REGRESSION TECHNIQUES PROJECT

PROJECT GUIDE : Dr. SWAGATA NANDI

AUTHORS :
ROLL No. 4 ANISHA GHOSH
ROLL NO. 14 PARAMITA ADHIKARI
ROLL NO. 15 PAYEL GHOSAL

INTRODUCTION :

Petroleum has revolutionized the entire transport network of the world. The generic product "Gasoline" more commonly known as Petrol is one of the most consumed petroleum product. However, burning of fossil fuels like petrol emits CO_2 which adds up to Global Warming. Also Petroleum is an exhaustible natural resource. Hence, we must be careful and wise about that consumption of petrol.

For one year, the consumption of petrol was measured in 48 states. The first few rows of our data is given below:

DATA :

Serial No.	Petrol tax (cents per gallon)	Average income (dollars)	Paved Highways (miles)	Proportion of population with driver's licenses	Consumption of petrol (millions of gallons)
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

The relevant variables are claimed to be

```
1)the petrol tax- It is intuitive that tax would discourage extravagant consumption of petrol.
2)the average income per capita- The economic condition of people affects the mode of
transportation(public/private or road/railway/air) they choose.
3)the number of miles of paved highway- This is a measure of how favoured road
transportation is in a particular state.
4)the proportion of the population with driver's licenses- It is again very intuitive that
the more the drivers, more is the consumption.
```

Now we name the covariates (independent variables) as:

- **Tax** : Petrol tax (cents per gallon)
- **Income** : Average income (dollars)
- **Miles** : Paved Highways (miles)
- **L_prop** : Proportion of population with driver's licenses

and the response (dependent) variable as:

- **Petrol** : Consumption of petrol (millions of gallons)

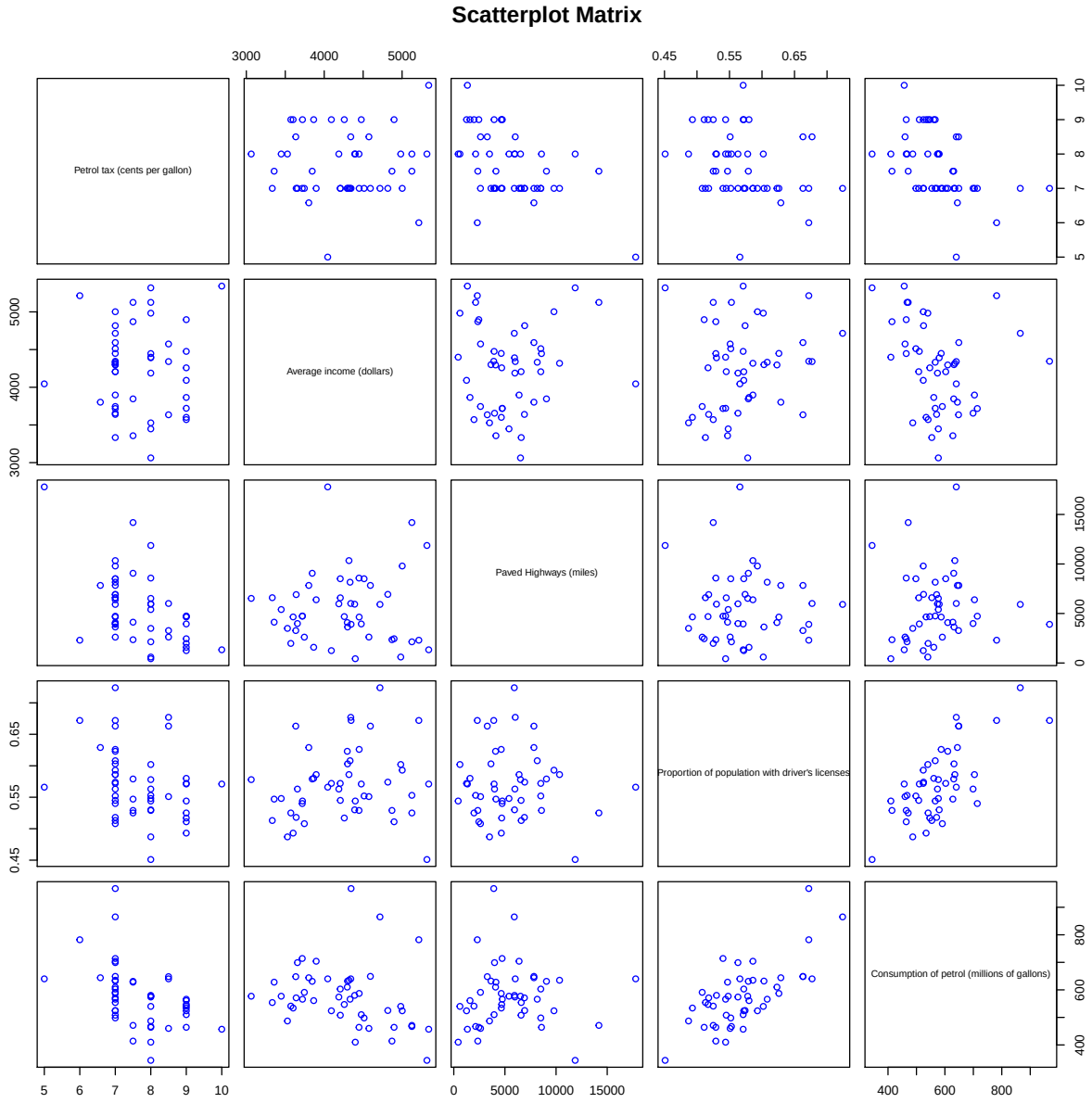
OBJECTIVE

Our aim is to carry out Statistical tests to investigate if the claim is valid. We then build a model for the consumption of petrol based on some or all of the four explanatory variables (petrol tax, average income per capita, number of miles of paved highway, and the proportion of the population with driver's licenses).

GRAPHICAL REPRESENTATION OF THE DATA :

First of all consider the following ***Scatterplot*** to get some general idea about the data :

```
plot(data[-1],col="blue",main="Scatterplot Matrix")
```



ANALYSIS :

From the scatter plot matrix, we can interpret the following things

- The variable tax does not perform like a continuous variable. It forms cluster around a finite set of points.
- Petrol consumption does not seem to be linearly related with “tax”,but the overall pattern is decreasing.

It is negatively (and seems to be linearly) related with the variable “average income”.

It is positively and linearly related with the “proportion of population with driver’s licenses”.It does not seem to be strongly related with the variable ‘miles’.

- Miles is not related with income and proportion, but it seems to be negatively related to the variable tax. There is no visible relationship between tax and income and between tax and proportion. Income and proportion may be related to each other but the relationship is not linear.

SELECTION OF APPROPRIATE MODEL :

We shall try to fit various linear models to the given data set by taking Petrol as response variable and selecting different subsets of explanatory variables from Tax, Income, Miles, L_prop. The ultimate model selection will be based on some popularly used criteria such as Residual Sum of Squares (RSS),R^2 , Adjusted R^2 (adjr),predicted R^2 , Akaike Information Criterion (AIC) and Mallow’s CP. We shall investigate the measures of all such criteria for all possible subsets of these explanatory variables. Then we will sort different models based on adjusted R^2 and choose such a model that has lower values of AIC and Mallow’s CP closer to the number of explanatory variables and high values of adjusted R^2. Here we consider scaled and centered model by defining a function f such that :

```
f<-function(x)
{
  x=(x-mean(x))/sqrt(sum((x-mean(x))^2))
  return (x)
}
```

POLYNOMIAL FITTING :

Taking a hint from the scatter plot, We fit a model for predicting the response variable depending upon income,miles,proportion and a polynomial of tax of degree 2 :

```
pfit=lm(Petrol~f(Tax)+I(f(Tax)^2)+f(Income)+f(Miles)+f(L_prop))
summary(pfit)
```

```
##
## Call:
## lm(formula = Petrol ~ f(Tax) + I(f(Tax)^2) + f(Income) + f(Miles) +
##     f(L_prop))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -119.66  -41.93  -12.40   33.61  237.07
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   571.55      11.70  48.862  < 2e-16 ***
## f(Tax)       -237.12      85.94  -2.759 0.008551 **
## I(f(Tax)^2)   250.83     319.84   0.784 0.437298
## f(Income)    -266.22      68.26  -3.900 0.000341 ***
## f(Miles)      -70.04      82.91  -0.845 0.403033
## f(L_prop)     503.34      73.72   6.827 2.57e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 66.61 on 42 degrees of freedom
## Multiple R-squared:  0.6833, Adjusted R-squared:  0.6456
## F-statistic: 18.13 on 5 and 42 DF,  p-value: 1.499e-09
```

Here p-values of Tax with degree 2 is larger than default level of significance(0.05).Hence this is insignificant . And as we are working with the scaled and centered model , we may reject this polynomial fitting and move on to linear regression fit. Moreover similarly the covariate Miles is also insignificant. So for further clarification we come up with the linear fitting . Here we would like to keep Miles for the time being to get a better overview whether rejecting that covariate is necessary even in linear regression or not .

FITTING LINEAR REGRESSION :

COMPARING ALL POSSIBLE SUBSETS :

Now consider all possible linear fits :

```
ols_step_all_possible(ffit)
```

Predictors	rsquare	adjr	predrsq	cp	aic
f(L_prop)	0.4885527	0.4774342	0.4317651	24.444837	561.9010
f(Tax)	0.2036539	0.1863420	0.1457033	62.571636	583.1549
f(Income)	0.0599574	0.0395217	-0.0259644	81.801925	591.1177
f(Miles)	0.0003626	-0.0213687	-0.0907809	89.777251	594.0681
f(Income) f(L_prop)	0.6175098	0.6005102	0.5492287	9.187046	549.9550
f(Tax) f(L_prop)	0.5566811	0.5369780	0.4947850	17.327496	557.0391
f(Miles) f(L_prop)	0.4926484	0.4700994	0.4151376	25.896721	563.5151
f(Tax) f(Miles)	0.2681444	0.2356175	0.1797086	55.941146	581.1012
f(Tax) f(Income)	0.2608541	0.2280032	0.1666368	56.916769	581.5770
f(Income) f(Miles)	0.0609412	0.0192052	-0.0642389	83.670277	593.0674
f(Tax) f(Income) f(L_prop)	0.6748583	0.6526896	0.6090890	3.512335	544.1578
f(Income) f(Miles) f(L_prop)	0.6249242	0.5993509	0.5473652	10.194807	551.0154
f(Tax) f(Miles) f(L_prop)	0.5669727	0.5374481	0.4891896	17.950210	557.9117
f(Tax) f(Income) f(Miles)	0.3177633	0.2712472	0.2054910	51.300857	579.7313
f(Tax) f(Income) f(Miles) f(L_prop)	0.6786867	0.6487971	0.5999074	5.000000	545.5892

NOTE : Predsq - The predictive R square is a measure that helps us to determine how well
the model predicts responses for new observations.
Predictive R-squared = [1 - (predicted RSS / sums of squares total) * 100 ,
the best model has the highest value of predsq.

The model with explanatory variables Tax , Income , L_prop has largest value of adjusted R^2 and the smallest AIC and CP values closer to the number of explanatory variables, than that of all the other models in the table.Therefore, we shall consider the model with explanatory variables Tax , Income , L_prop as our final model. If Y denote the values of the response variable Petrol, the considered model is given below:-

MODEL M :

$$Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 + \epsilon$$

Here X_1, X_2, X_3 denote the values corresponding to the scaled and centered values of the columns Tax , Income , L_prop respectively and $\gamma_1, \gamma_2, \gamma_3$ are corresponding parameters and γ_0 is the intercept . We assume that the random error ϵ has the following properties :

- it has 0 expectation
- variance covariance matrix of ϵ vector is $\sigma^2 I_n$
- It is normally distributed

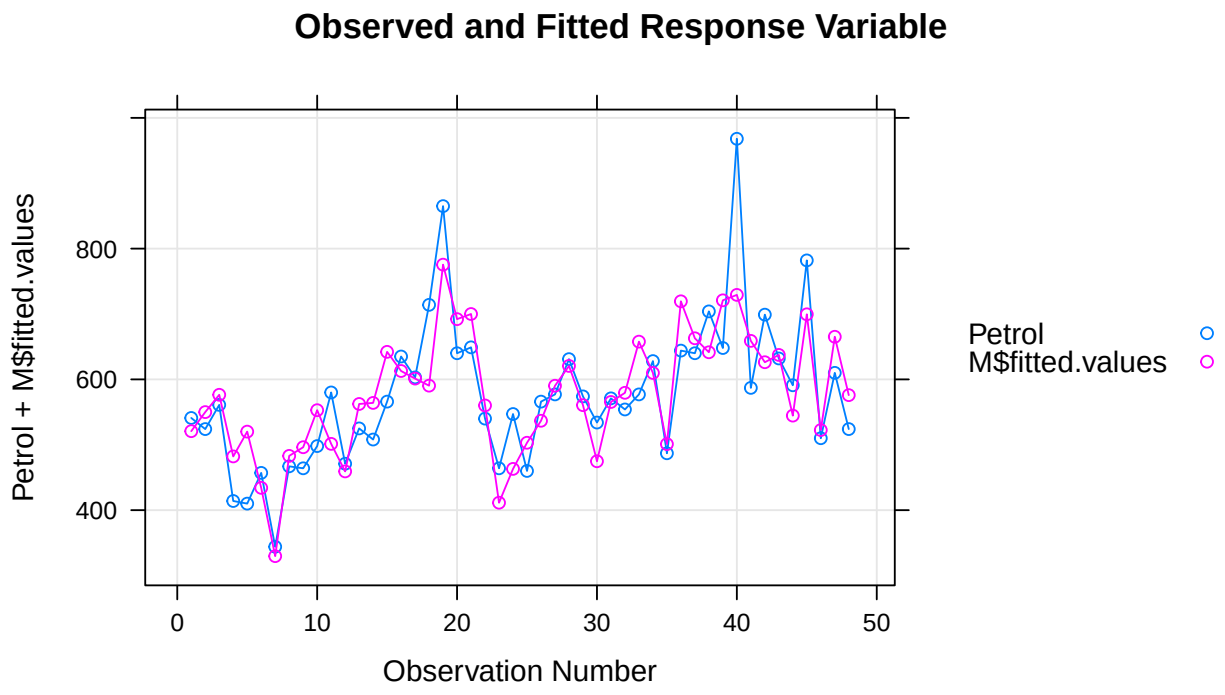
SUMMARY OF THE MODEL :

```
M<-lm(Petrol~f(Tax)+f(Income)+f(L_prop))
summary(M)
```

```
##
## Call:
## lm(formula = Petrol ~ f(Tax) + f(Income) + f(L_prop))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -110.10  -51.22  -12.89   24.49  238.77
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  576.771      9.517  60.602  < 2e-16 ***
## f(Tax)      -192.180     68.985  -2.786  0.00785 **
## f(Income)   -267.504     66.892  -3.999  0.00024 ***
## f(L_prop)    522.804     69.847   7.485 2.24e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65.94 on 44 degrees of freedom
## Multiple R-squared:  0.6749, Adjusted R-squared:  0.6527
## F-statistic: 30.44 on 3 and 44 DF,  p-value: 8.235e-11
```

In order to show how better the fit is consider the following plot :

```
xyplot(Petrol + M$fitted.values ~1:48, auto.key = list(space = "right"),grid = TRUE,
       main = "Observed and Fitted Response Variable",xlab = "Observation Number", type = "b" )
```

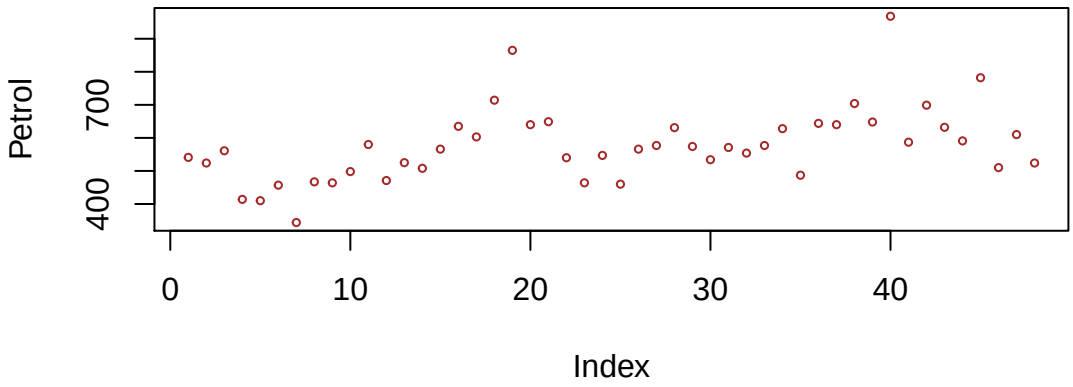


Clearly , the fit is not good .

PRELIMINARY ANALYSIS OF THE SELECTED MODEL :

Consider the index plot of Petrol Consumption :

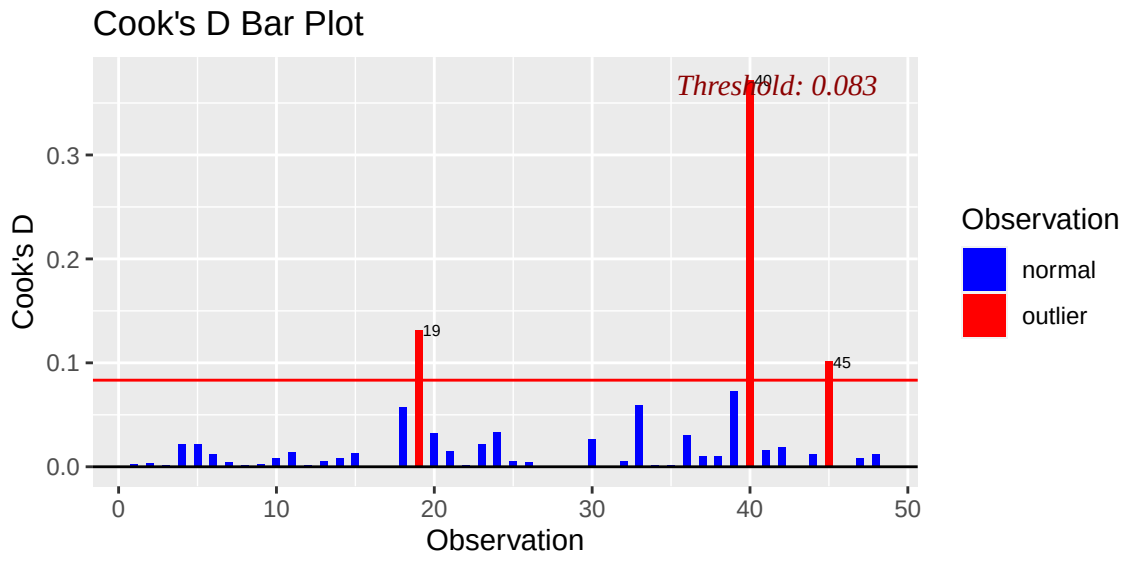
```
plot(Petrol,cex=0.5,col="brown")
```



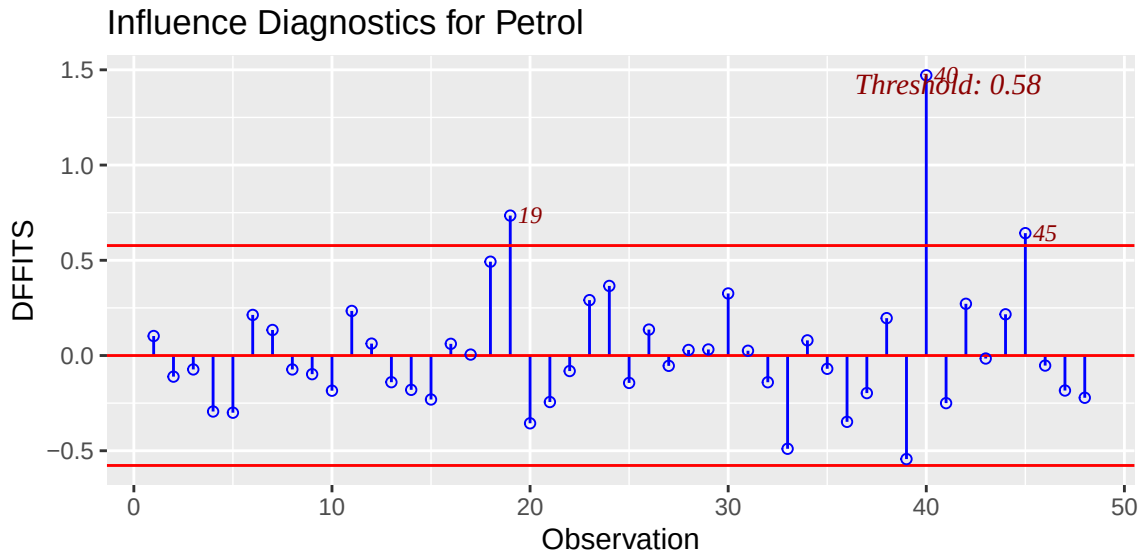
This plot easily says there are outliers in the data .

We will try to detect influential points in our model with the help of Cook's D Bar Plot,DFFIT plot,DFBETAS plot, Studentized Residual Plot and Hatmatrix diagonals Plot .

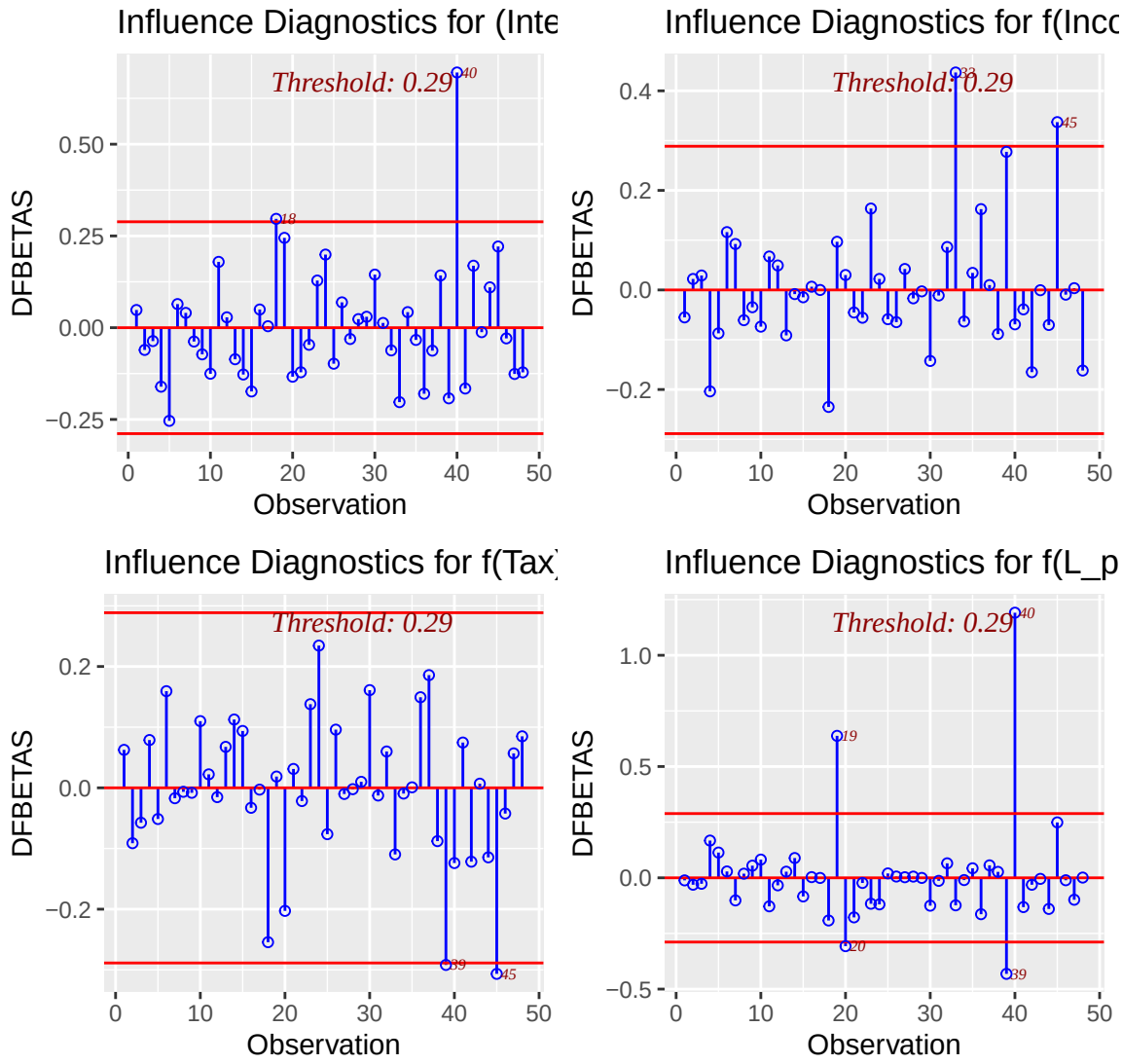
```
ols_plot_cooksd_bar(M)
```



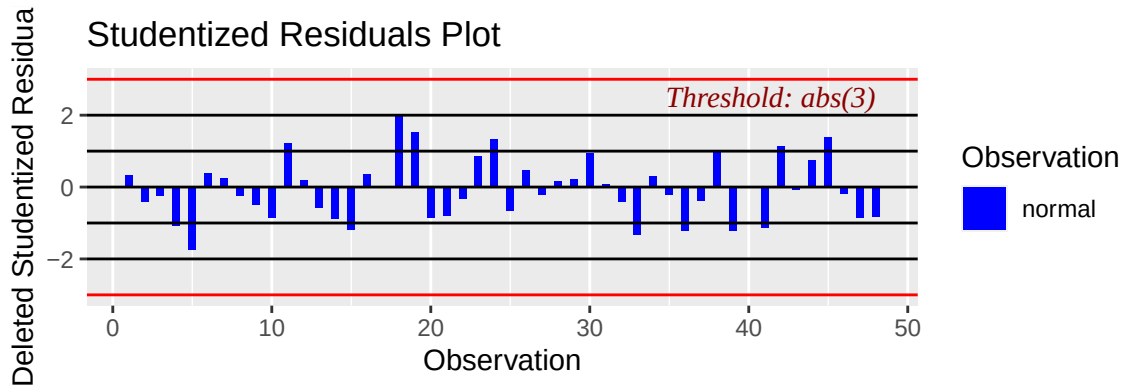

```
ols_plot_dffits(M)
```



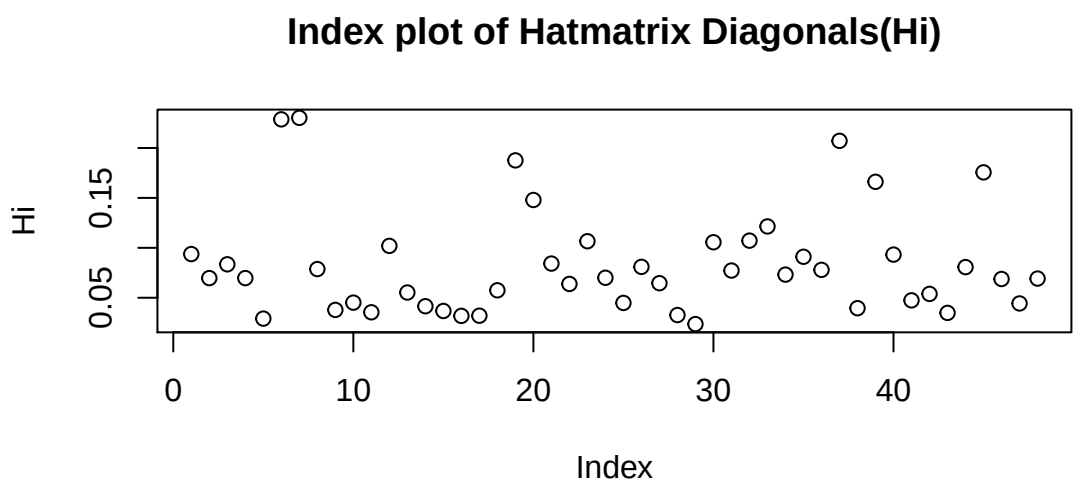

```
ols_plot_dfbetas(M)
```




```
ols_plot_resid_stud(M)
```




```
ans=ls.diag(M)
hi=ans$hat
plot(hi,ylab="Hi",xlab="Index",typ="p",main="Index plot of Hatmatrix Diagonals(Hi)")
```




```
which(hi>8/48)
```

```
## [1]  6  7 19 37 45
```

```
which(covratio(M)<0.75&&covratio(M)>1.25)
```

```
## integer(0)
```

We have seen that observations 19 ,40 and 45 appear as influential points in both Cook's D bar plot and DFFIT plot. The studentized residual plot and covratio do not suggest presence of any outlier.Lastly using Hat matrix diagonals we get that the points in position no. 6,7,19,37,45 are high leverage points .So we remove all those points for which we have more evidences to be outliers . Because removing less number of points will keep on giving the same trouble of outlying values and removing more of them will eventually abolish out important information from the given data resulting a bad fitting in both the cases. Hence we will remove the $19^{th}, 40^{th}, 45^{th}$ points and check if the fit improves anymore. We compare adjusted R^2 and $\hat{\sigma}^2$ values for these two models with (i.e M) and without (i.e M1) outliers respectively.

The adjusted R^2 values of the two models with and without outliers are respectively :

```
summary(M)$adj.r.squared
```

```
## [1] 0.6526896
```

```
summary(M1)$adj.r.squared
```

```
## [1] 0.6789824
```

The $\hat{\sigma}^2$ values of the two models with and without outliers are respectively :

```
summary(M)$sigma^2
```

```
## [1] 4347.783
```

```
summary(M1)$sigma^2
```

```
## [1] 2135.631
```

The $\hat{\sigma}^2$ and adjusted R^2 values have also improved.Therefore clearly the model without outliers is behaving better than that of the former one .

SUMMARY OF THE MODEL :

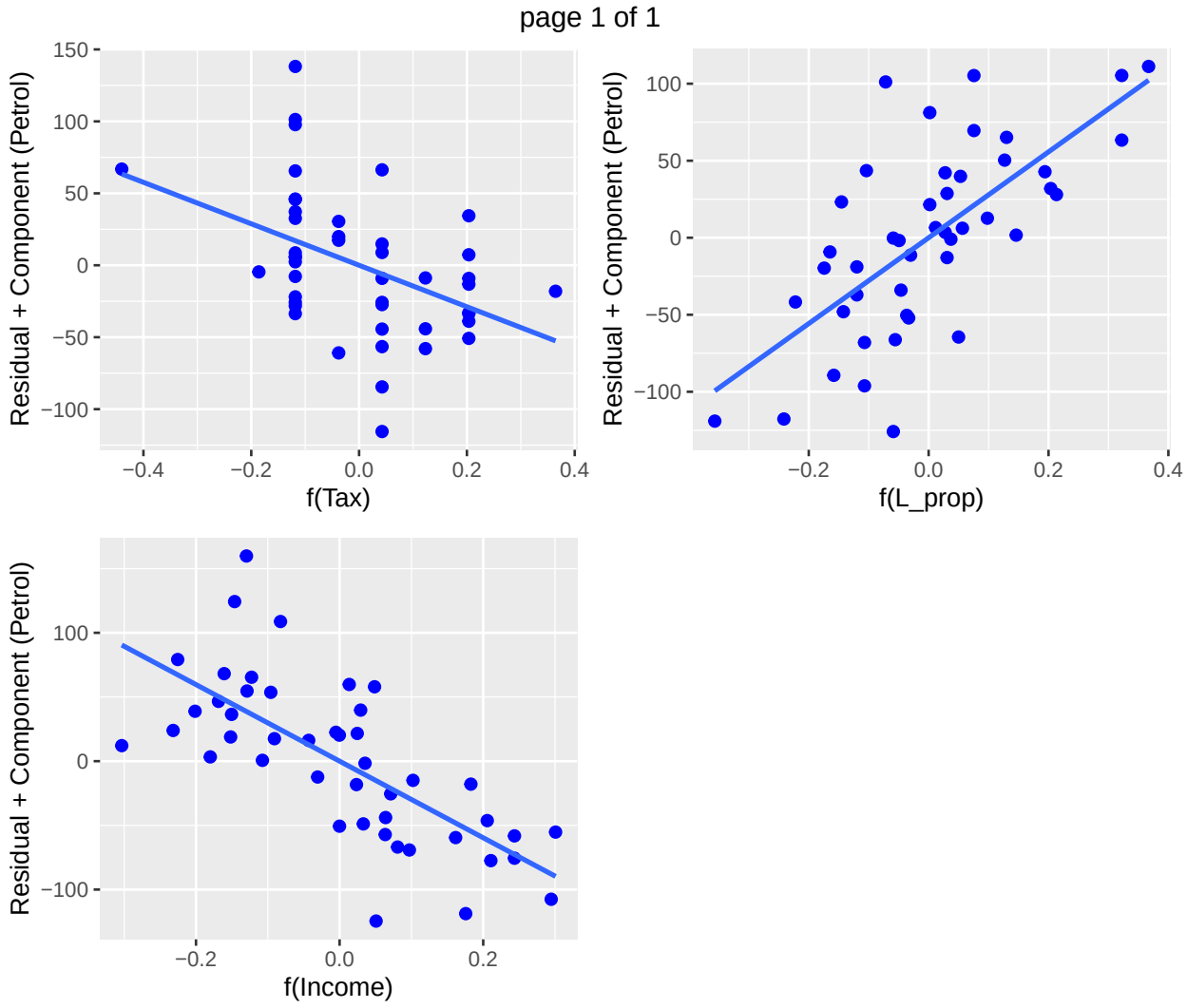
The summary of the outlier deleted model is given by :

```
summary(M1)
```

```
##
## Call:
## lm(formula = Petrol ~ f(Tax) + f(Income) + f(L_prop))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -109.392  -26.406   -3.717   20.966  121.123
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  557.111      6.889  80.870  < 2e-16 ***
## f(Tax)      -144.155     47.281  -3.049  0.00401 **
## f(Income)   -298.311     46.533  -6.411 1.13e-07 ***
## f(L_prop)    278.649     47.052   5.922 5.58e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 46.21 on 41 degrees of freedom
## Multiple R-squared:  0.7009, Adjusted R-squared:  0.679
## F-statistic: 32.02 on 3 and 41 DF,  p-value: 7.913e-11
```

```
ols_plot_comp_plus_resid(M1)
```

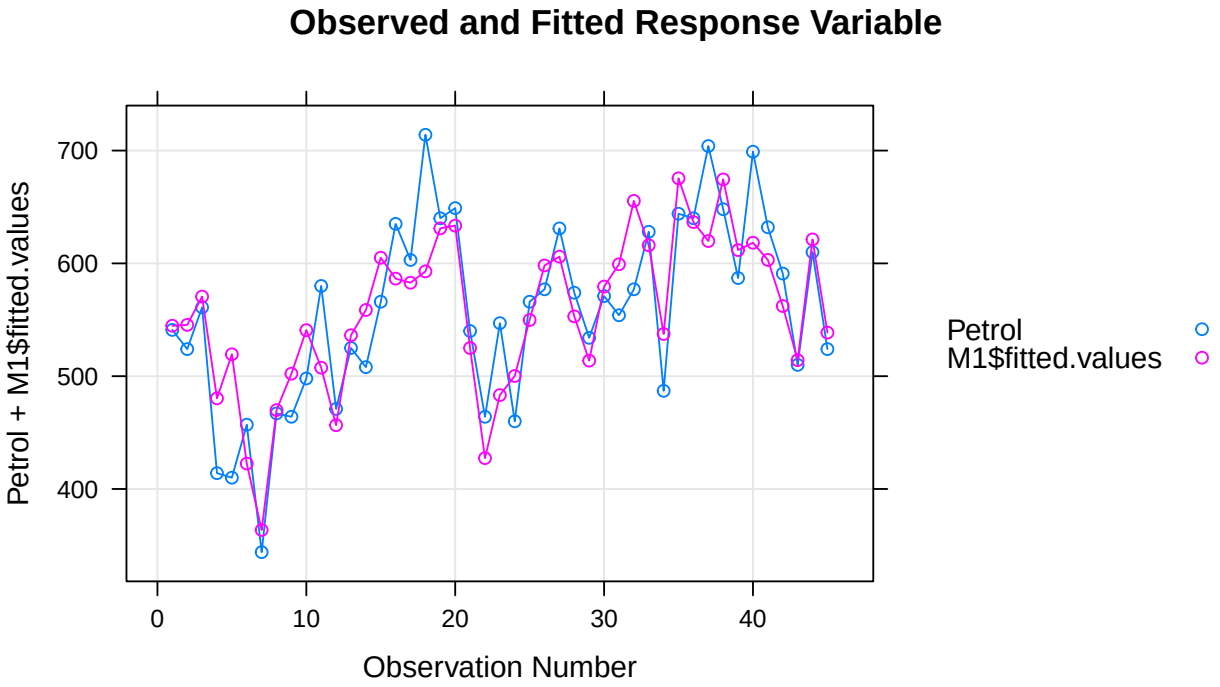
```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



If the response variable y is linearly depending upon the function (say g) of an independent variable x, then the partial residual plot corresponding to that covariate helps us to understand the shape of "g". Here the shape of 'g' corresponding to the covariate tax is not clear. the shape of 'g' corresponding to the covariates 'average income' and 'proportion' is almost linear.

In order to show how better the fit is consider the following plot :

```
xyplot(Petrol + M1$fitted.values ~1:45, auto.key = list(space = "right"),grid = TRUE,
       main = "Observed and Fitted Response Variable",xlab = "Observation Number", type = "b" )
```



Now we can see that the fit got better than that of the previous one .

Modification :

Here from the above partial residual plots we come to know that the variable Tax is behaving like a categorical variable , which is not true in general . But this can hold if there are some notion of rounding of in the given Tax values . Hence we use the method of adding noise to the Tax covariate to modify the overall view point for fitting linear regression .

```
set.seed(100)
Tax.n<-Tax+rnorm(length(Tax),0,0.2)
```

Now fitting the same linear regression using the modified Tax in the model we get :

```
Model<-(lm(Petrol~f(Tax.n)+f(Income)+f(L_prop)))
summary(Model)
```

```
##
## Call:
## lm(formula = Petrol ~ f(Tax.n) + f(Income) + f(L_prop))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -109.025  -25.794   -5.981   21.675  123.303
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  557.111      6.911  80.614  < 2e-16 ***
## f(Tax.n)    -141.681     47.283  -2.996  0.00462 **
## f(Income)   -298.307     46.689  -6.389 1.21e-07 ***
## f(L_prop)    281.750     47.041   5.989 4.48e-07 ***
```

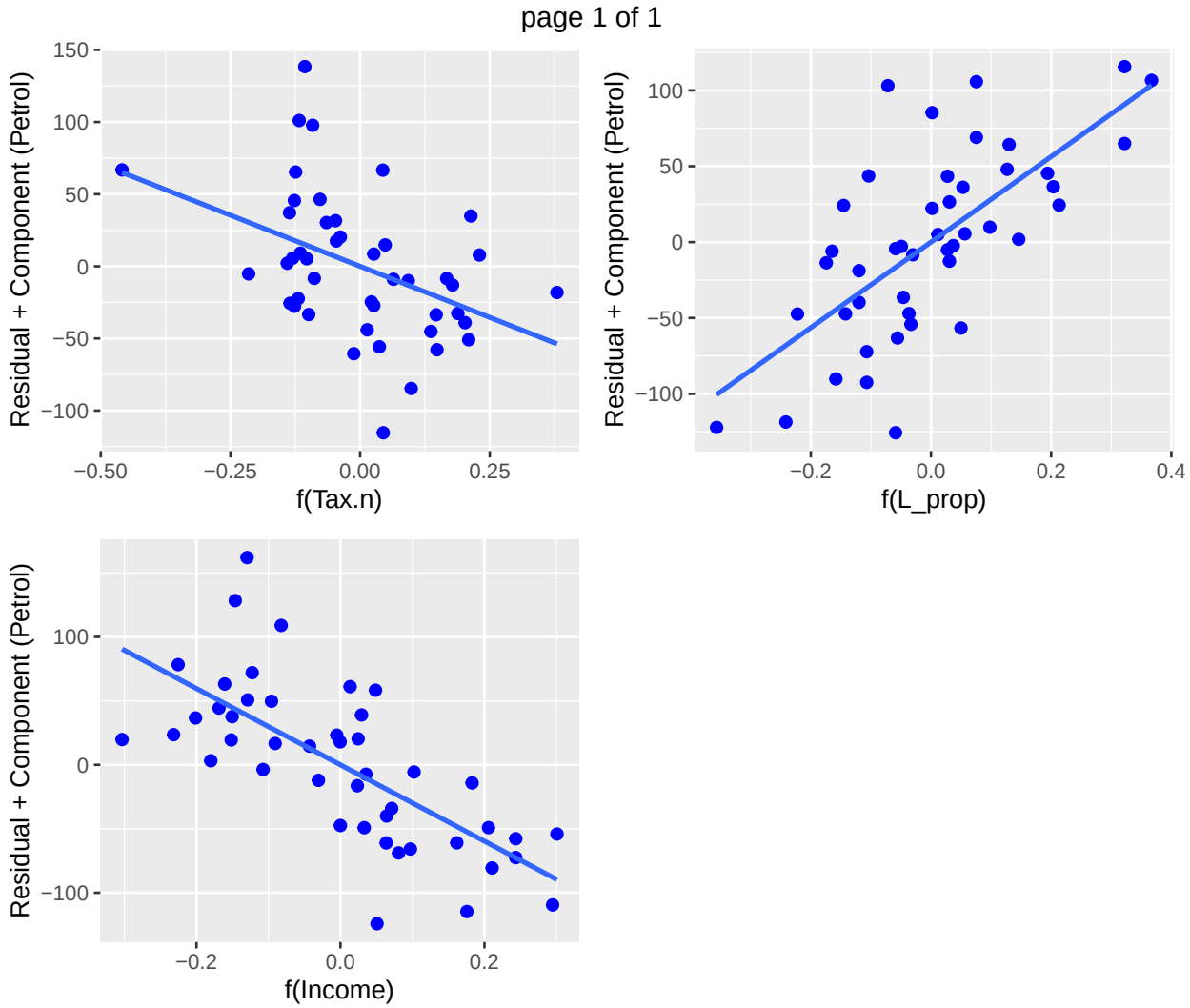
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 46.36 on 41 degrees of freedom
## Multiple R-squared:  0.699,  Adjusted R-squared:  0.6769
## F-statistic: 31.73 on 3 and 41 DF,  p-value: 8.997e-11
```

```
summary(Model)
```

```
##
## Call:
## lm(formula = Petrol ~ f(Tax.n) + f(Income) + f(L_prop))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -109.025  -25.794   -5.981   21.675  123.303
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  557.111      6.911  80.614  < 2e-16 ***
## f(Tax.n)    -141.681     47.283  -2.996  0.00462 **
## f(Income)   -298.307     46.689  -6.389 1.21e-07 ***
## f(L_prop)    281.750     47.041   5.989 4.48e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 46.36 on 41 degrees of freedom
## Multiple R-squared:  0.699,  Adjusted R-squared:  0.6769
## F-statistic: 31.73 on 3 and 41 DF,  p-value: 8.997e-11
```

```
ols_plot_comp_plus_resid(Model)
```

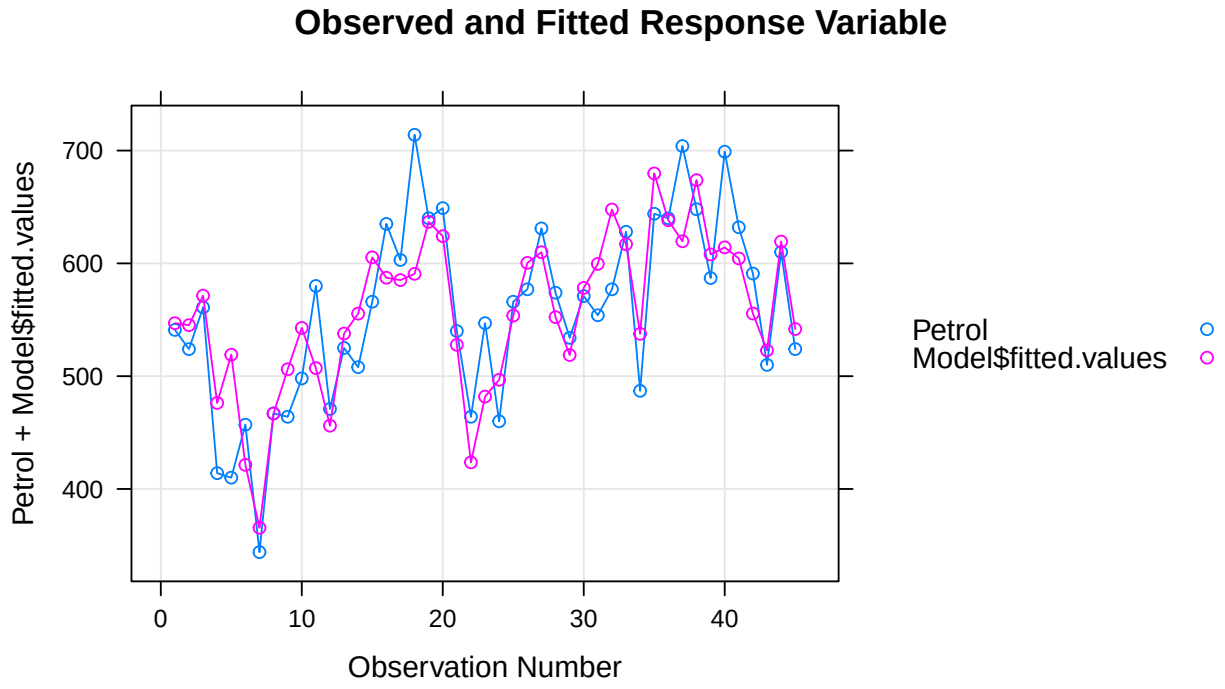
```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



Now the response variable (Petrol) is more or less linear with the variable Tax (actually Tax.n i.e noise added Tax).

In order to show how better the fit is consider the following plot :

```
xyplot(Petrol + Model$fitted.values ~1:45, auto.key = list(space = "right"),grid = TRUE,
       main = "Observed and Fitted Response Variable",xlab = "Observation Number", type = "b" )
```



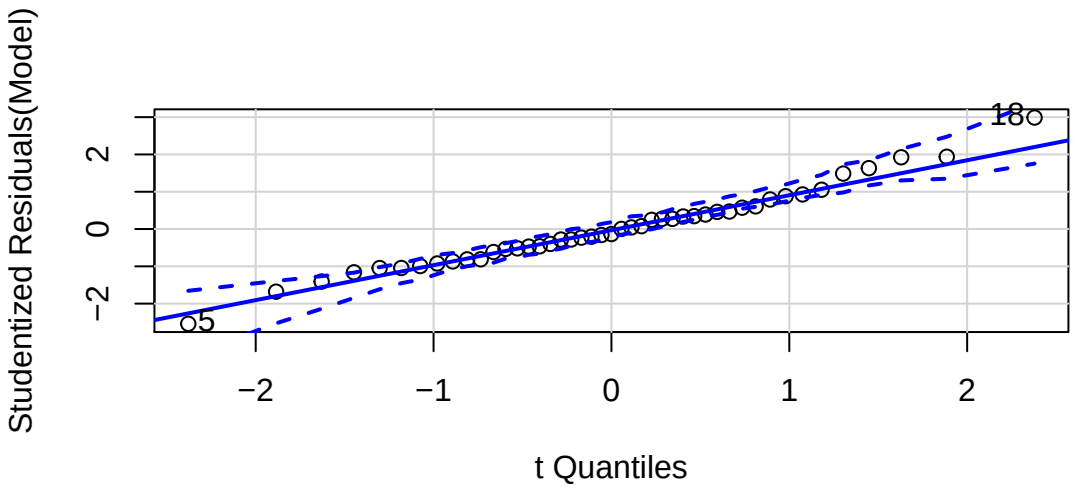
Next we will like to check the normality and homoscedasticity assumptions for our model.

CHECKING FOR ASSUMPTIONS :

NORMALITY:

The normal **qqplot** of the model after removal of the outliers is presented below :

```
qqPlot(Model)
```




```
## [1]  5 18
```

From the qqplot it can be observed that almost all the points are within the 95% confidence band, but still we will perform Shapiro-Wilk test for normality to have a better idea.

```
ols_test_normality(Model)$shapiro
```

```
##
##  Shapiro-Wilk normality test
##
## data:  y
## W = 0.98153, p-value = 0.6818
```

It shows that we do not have any strong evidence for rejecting the null hypothesis that the errors are normally distributed.

HOMOSCEDASTICITY:

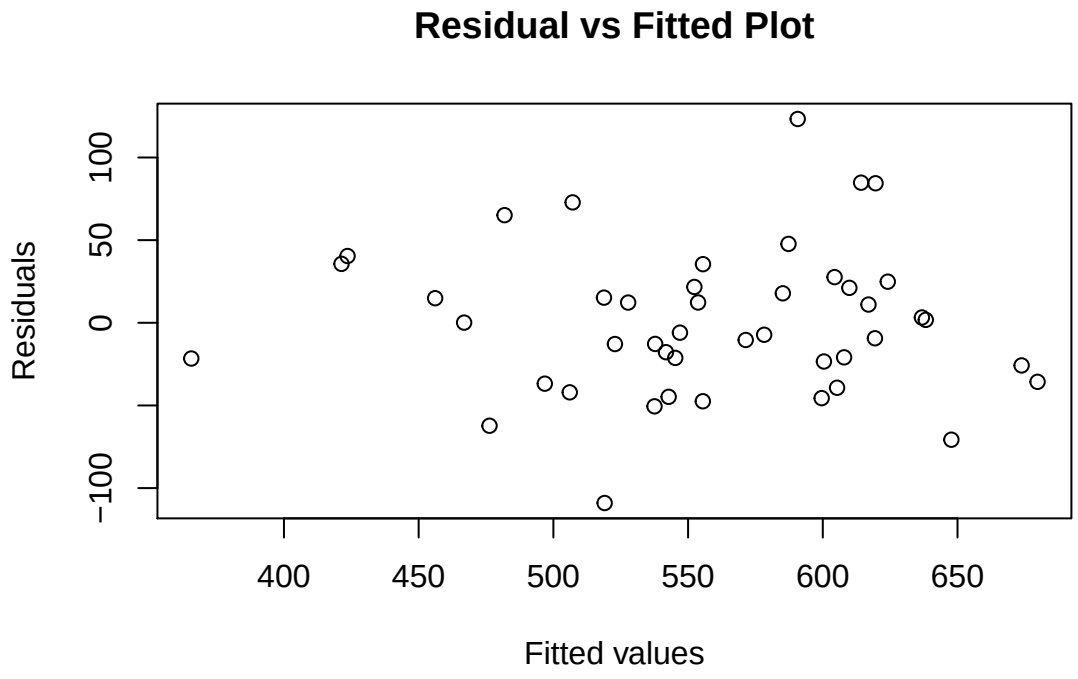
We will perform Breusch-Pagan test to check for homoscedasticity. The p-value for this test statistic is given below.

```
ols_test_breusch_pagan(Model)$p
```

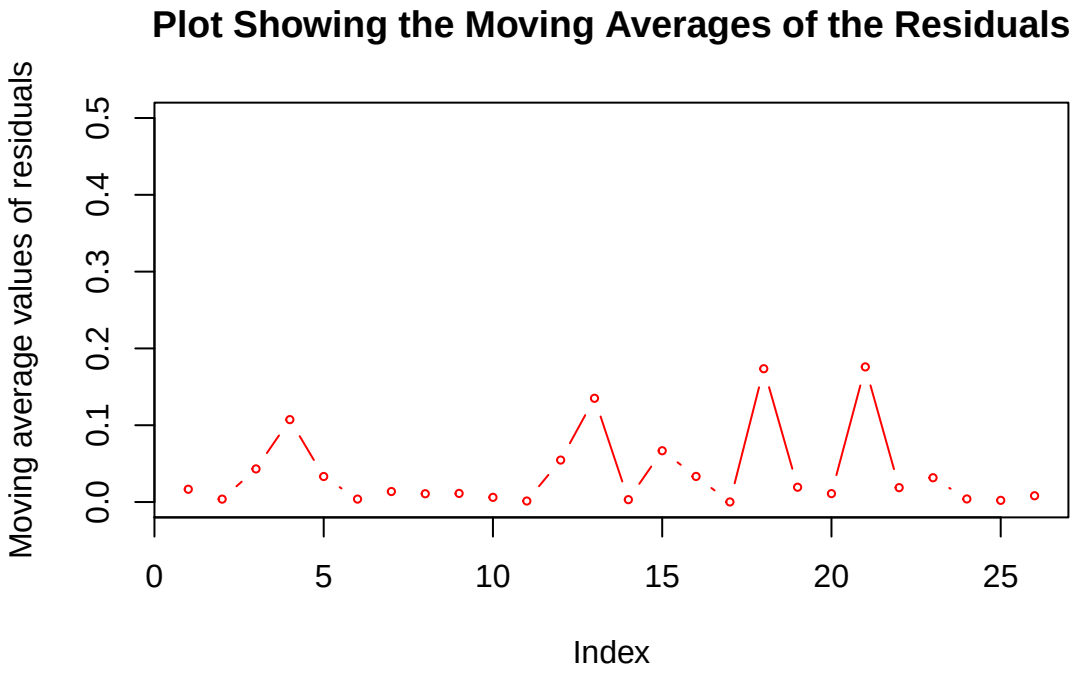
```
## [1] 0.8026934
```

The p-value is very large compared to 0.05(default level of significance) which indicates the absence of heteroscedasticity in our data. To confirm our hunch we present the Residual vs Fitted plot which will enable us to visualize presence or absence of heteroscedasticity more clearly.

```
plot(Model$fitted.values,Model$residuals, xlab = "Fitted values", ylab = "Residuals",
     main = "Residual vs Fitted Plot")
```



Now this plot shows more scatteredness towards right resulting a fan-like structure . Now to find more evidence in a single direction we use Moving Average method on the residuals of the fitted regression .

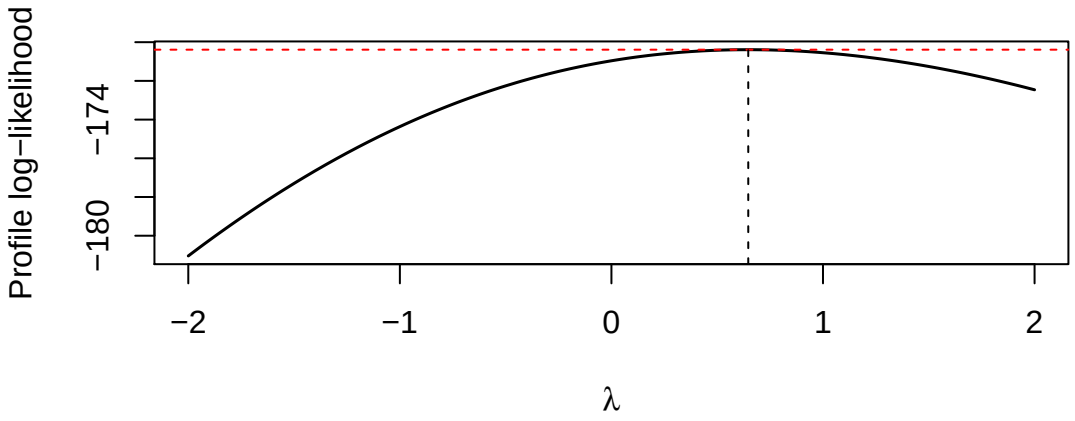


Now this is not at all a horizontal line . Hence we may conclude there are significant heterosacdasticity issues in the data . Hence we use the following transformations:

TRANFORMATIONS :

Boxcox Transformation :

```
boxcox(Model)$lambdahat
```




```
## [1] 0.6467182
```

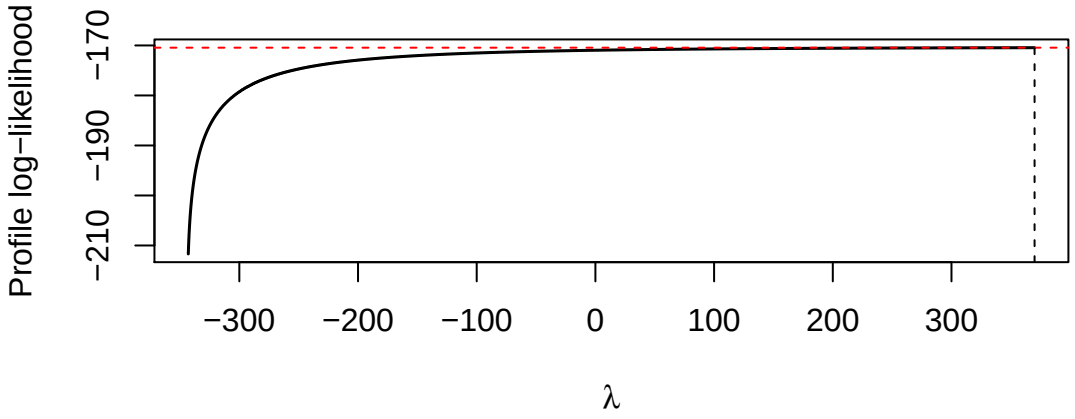
```
Petrol.tb=((Petrol^ 0.672239)-1)/ 0.672239
M.tb=lm(Petrol.tb~f(Tax.n)+f(Income)+f(L_prop))
summary(M.tb)
```

```
##
## Call:
## lm(formula = Petrol.tb ~ f(Tax.n) + f(Income) + f(L_prop))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.3116  -3.9977  -0.6264   2.9475  14.9284
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 102.5923     0.8702 117.900  < 2e-16 ***
## f(Tax.n)    -17.5477     5.9535  -2.947  0.00527 **
## f(Income)   -38.4391     5.8788  -6.539 7.42e-08 ***
## f(L_prop)    36.2274     5.9231   6.116 2.96e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.837 on 41 degrees of freedom
## Multiple R-squared:  0.7058, Adjusted R-squared:  0.6843
## F-statistic: 32.79 on 3 and 41 DF,  p-value: 5.638e-11
```

Log-shift Model:

```
logshiftopt(Model)$lambdahat
```

```
## The default lambdarange for the Log shift opt transformation is calculated dependent on the data ran
```




```
## [1] 369.9999
```

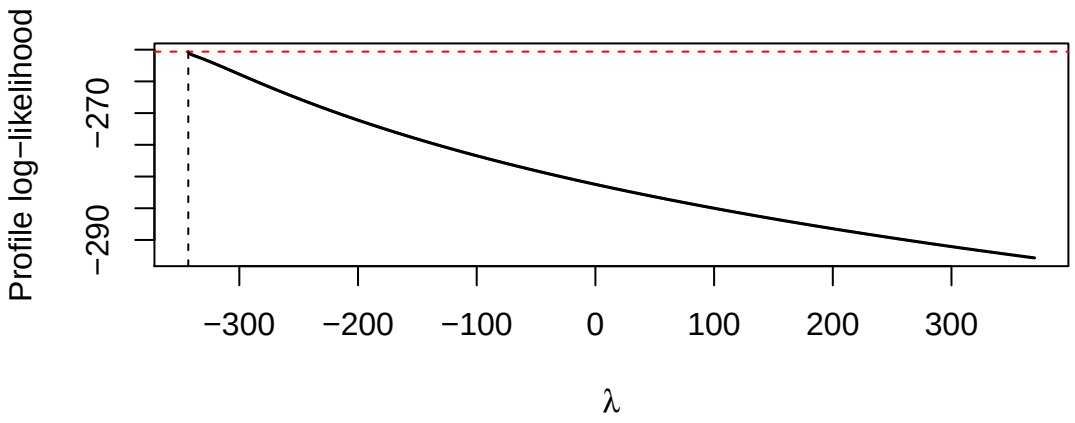
```
Petrol.tl=log(Petrol+369.9999)
M.tl=lm(Petrol.tl~f(Tax.n)+f(Income)+f(L_prop))
summary(M.tl)
```

```
##
## Call:
## lm(formula = Petrol.tl ~ f(Tax.n) + f(Income) + f(L_prop))
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.126653 -0.039240 -0.003908  0.024014  0.123558
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.828186   0.007468 914.367  < 2e-16 ***
## f(Tax.n)    -0.148289   0.051093  -2.902  0.00594 **
## f(Income)   -0.335243   0.050451  -6.645 5.24e-08 ***
## f(L_prop)    0.315378   0.050832   6.204 2.22e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05009 on 41 degrees of freedom
## Multiple R-squared:  0.7103, Adjusted R-squared:  0.6891
## F-statistic: 33.52 on 3 and 41 DF,  p-value: 4.118e-11
```

Square Shift Model :

```
sqrtshift(Model)$lambdahat
```

```
## The default lambdarange for the Square-root shift transformation is calculated dependent on the data
```




```
## [1] -342.9999
```

```
Petrol.ts=sqrt(Petrol-342.99)
M.ts=lm(Petrol.ts~f(Tax.n)+f(Income)+f(L_prop))
summary(M.ts)
```

```
##
## Call:
## lm(formula = Petrol.ts ~ f(Tax.n) + f(Income) + f(L_prop))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0856 -1.0909  0.1666  0.9597  3.7716
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.2561     0.2799  50.924  < 2e-16 ***
## f(Tax.n)     -4.3288     1.9154  -2.260   0.0292 *
## f(Income)   -12.6764     1.8913  -6.702 4.34e-08 ***
## f(L_prop)    12.1529     1.9056   6.378 1.26e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.878 on 41 degrees of freedom
## Multiple R-squared:  0.7048, Adjusted R-squared:  0.6832
## F-statistic: 32.63 on 3 and 41 DF,  p-value: 6.036e-11
```

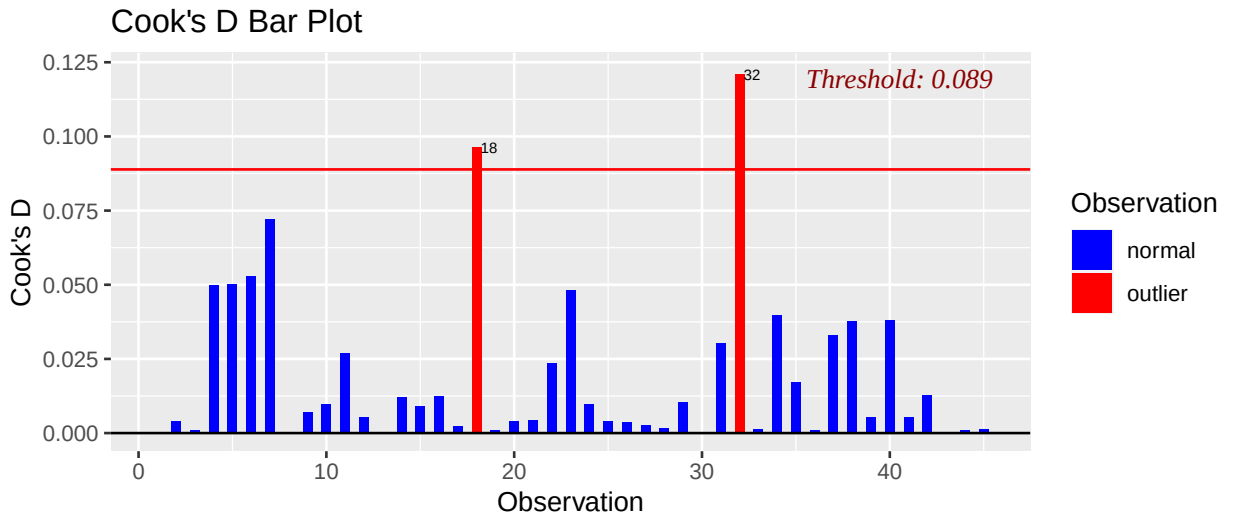
Adjusted R^2 is maximum in `Log-shift Model` . So consider further analysis on the model M.tl :

ANALYSIS OF THE TRANSFORMED MODEL :

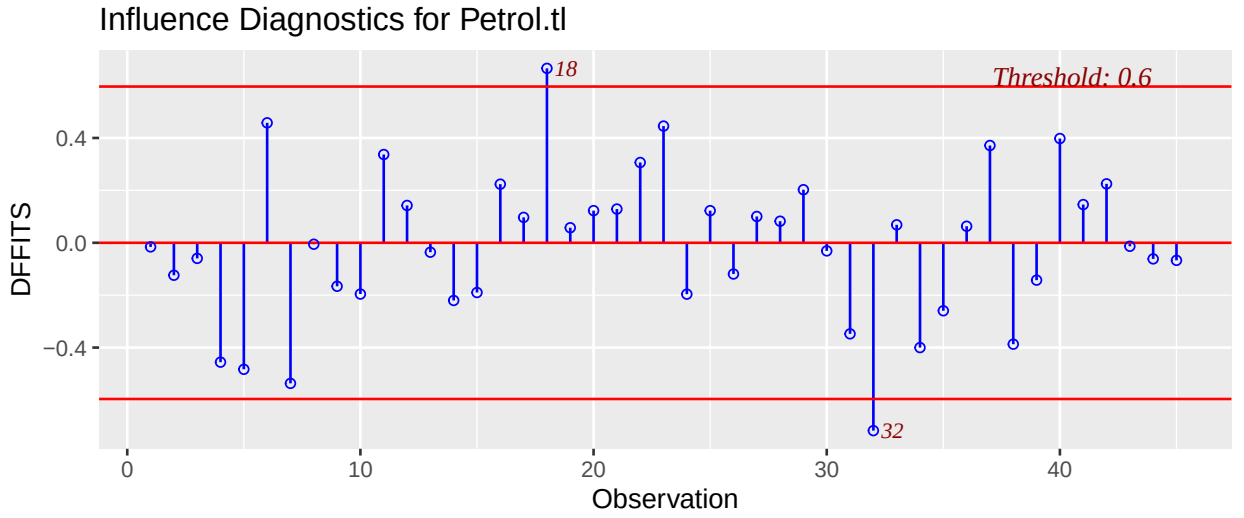
Outlier DETECTION :

We will try to detect influential points in our model with the help of Cook's D Bar Plot,DFFIT plot.

```
ols_plot_cooksd_bar(M.tl)
```




```
ols_plot_dffits(M.tl)
```



We have seen that observations 18 and 32 appear as influential points in both Cook's D bar plot and DFFIT plot.So we remove all those points for which we have evidences to be outliers . Because removing less number of points will keep on giving the same trouble of outlying values and removing more of them will eventually remove important information from the given data resulting a bad fitting in both the cases. Hence we will remove the $18^{th}, 32^{nd}$ points and check if the fit improves anymore. We compare adjusted R^2 and $\hat{\sigma}^2$ values for these two models with (i.e M.tl) and without(i.e Model.tl) outliers respectively.

The adjusted R^2 and $\hat{\sigma}^2$ values of the two models with and without outliers are respectively :

```
summary(M.tl)$adj.r.squared
```

```
## [1] 0.6905518
```

```
summary(Model.tl)$adj.r.squared
```

```
## [1] 0.7352163
```

The adjusted R^2 value gets enhanced after outlier deletion.

The $\hat{\sigma}^2$ values of the two models with and without outliers are respectively :

```
summary(M.tl)$sigma^2
```

```
## [1] 0.002498121
```

```
summary(Model.tl)$sigma^2
```

```
## [1] 0.002068479
```

Both the $\hat{\sigma}^2$ and adjusted R^2 values have also improved.Therefore clearly the model without outliers is behaving better than that of the former one .

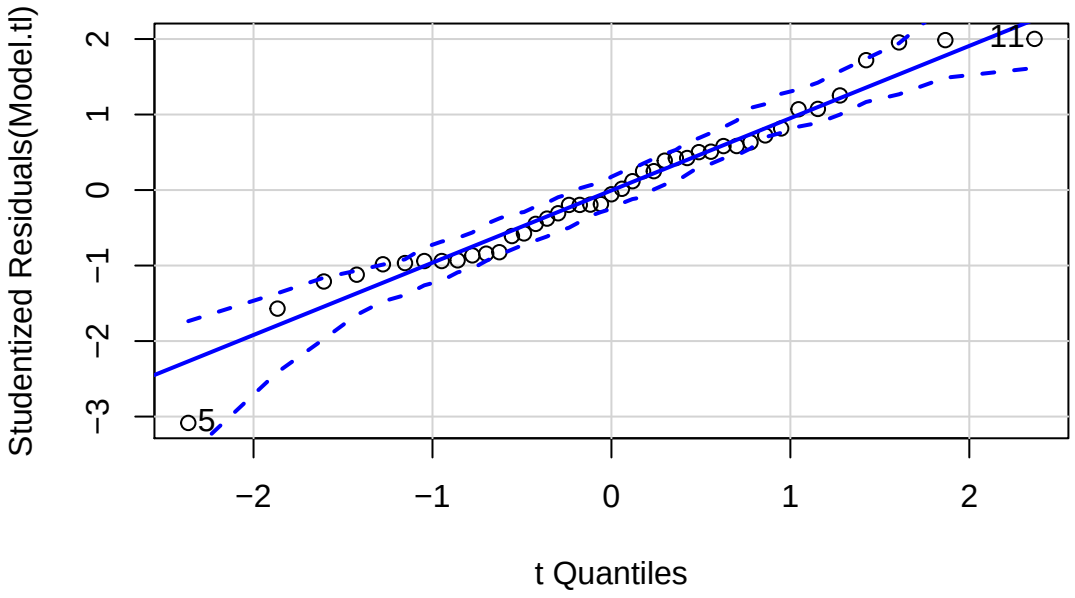
Next we will like to check the normality and homoscadasticity assumptions for our model.

CHECKING FOR ASSUMPTIONS :

NORMALITY:

The normal **qqplot** of the model after removal of the outliers is presented below :

```
qqPlot(Model.tl)
```




```
## [1]  5 11
```

From the qqplot it can be observed that almost all the points are within the 95% confidence band, but still we will perform Shapiro-Wilk test for normality to have a better idea.

```
ols_test_normality(Model.tl)$shapiro
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  y
## W = 0.97038, p-value = 0.3259
```

It shows that we do not have any strong evidence for rejecting the null hypothesis that the errors are normally distributed.

HOMOSCEDASTICITY:

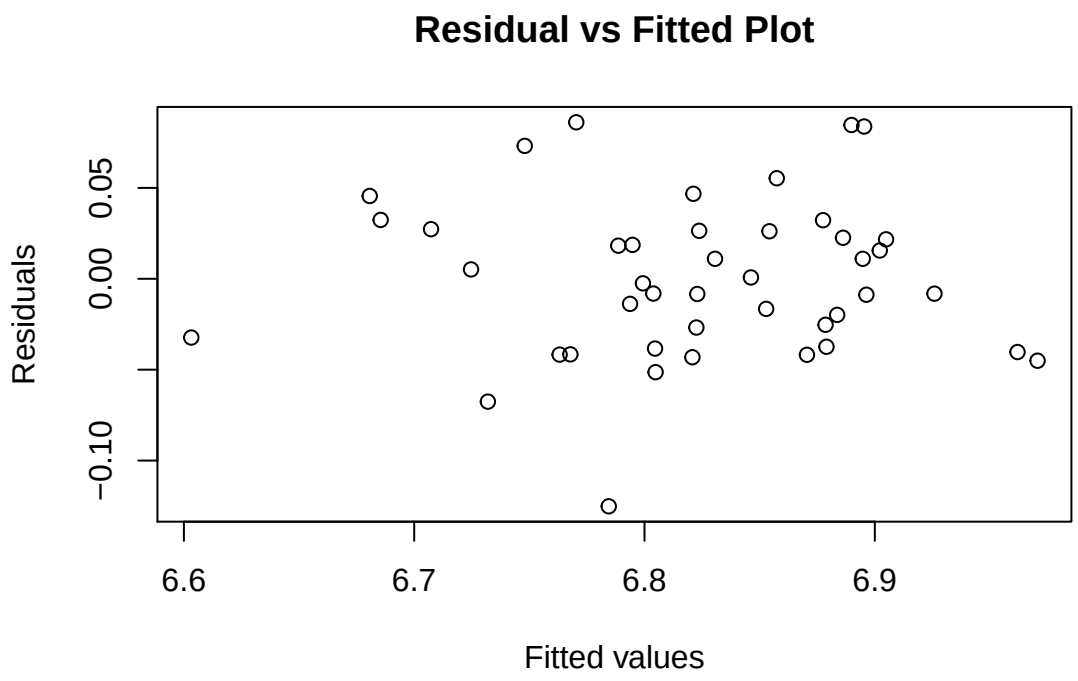
We will perform Breusch-Pagan test to check for homoscedasticity. The p-value for this test statistic is given below.

```
ols_test_breusch_pagan(Model.tl)$p
```

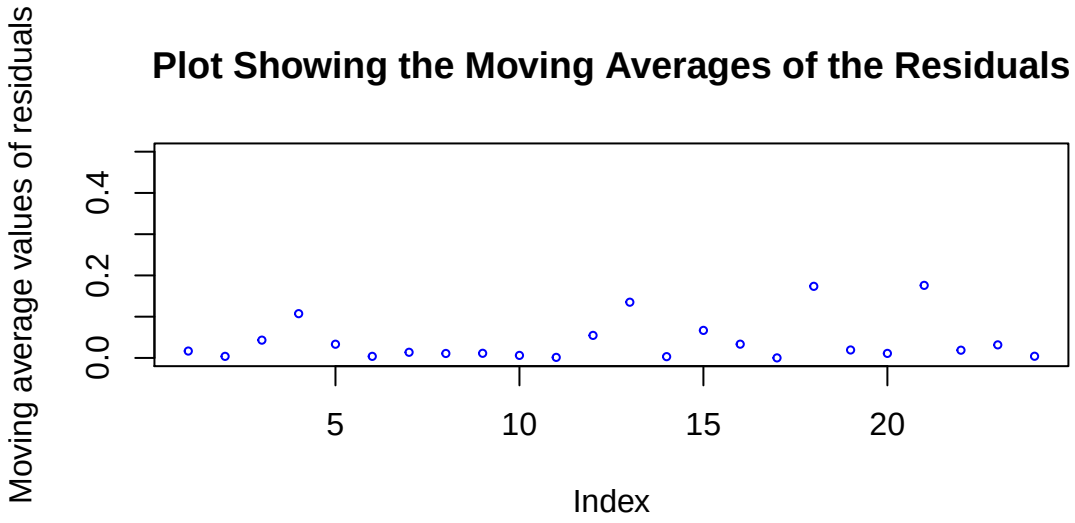
```
## [1] 0.4889326
```

The p-value is large compared to 0.05(default level of significance) which indicates the absence of heteroscedasticity in our data. To confirm our hunch we present the Residual vs Fitted plot which will enable us to visualize presence or absence of heteroscedasticity more clearly.

```
plot(Model.tl$fitted.values,Model.tl$residuals, xlab = "Fitted values", ylab = "Residuals",
     main = "Residual vs Fitted Plot")
```



Now this plot shows more scatteredness than the previous one . Now to compare the moving average with the previous one we plot Moving Averages of the residuals for the newly fitted regression (using 20 points at a time) .



Now the fluctuations among the points in moving average got decreased . Hence we may conclude there are no significant heterosacdasticity issues in the transformed data .

TESTING FOR COLLINEARITY :

In order to check for collinearity i.e. presence of any linear or almost linear relationship among the covariates, we use variance inflation factors and condition number(κ) respectively.

```
ols_coll_diag(Model.tl)
```

```
## Tolerance and Variance Inflation Factor
## ------------------------------------
##     Variables Tolerance      VIF
## 1   f(Tax.nl) 0.9494026 1.053294
## 2 f(Income.l) 0.9782329 1.022251
## 3 f(L_prop.l) 0.9643805 1.036935
##
##
## Eigenvalue and Condition Index
## ------------------------------
##   Eigenvalue Condition Index intercept   f(Tax.nl) f(Income.l) f(L_prop.l)
## 1   1.201306        1.000000         0 0.429557170   0.1063191    0.261629
## 2   1.040565        1.074465         0 0.003199209   0.6124782    0.319731
## 3   1.000000        1.096041         1 0.000000000   0.0000000    0.000000
## 4   0.758129        1.258796         0 0.567243620   0.2812027    0.418640
```

Both Condition Number and VIF(or Tolerance = 1/VIF) values are sufficiently lower(or larger) than their respective cut off points . So there is no indication of severe multi-collinearity in our data.

TESTING FOR AUTOCORRELATION :

To detect the presence of autocorrelation in the data consider the Darbin Watson Test :

```
dwtest(Model.tl)
```

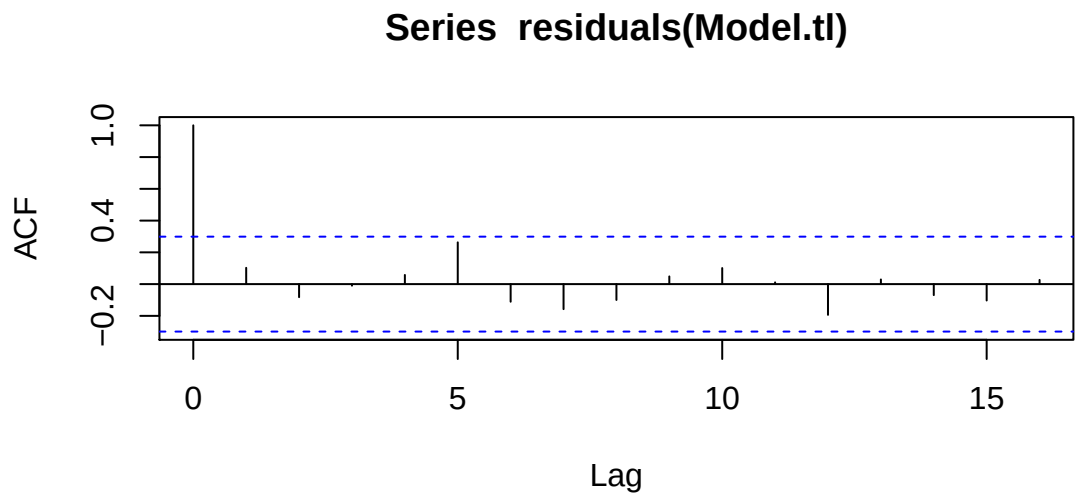
```
##
##  Durbin-Watson test
##
## data:  Model.tl
## DW = 1.7931, p-value = 0.1802
## alternative hypothesis: true autocorrelation is greater than 0
```

According to the high p-value we may conclude that no significant auto-correlation issues in the errors .

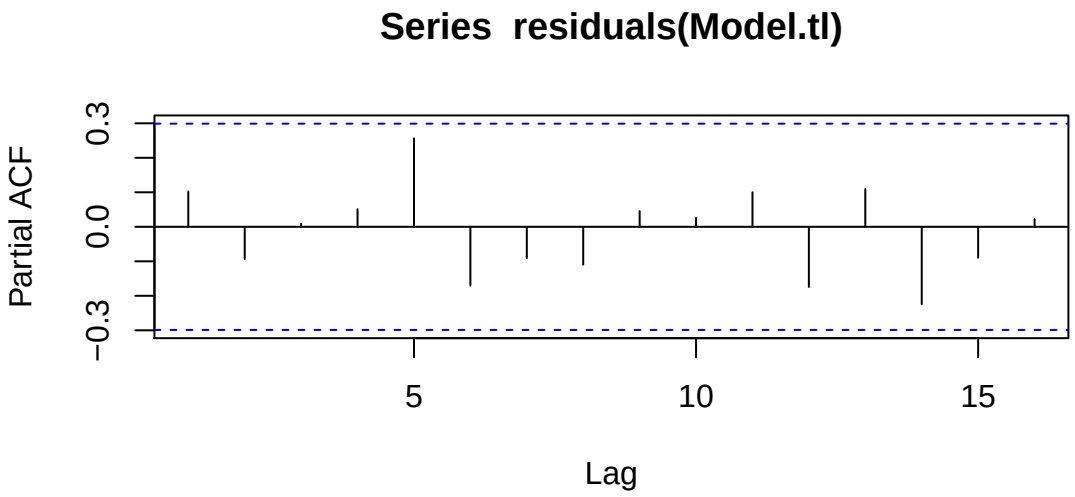
But we will check the ACF plot and PACF plot for conformation .

ACF PACF PLOTS :

```
acf(residuals(Model.tl))
```




```
pacf(residuals(Model.tl))
```



From the plots we get the same conclusion of absence of autocorrelation in the error terms .

SUMMARY OF THE MODEL :

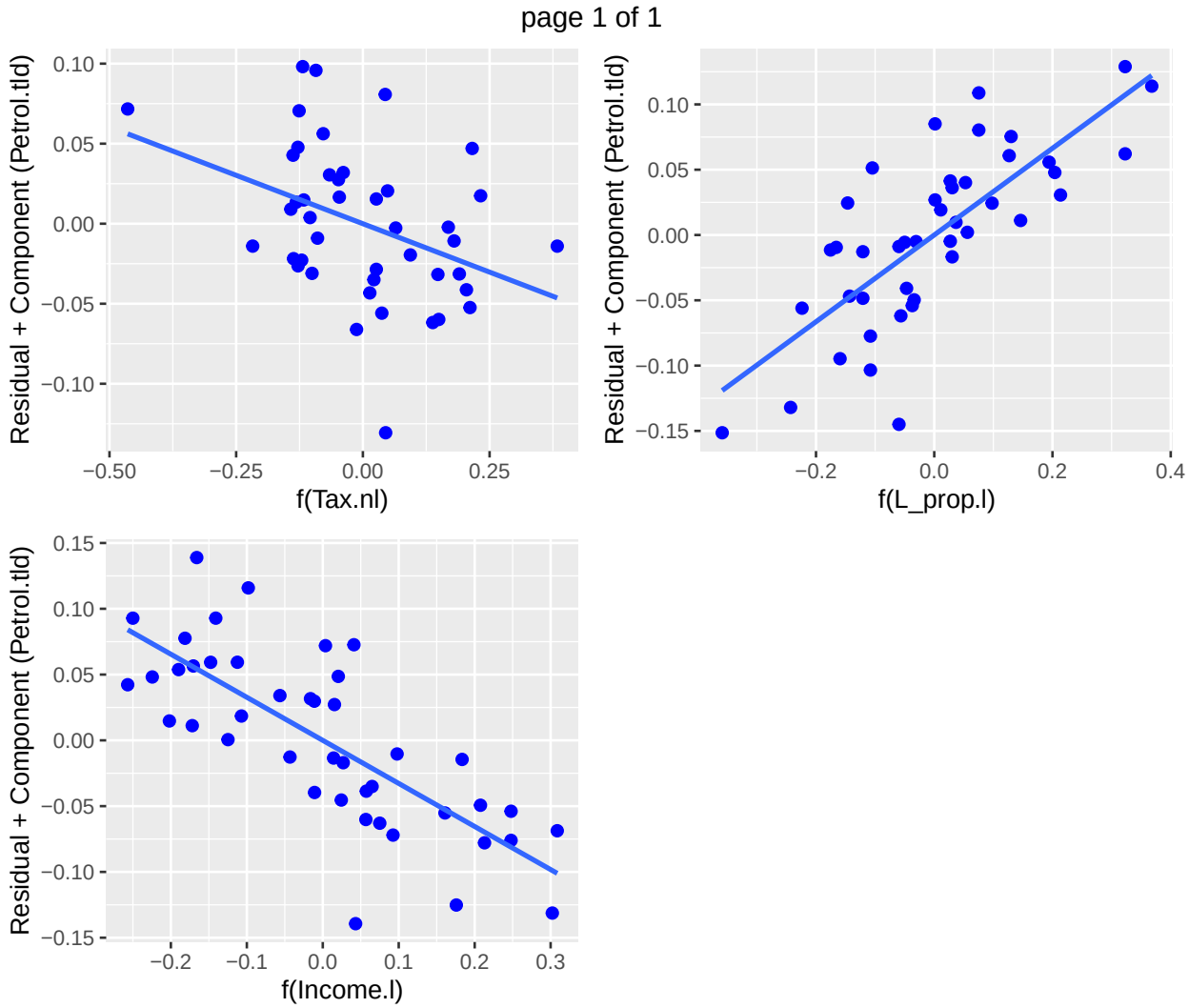
The summary of the finally selected model is given below:

```
summary(Model.tl)
```

```
##
## Call:
## lm(formula = Petrol.tld ~ f(Tax.nl) + f(Income.l) + f(L_prop.l))
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.125174 -0.034859 -0.002495  0.026263  0.086092
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.823876   0.006936 983.875  < 2e-16 ***
## f(Tax.nl)   -0.120968   0.046677  -2.592   0.0134 *
## f(Income.l) -0.327372   0.045984  -7.119 1.47e-08 ***
## f(L_prop.l)  0.332035   0.046313   7.169 1.26e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04548 on 39 degrees of freedom
## Multiple R-squared:  0.7541, Adjusted R-squared:  0.7352
## F-statistic: 39.87 on 3 and 39 DF,  p-value: 5.843e-12
```

```
ols_plot_comp_plus_resid(Model.tl)
```

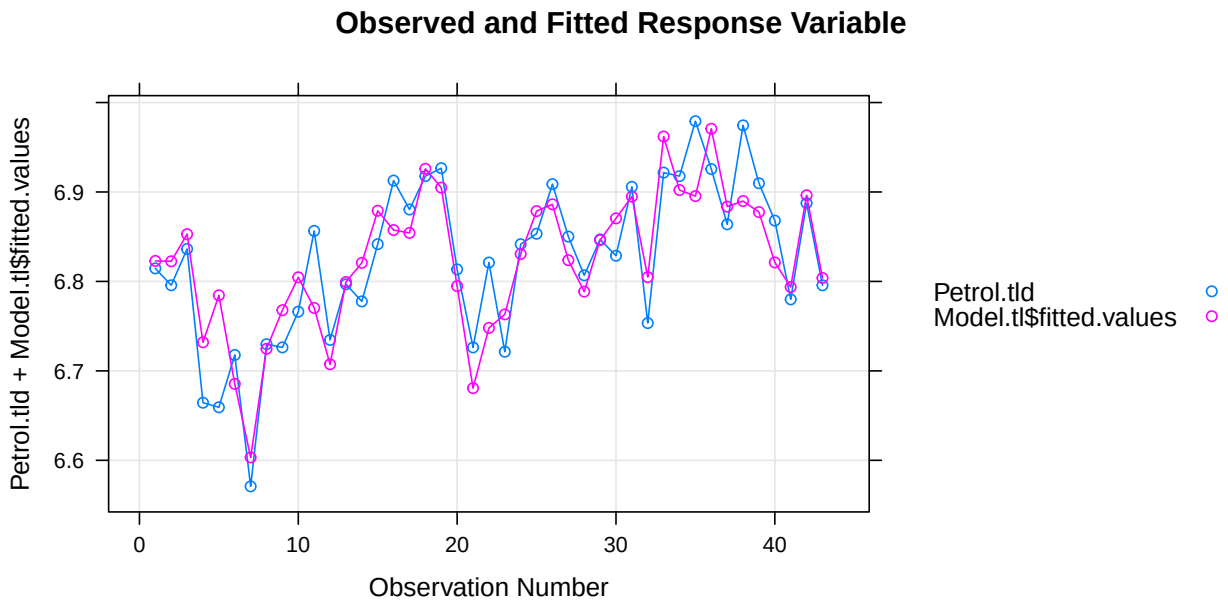
```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



The partial residual plots do not show any specific curvatures.

In order to show how better the fit is consider the following plot :

```
xyplot(Petrol.tld + Model.tl$fitted.values ~1:43, auto.key = list(space = "right"),grid = TRUE,
       main = "Observed and Fitted Response Variable",xlab = "Observation Number", type = "b" )
```



CONCLUSION :

From the above two plots we can say that our fit is good. The fiitted model can explain 74% of the total variability in the response. So, we can claim that petrol consumption depends on “tax on petrol”, “average income”, “proportion of population with driving licenses”.When tax increases, it discourages the consumption of petrol. when income increases petrol consumption decreases (It indicates the absence of other explanatory variables having negative impact on petrol consumption because income should have positive influence on consumption of petrol in real sense).Increment in the proportion of population with driving license encourages the consumption of petrol.