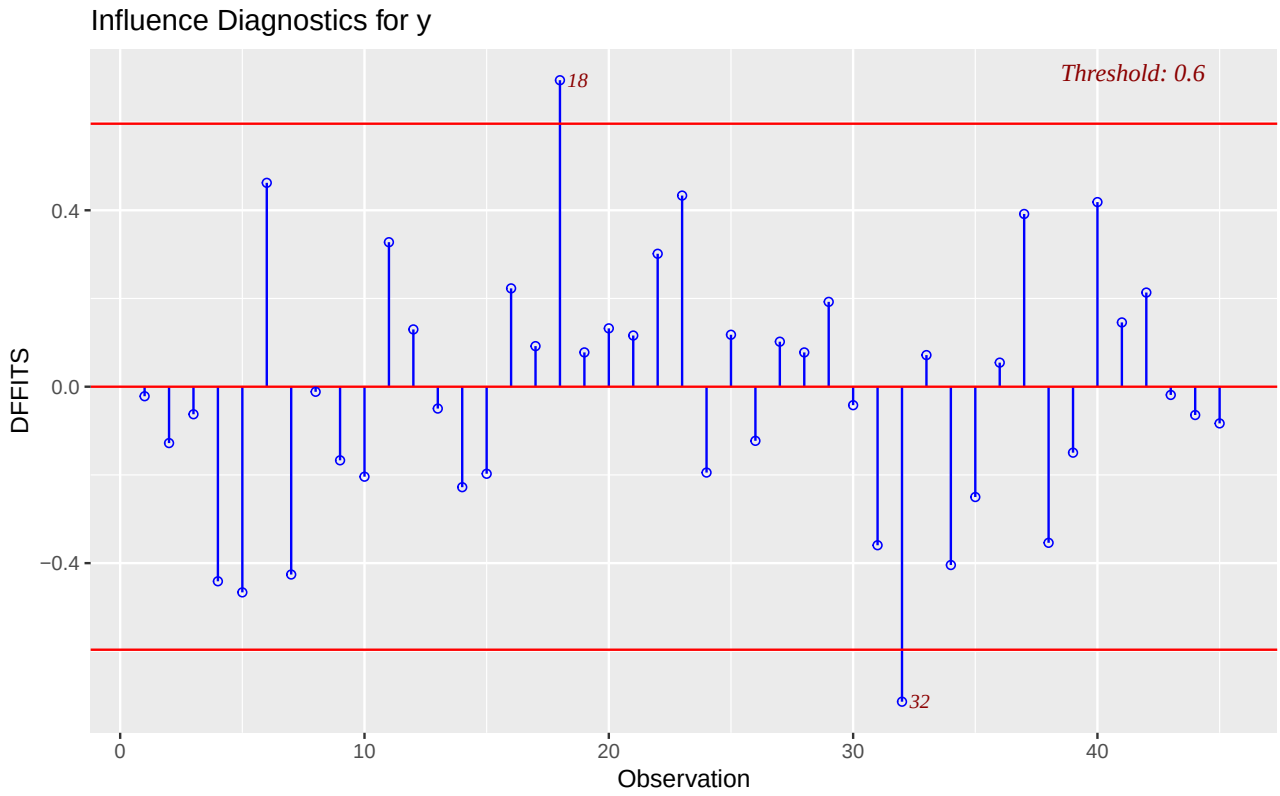
Clearly we can see that there are two potential outliers,but again we will see some more plots befor concluding.


Studentized Residuals Plot
Threshold: abs(3)
Deleted Studentized Residuals
2
0
−2
−4
0
10
20
30
40
Observation
Observation
normal


Standardized Residuals Chart
Threshold: abs(2)
18
5
Standardized Residuals
2
1
0
−1
−2
0
10
20
30
40
Observation


Outlier and Leverage Diagnostics for y
Threshold: 0.178
RStudent
Leverage
18
5
19
38
36
6
7
Observation
normal
leverage
outlier


Cook's D Bar Plot
Threshold: 0.089
Cook's D
Observation
18
32
Observation
normal
outlier



Both Cook's D and DFFITS identify the 18^{th} and 32^{nd} observations to be outliers.Thus,we drop them and fit our model.

```
Call:
lm(formula = y ~ x1 + x2 + x4, data = df2)

Residuals:
     Min       1Q   Median       3Q      Max
-13.6874  -3.5633  -0.2955   2.8905   9.8120

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  94.819005  13.191670   7.188 1.19e-08 ***
x1           -2.335081   0.839784  -2.781  0.00831 **
x2           -0.010258   0.001431  -7.168 1.26e-08 ***
x4          117.570783  16.559112   7.100 1.56e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.043 on 39 degrees of freedom
Multiple R-squared:  0.7562,Adjusted R-squared:  0.7374
F-statistic: 40.31 on 3 and 39 DF,  p-value: 4.974e-12
```

Finally,we will check whether the standard assumptions of classical linear regression model hold or not.

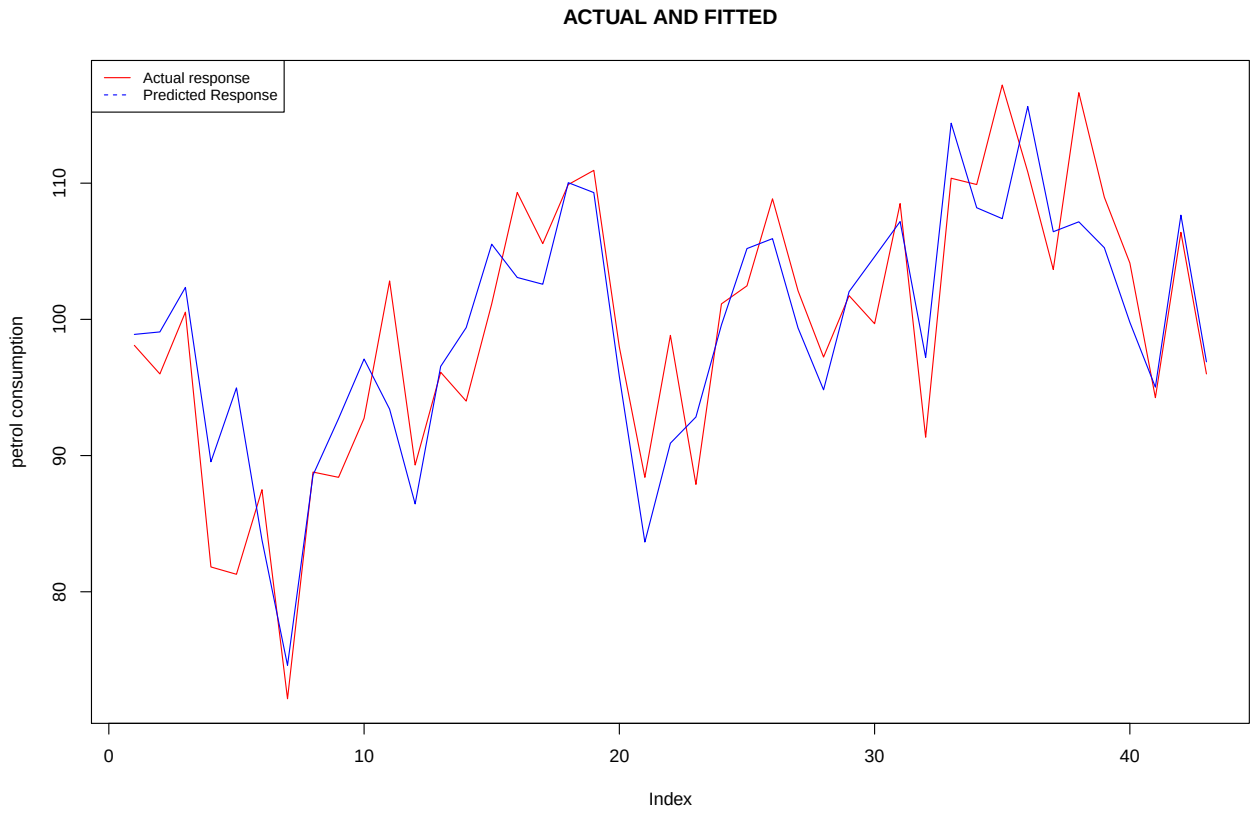
Test	p-value
Breusch Pagan Test (for heteroscedasticity)	0.726668
Shapiro Wilks Test (for Normality)	0.4685
Durbin Watson Test (for Autocorrelation against $H_1 : \rho \neq 0$)	0.386

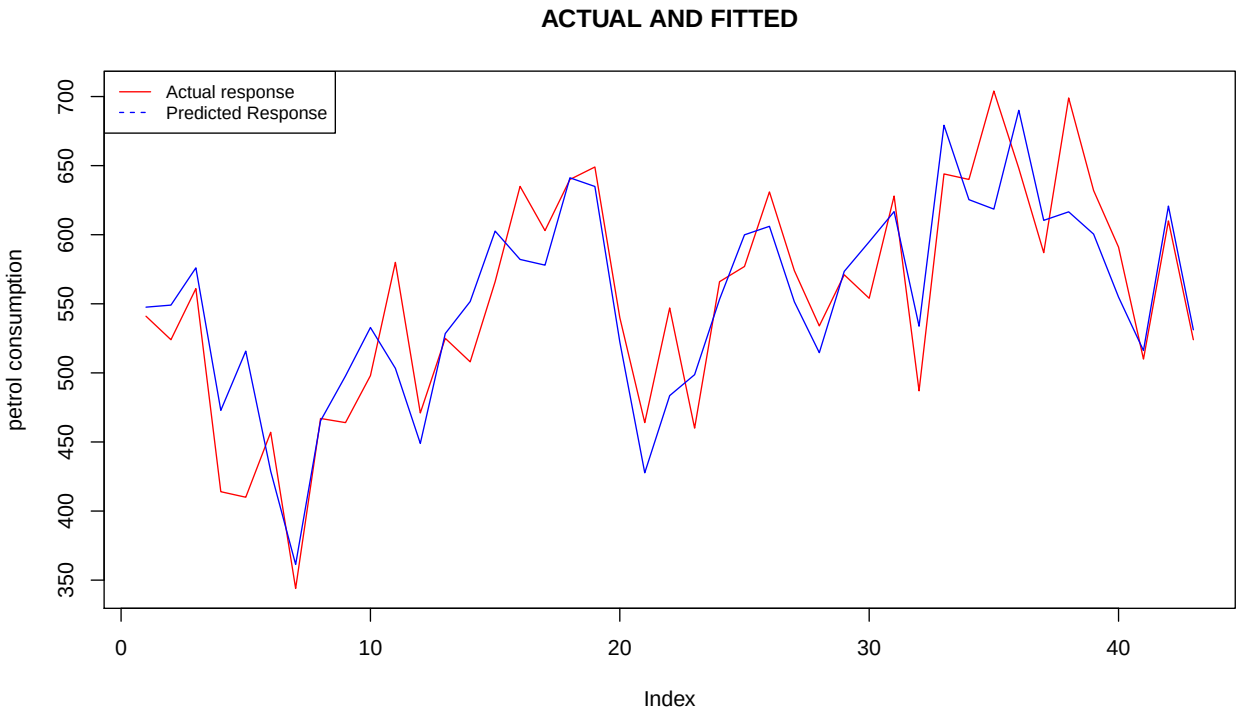
Next we do the collinearity diagnostics.

```
$vif_t
  Variables Tolerance      VIF
1        x1 0.9491353 1.053591
2        x2 0.9841836 1.016071
3        x4 0.9585887 1.043200
```

The values of the VIF's are not at all large and the tolerances are quite large,which gives no evidence to suspect collinearity to be present in our model. hence, all the conditions needed are satisfied.

We finally see the plot of predicted values and observed values.





Comment:

Thus,we see that there is a nice agreement between the observed and the predicted values and the coefficient of edetermination R^2 is also moderately high, so we can say that this is a good fit.

Conclusion:

We started with the very basic classical linear regression model and took the necessary steps to make our model better and ensured all the standard assumptions namely homoscedasticity,normality,no presence of autocorrelation and multicollinearity to be satisfied and came up with this final model, eventually in due course we also got rid of influential points and one insignificant explanatory variable namely x_3:the number of miles of paved highway(in miles). Hence,we conclude the optiomal model to be:

$$E(y^{(\lambda)}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4$$

where
y : consumption of petrol (in gallons)
$y^{(\lambda)}$:Transformed y after Box Cox transformation
x_1:the petrol tax(in cents per gallon)
x_2:the average income per capita(in dollars)
x_4:the proportion of the population with driver's licenses

References

[1] Linear Regression Analysis by George A.F Seber , Alan J Lee

[2] Introduction to Statistical Learning with Applications in R by Gareth James,Daniela Witten,Trevor Hastie,Robert Tibshirani

[3] https://www.isid.ac.in/~deepayan/Mysore-University-2019/rvisualization.html

[4] https://cran.r-project.org/web/packages/olsrr/vignettes/intro.html

Acknowledgement

We are grateful to our professor Dr.Swagata Nandi, ISI Delhi, for the timely guidance without which the project would not have been completed on time.We also thank her for giving us the exposure regarding how to use our theoritical knowledge to real life data.