Regression Project

Analysis of Petrol Consumption Data

Debarshi Chakraborty,Sagar Dey,Nikhil Bhardwaj

Master of Statistics
Indian Statistical Institue,Delhi

Dec 22, 2021

Abstract

We are given a data , our goal is to develop a model which can predict the petrol consumption of a state based on several explanatory variables.We try to use the knowledge gained in our Regression Techniques course and implement that practically.

Data Description

- y:Consumption of Petrol (in millions of gallons)
- x_1:Petrol Tax (in cents per gallons)
- x_2:Average Income per capita (in dollars)
- x_3:Paved Highways (in miles)
- x_4:Proportion of population having driver's license

Multiple Linear Regression

We start with the most basic model. We include all the 4 covariates and the intercept term.

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

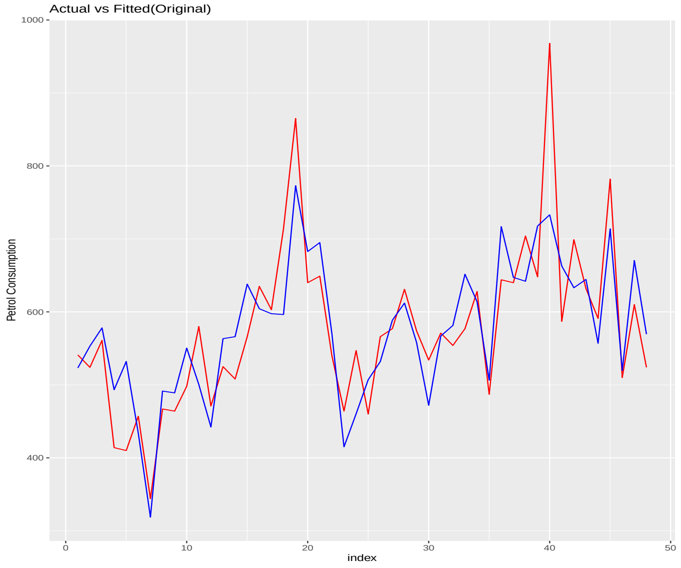
The estimates obtained are the ordinary least square estimates.We will take a quick look at the fitted model.

Fitted Model

```
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
```

Fitted Model



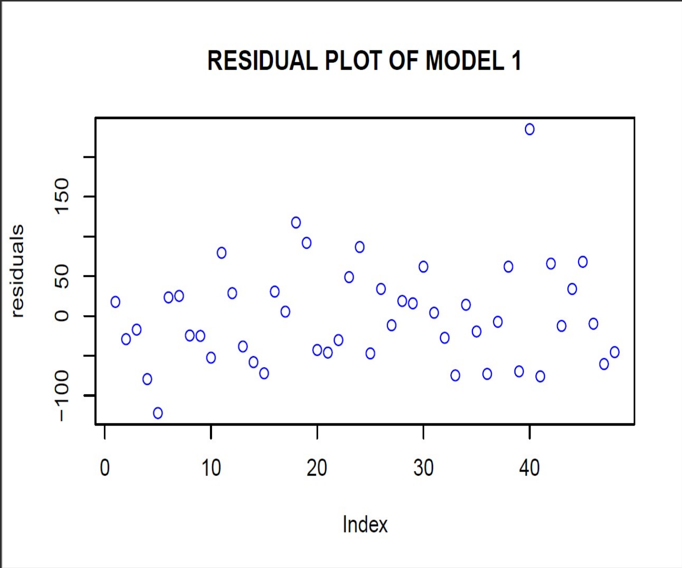
Interpretation

- $R^2 = 0.6787$
- Actual vs Fitted plot does not show very nice agreement
- Scope of improvement
- Next task is to check model assumptions

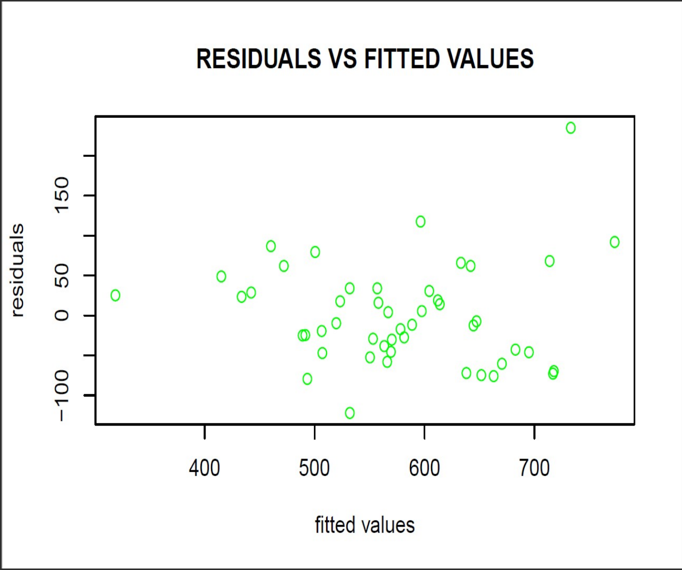
Model Assumptions

- **Homoscedasticity** : errors are assumed to have equal variances
- **Normality**: $\epsilon_i \sim N(0, \sigma^2)$ independently
- **No Autocorrelation**
- **No Collinearity**: a desirable scenario

Residual Plot



Residuals vs Fitted Values



Interpretation

It is desirable that residuals should not exhibit any particular pattern,they should be randomly scattered.Although it seems to be random, we perform a One Sample Runs Test where

H_0:observations arise from a random process
H_1:observations are not random

p-value = 0.7704 $>$ 0.05 $= \alpha$

Testing for Homoscedasticity

We will use the Breusch Pagan Test.

H_0:Errors have equal variances.

H_1:Errors have unequal variances.

p-value

$$\mathbf{0.007 < 0.05}$$

Interpretation

- Should we conclude presence of heteroscedasticity from here?

  NO

- Why?

  The test is very sensitive to violation of normality.

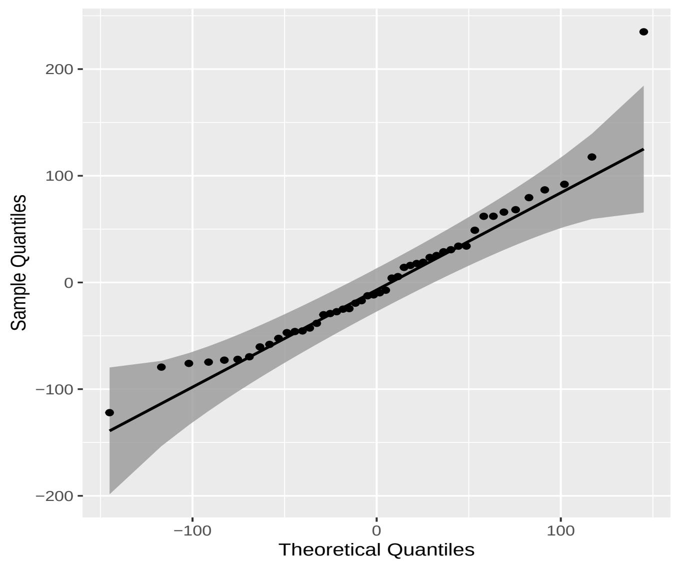
  Hence,take care of normality assumption first.

Testing for Normality

We will do it in 2 steps.

- Look at a Normal Quantile-Quantile Plot (QQ PLOT) to get some insight
- Do a formal test

The QQ Plot



200
100
0
−100
−200
Sample Quantiles
−100
0
100
Theoretical Quantiles

Interpretation

- Could not get much intuitive insights,like whether satisfies normality or skewed or heavy tailed.
- Let's proceed to the formal test.

Test for Normality

We will use the Shapiro Wilk Test for Normality

H_0:Errors are normally distributed.

H_1:Errors are non normal.

p-value=0.0151<0.05

Assumption of Normality does not hold.
We need some remedial measure.

Remedial measure for non normality

- We want to do the Box-Cox Transformation
- May not work well in presence of outliers
- Need to remove outliers first
- Do the influential diagnostics

Influential Diagnostics

We are now going to encounter the following notations,terminologies and measures very often for a while

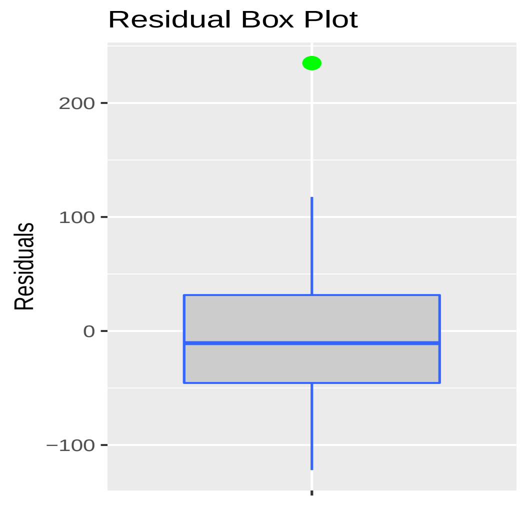
- y_i:actual value of the response in the i^{th} observation
- $\hat{y_i}$:fitted value for the i^{th} observation
- H:hat matrix and h_i:i^{th} diagonal element of the hat matrix
- $\hat{\beta}$:the usual LSE
- $S^2 = \sum_{i=1}^{n} \frac{e_i^2}{n-p}$:usual unbiased estimate of the error variance
- $\beta(\hat{i})$ and $S(i)^2$:same quantities without the i^{th} observation

Influential Diagnostics

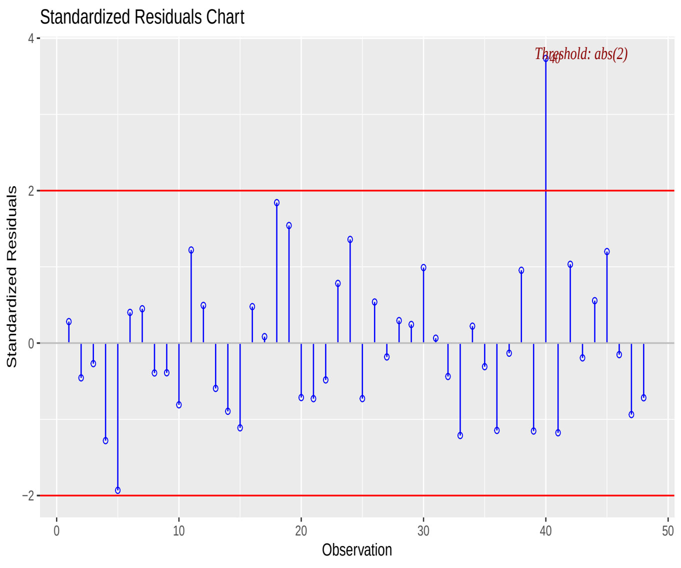
- Residual: $e_i = y_i - \hat{y}_i$
- Internally studentized residual: $r_i = \frac{e_i}{S(1-h_i)^{\frac{1}{2}}}$
- Externally studentized residual: $t_i = \frac{e_i}{S(i)(1-h_i)^{\frac{1}{2}}}$

Now let us look at some graphs

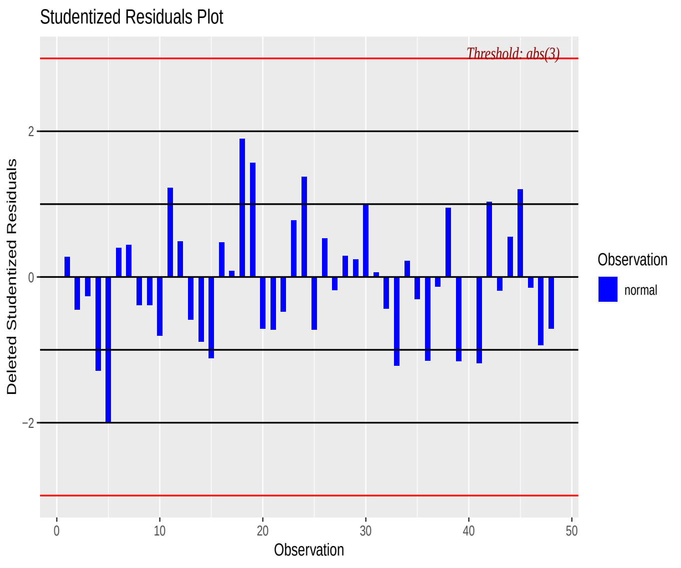
Influential Diagnostics-Plots



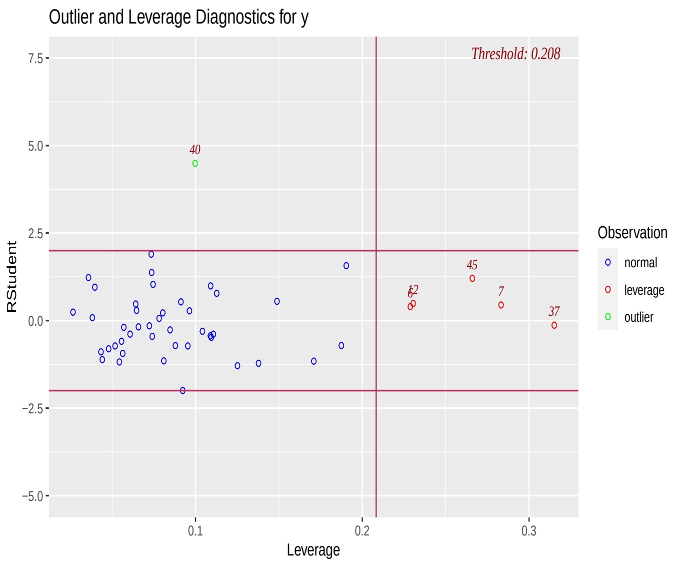
Influential Diagnostics-Plots



Influential Diagnostics-Plots



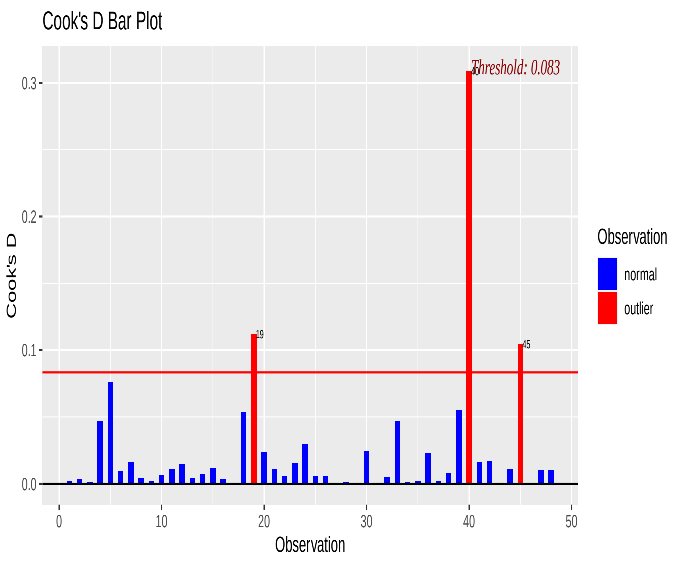
Influential Diagnostics-Plots



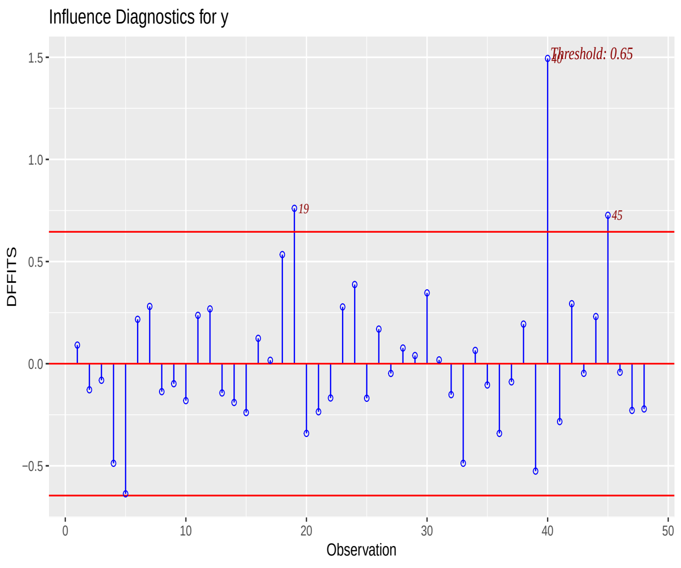
Interpretation

- The 40^{th} point is a good candidate for an outlier.
- Nothing else is clear
- We need some improved measures
- Let us look at Cook's D and DFFITS.

Infulential Diagnostics-Plots



Infulential Diagnostics-Plots



Interpretation

Both Cook's D and DFFITS identify the $19^{th}, 40^{th}, 45^{th}$ data points to be influential.This is natural since the mathematical expressions of both the measures look quite similar.

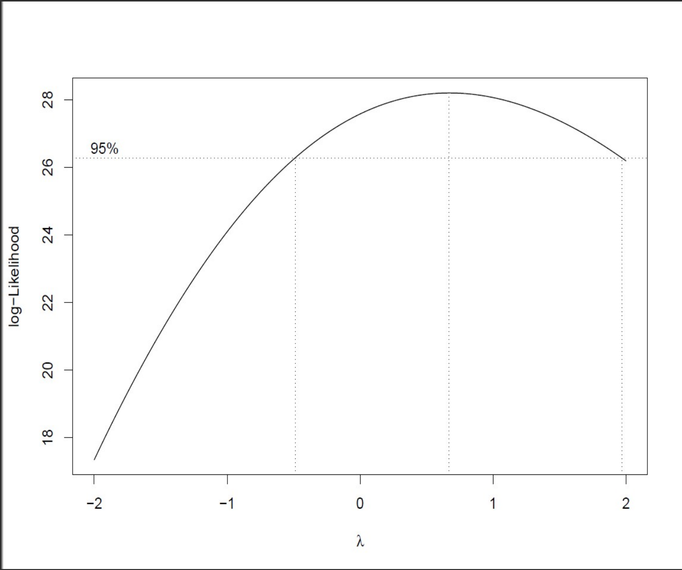
Thus we discard them from our data and continue our analysis.

Box Cox Transformation

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

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

Selecting λ



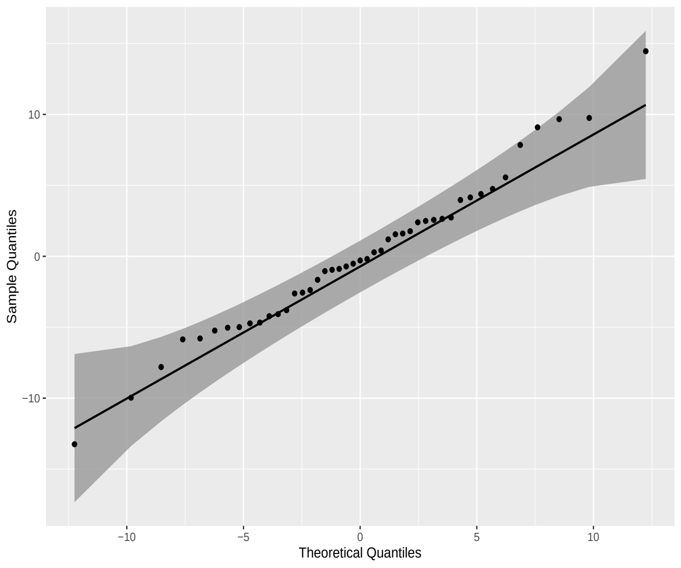
Model with transformed response

We fit the same model i.e. with all the 4 covariates , just the response being transformed.

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

Now,let us check whether the Normality assumption holds.

Normal QQ plot



Test for Normality

We will use the same Shapiro Wilk test.

p-value=$0.8021>0.05$

Hence,the normality assumptions holds in this model.

What about homoscedasticity now?

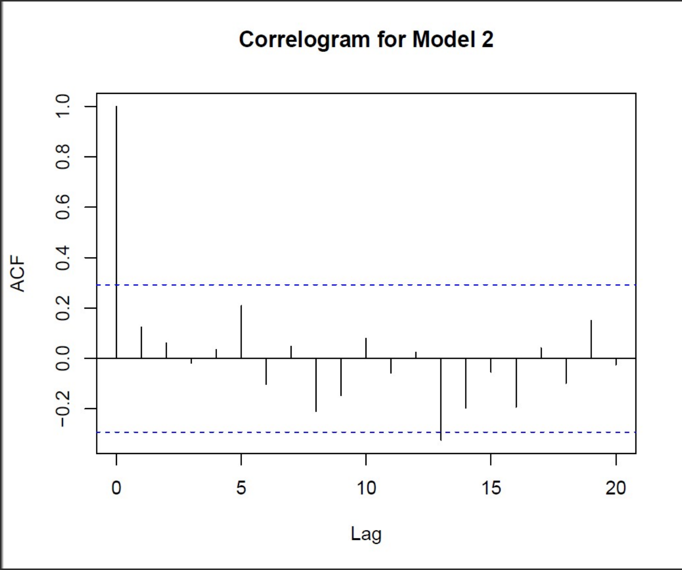
Now since the normality assumption holds,we can do the
Breusch Pagan Test

p-value=0.86>0.05

Hence, now we can say the errors are homoscedastic.

Now,we have to check another important assumption which the classical linear regression model should satisfy-no autocorrelation.

Presence of Autocorrelation?



Test for Autocorrelation

We shall use the Durbin Watson Test.

$H_0 : \rho = 0$
$H_1 : \rho \neq 0$

p-value=0.318>0.05

Hence,not enough evidence to suspect autocorrelation.

Detecting Collinearity

At first we check the pairwise correlations to get an idea.

	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 covariates exhibit high enough correlation among them to suspect collinearity.

Detecting Collinearity

Now,we examine a more popular measure for detecting collinearity-
High VIF or equivalently low tolerance

	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

VIFs are not at all high to suspect collinearity.

Is this the correct model?

Let us take a close look at our new model once again.

```
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
```

Interpretation

Some points to be noted:

- R^2 has increased which is an indication that we were successful in discarding influential points.
- Apparently the model seems to be fine.
- If we want to check the significance of β_is,we observe that β_3 is not significant.
- This observation leads us to think about selecting the best subset of explanatory variables.

Model Selection

We are going to use forward selection method.

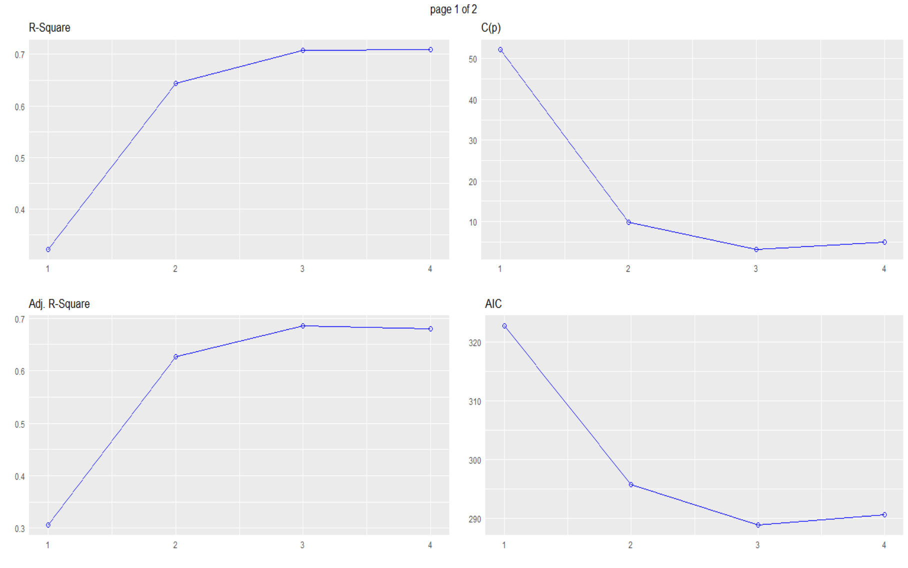
To assess how good a model is , we will use different criterions like

- Adjusted R^2
- Mallow's C_p
- Akaike Information Criterion (AIC)

Model Selection

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

Model Selection



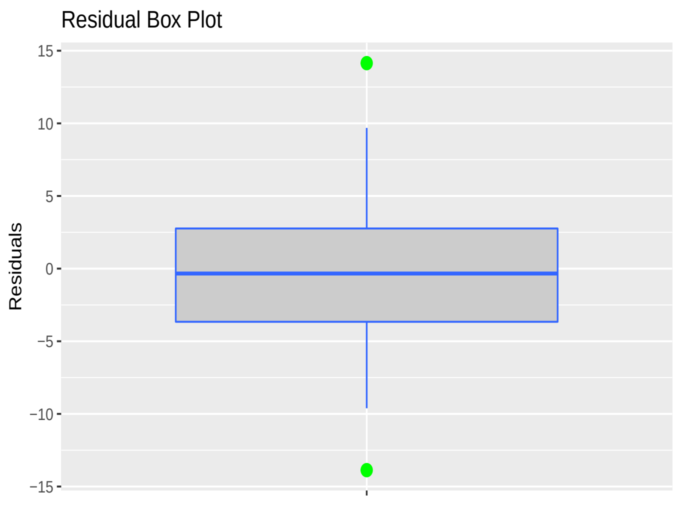
Model Selection

We choose our optimum model to be

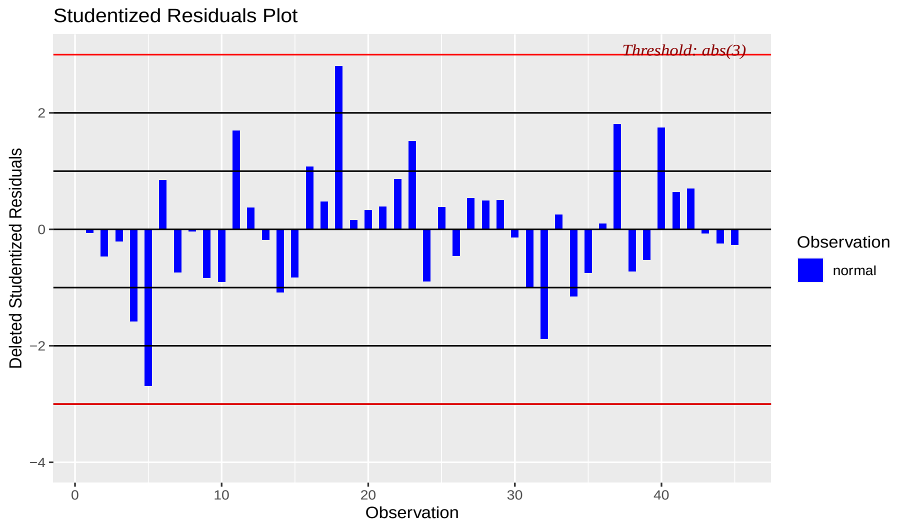
$$E(y^{(\lambda)}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4$$

Finally,we will check the standard assumptions and since we have fitted a new model by dropping one predictor,we will try to get rid of influential points too.

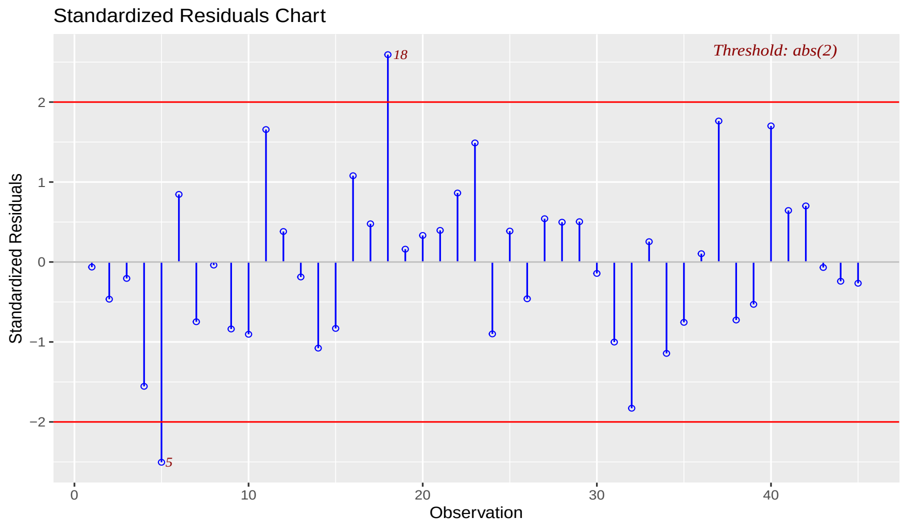
Influential diagnostics for final model



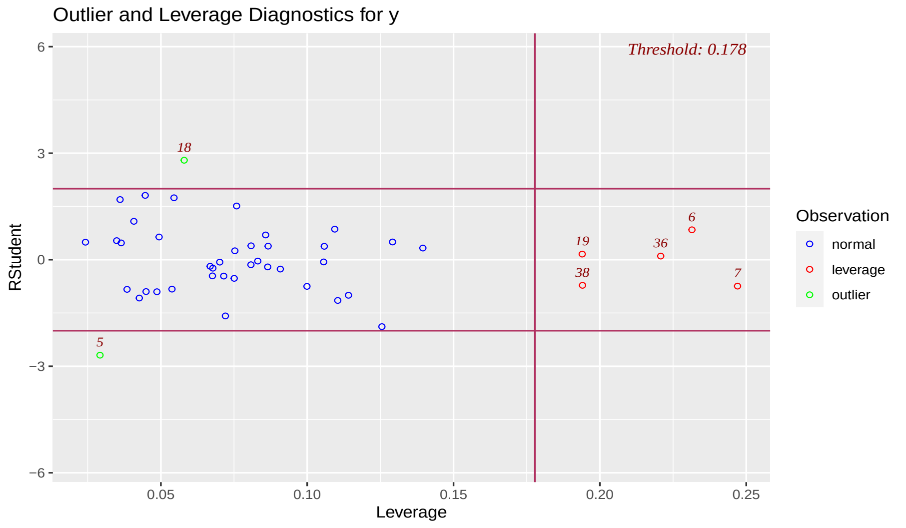
Influential diagnostics for final model



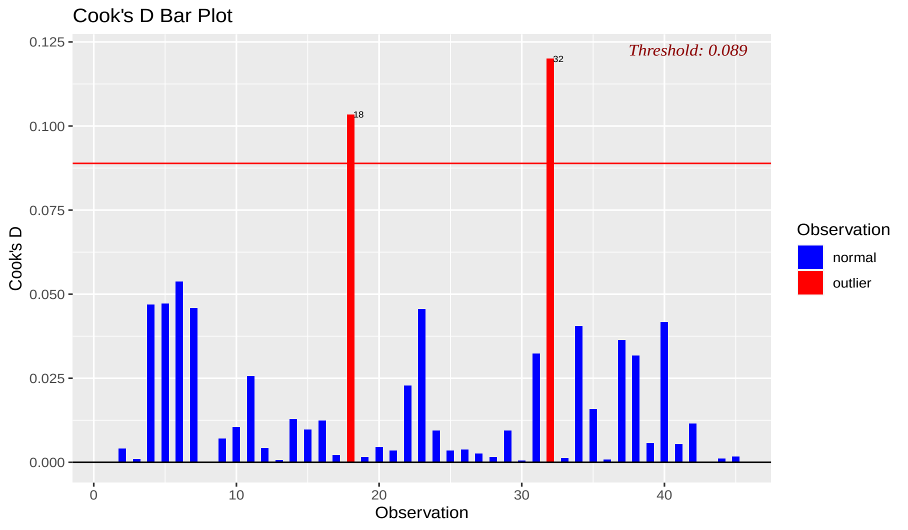
Influential diagnostics for final model



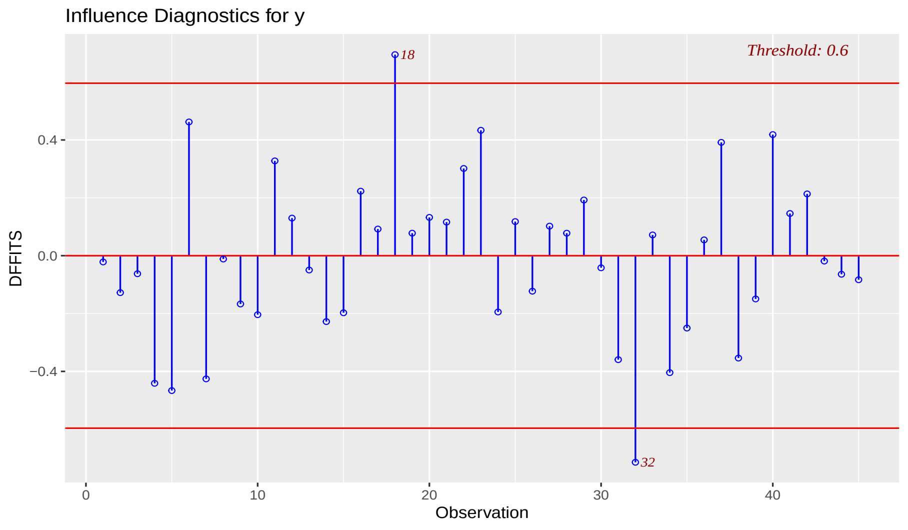
Influential diagnostics for final model



Influential diagnostics for final model



Influential diagnostics for final model



Final Model

We discard the 18^{th} and 32^{nd} point as they turn out to be influential.

Now we finally check whether our model violates any of the standard assumptions.

Checking Standard Assumptions

Table: Test Results for Assumptions

Assumption	**Name of Test**	**p-value**
Heteroscedasticity	Breusch Pagan Test	0.72
Normality	Shapiro Wilk Test	0.47
Autocorrelation	Durbin Watson Test	0.39

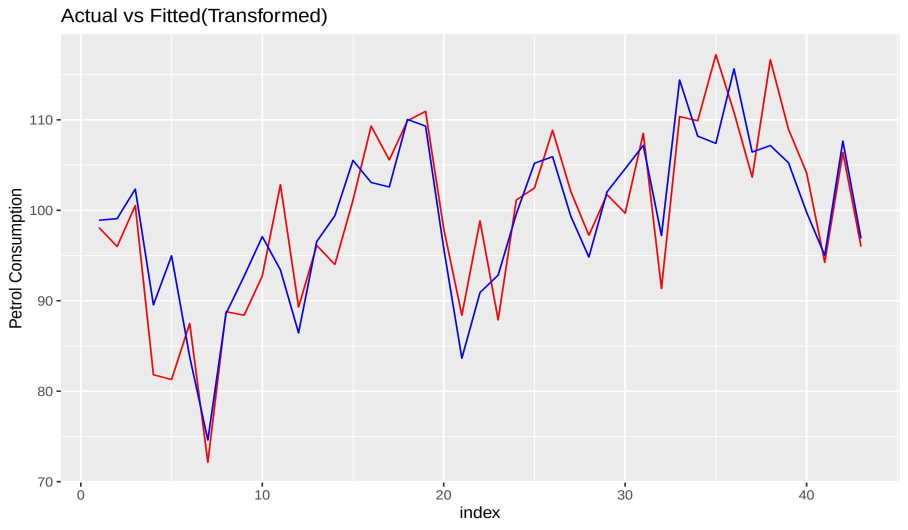
Table: Collinearity diagnostics

Variables	Tolerance	VIF
x_1	0.95	1.05
x_2	0.98	1.02
x_4	0.96	1.04

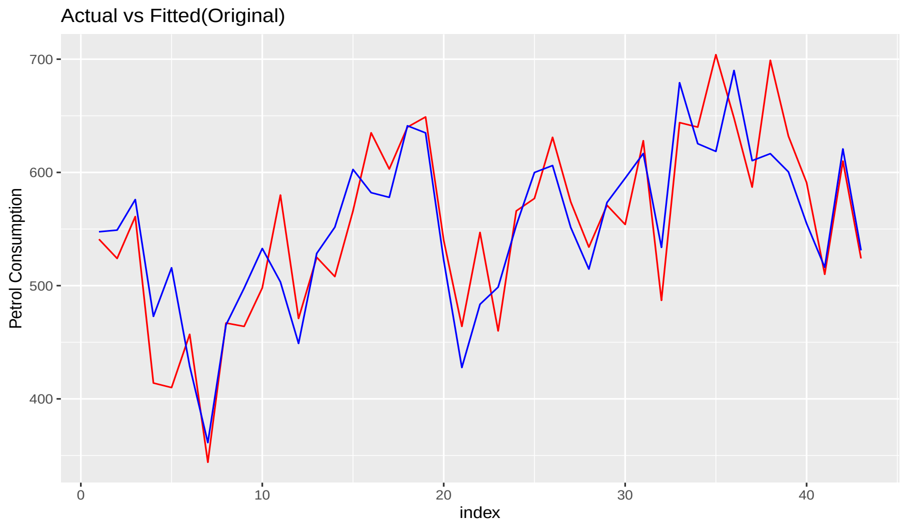
Conclusion

- The errors are homoscedastic.
- The errors are normally distributed.
- No evidence of autocorrelation.
- No evidence to suspect collinearity.
- $R^2 = 0.7562$: far better than from where we started.

Visualizing Agreement between predicted and actual



Visualizing Agreement between predicted and actual



Conclusion

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.

Thank You