PROJECT REPORT On:

# PREDICTION OF RESALE VALUE OF USED CARS

A Regression Case Study Submitted By;

**Ramit Nandi (MD2211)**

Course Project of Regression Techniques

under supervision of Dr. Deepayan Sarkar

STATISTICS AND MATHEMATICS UNIT, ISI DELHI

# Acknowledgement

Real learning comes from a practical work. We would like to thank our instructor of the course Dr. Deepayan Sarkar (Statistics & Mathematics Unit, ISI Delhi), for providing us constant guidance and motivation for this project, without which it would have been an impossible task to accomplish.We would like to thank our department professors for teaching all the necessary topics with immense care which was needed to make the project fruitful.

We would also like to thank our seniors for their extensive support throughout the session. Their constant encouragement has enabled us to complete the project within the stipulated time period. We also take this opportunity to thank the authors and publishers of the various books and journals we have consulted .Without those this work would not have been completed. It has been a great learning experience and has also provided us with a practical insight of the theoretical knowledge gathered during the course lecture.

## Contents:- pg. no.

## Introduction :

In many developed countries, it is common to lease a car rather than buying it outright. A lease is a binding contract between a buyer and a seller (or a third party – usually a bank, insurance firm or other financial institutions) in which the buyer must pay fixed instalments for a pre-defined number of months/years to the seller/financer. After the lease period is over, the buyer has the possibility to buy the car at its residual value, i.e. its expected resale value. Thus, it is of commercial interest to seller/financers to be able to predict the salvage value (residual value) of cars with accuracy.

Predicting the price of used cars in both an important and interesting problem .According to data obtained from the National Transport Authority , the number of cars registered between 2003 and 2013 had witnessed a spectacular increase of 234%. From 68, 524 cars registered in 2003, this number reached 160, 701. It is reported that the sales of new cars had registered a decrease of 8% in 2013.With difficult economic conditions, it is likely that sales of second-hand imported (reconditioned) cars and used cars will increase. If the residual value is under-estimated by the seller/financer at the beginning, the instalments will be higher for the clients who will certainly then opt for another seller/financer. If the residual value is over-estimated, the instalments will be lower for the clients but then the seller/financer may have much difficulty at selling these high-priced used cars at this over-estimated residual value. Thus, we can see that estimating the price of used cars is of very high commercial importance as well.

## Objective :

It is trite knowledge that the value of used cars depends on a number of factors. The most important ones are usually the age of the car, its make (and model), the origin of the car (the original country of the manufacturer), its mileage and its horsepower. Due to rising fuel prices, fuel economy is also of prime importance. Other factors such as the interior style, the braking system, acceleration, the volume of its cylinders, safety index, its size, number of doors, paint colour, weight of the car, consumer reviews, prestigious awards won by the car manufacturer, its physical state, whether it is a sports car, whether it has cruise control, whether it is automatic or manual transmission, whether it belonged to an individual or a company and other options such as air conditioner, sound system, power steering, cosmic wheels, GPS navigator all may influence the price as well.

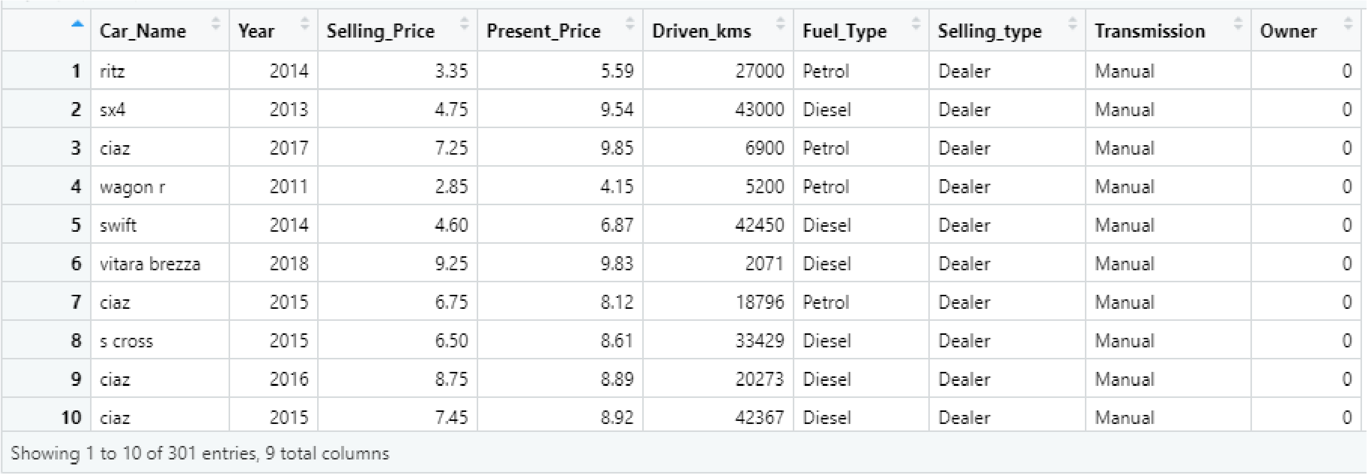
As we can see, the price depends on a large number of factors. Unfortunately, information about all these factors are not always available and the buyer must make the decision to purchase at a certain price based on few factors only. In this work, we have considered only a small subset of the factors mentioned above and tried to give reasonable prediction for the resale price.

## Data Source :

Here our data is collected from https://www.kaggle.com/datasets/vijayaadithyanvg/car-price-predictionused-cars

## Dataset Description:

Our dataset contains 301entries of 9variables….

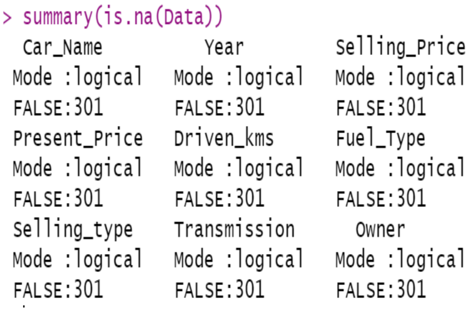


| | Car_Name | Year | Selling_Price | Present_Price | Driven_kms | Fuel_Type | Selling_type | Transmission | Owner |
|---|---|---|---|---|---|---|---|---|---|
| 1 | ritz | 2014 | 3.35 | 5.59 | 27000 | Petrol | Dealer | Manual | 0 |
| 2 | sx4 | 2013 | 4.75 | 9.54 | 43000 | Diesel | Dealer | Manual | 0 |
| 3 | ciaz | 2017 | 7.25 | 9.85 | 6900 | Petrol | Dealer | Manual | 0 |
| 4 | wagon r | 2011 | 2.85 | 4.15 | 5200 | Petrol | Dealer | Manual | 0 |
| 5 | swift | 2014 | 4.60 | 6.87 | 42450 | Diesel | Dealer | Manual | 0 |
| 6 | vitara brezza | 2018 | 9.25 | 9.83 | 2071 | Diesel | Dealer | Manual | 0 |
| 7 | ciaz | 2015 | 6.75 | 8.12 | 18796 | Petrol | Dealer | Manual | 0 |
| 8 | s cross | 2015 | 6.50 | 8.61 | 33429 | Diesel | Dealer | Manual | 0 |
| 9 | ciaz | 2016 | 8.75 | 8.89 | 20273 | Diesel | Dealer | Manual | 0 |
| 10 | ciaz | 2015 | 7.45 | 8.92 | 42367 | Diesel | Dealer | Manual | 0 |

Showing 1 to 10 of 301 entries, 9 total columns

The variables are as follow-

- **Car_Name** : model name of the car
- **Year** : year of purchase
- **Selling_Price** : resale value of the car
- **Present_Price** : price of a new car of same model
- **Driven_kms** : how many kms the car already ran
- **Fuel_type** : type of fuel used, i.e. petrol, diesel or CNG
- **Selling_type** : the used car is sold through a dealer or not
- **Transmission** : the car has automatic transmission or not
- **Owner** : number of owner of the car



```
> summary(is.na(Data))
  Car_Name           Year          Selling_Price
 Mode :logical   Mode :logical   Mode :logical
 FALSE:301       FALSE:301       FALSE:301
 Present_Price   Driven_kms      Fuel_Type
 Mode :logical   Mode :logical   Mode :logical
 FALSE:301       FALSE:301       FALSE:301
 Selling_type    Transmission      Owner
 Mode :logical   Mode :logical   Mode :logical
 FALSE:301       FALSE:301       FALSE:301
```

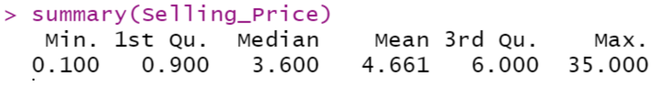
and clearly the dataset contains no missing value .

Here Selling_price is the variable of interest to be predicted & others are potential predictors.

## EDA:

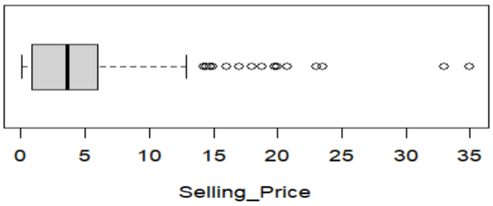
We start with examining our variables-

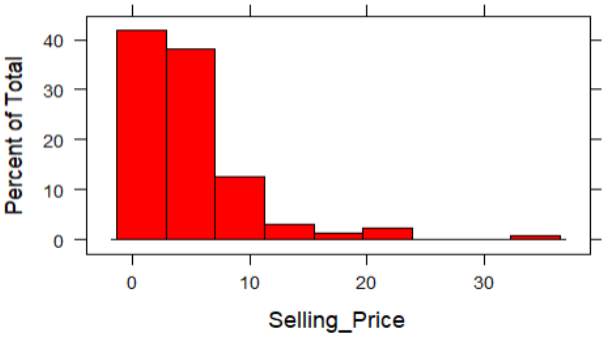
- **Selling_Price** :



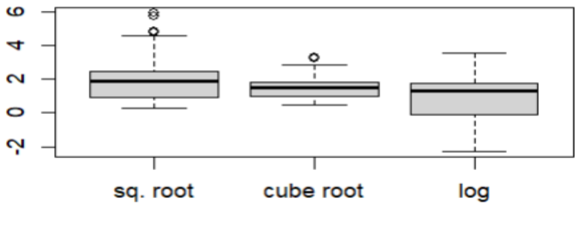
```
> summary(Selling_Price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.100   0.900   3.600   4.661   6.000  35.000 
```

this variable is positively skewed , has extreme observations at right end.



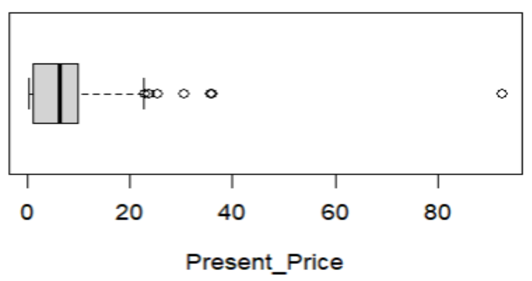


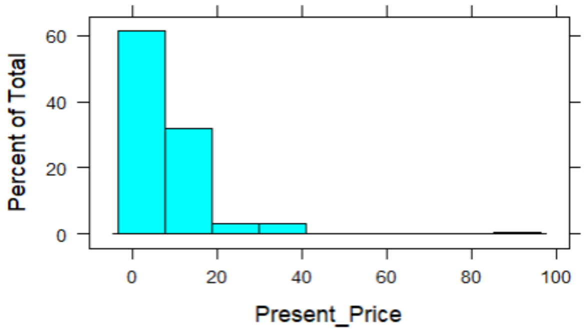
the log-transformation of the variable removes outliers, but to some extent it over-corrects the positive skewness.



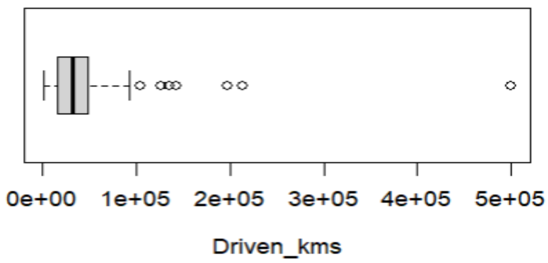
boxplot for various transformation

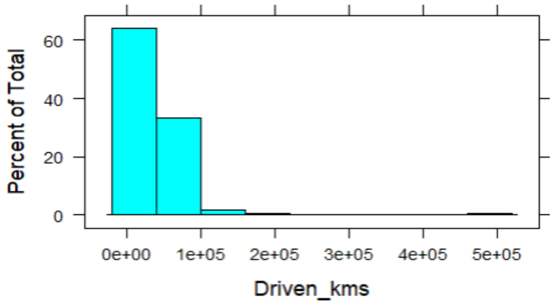
- **Present_Price** : It is also positively skewed like Selling_Price.



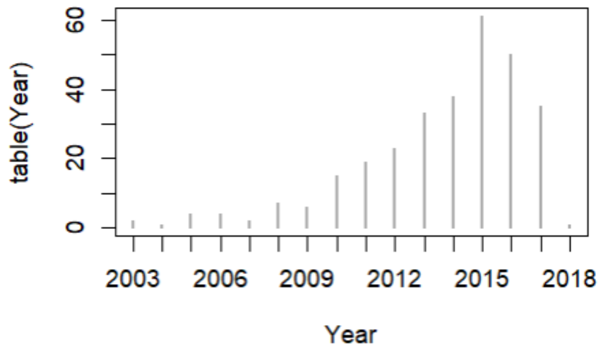


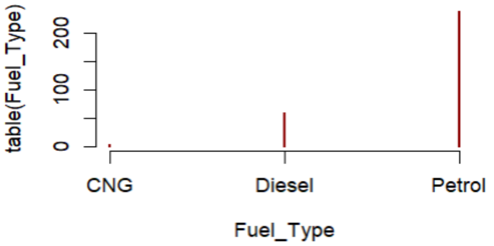
- **Driven_kms** : It is positive skewed, with extreme observations at right side.

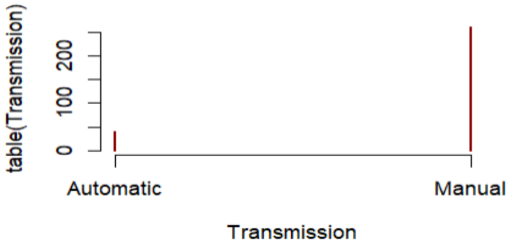


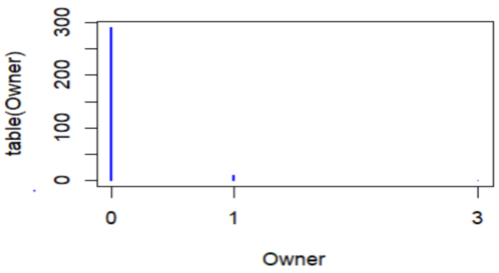


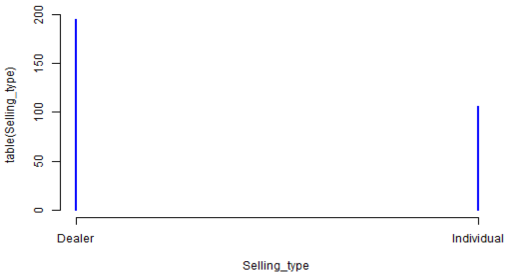
- **Year** : Note that Year is actually a categorical variable, but it gives information about how old the car is, which is a numerical variable. We notice, there is very few observations available for distant past or most recent years.





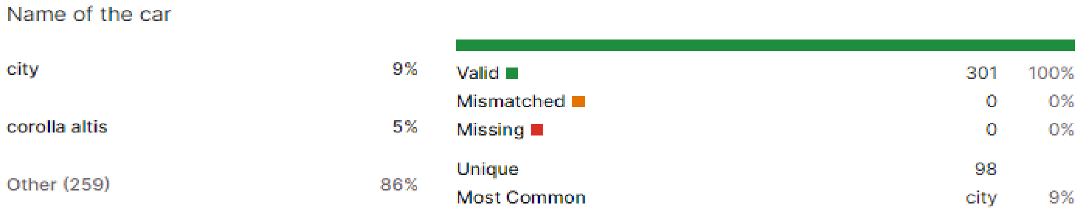






- **Fuel_Type** : Very few observation available for CNG compared to petrol and diesel.

- **Transmission** : Most car is of manual transmission type.
- **Selling_Type** : More cars are sold through dealer.
- **Owner** : Most observation is from one category, so we exclude this from further discussion.
- **Car_Name** : There are too many categories (98 categories). We will exclude this variable from further discussion.



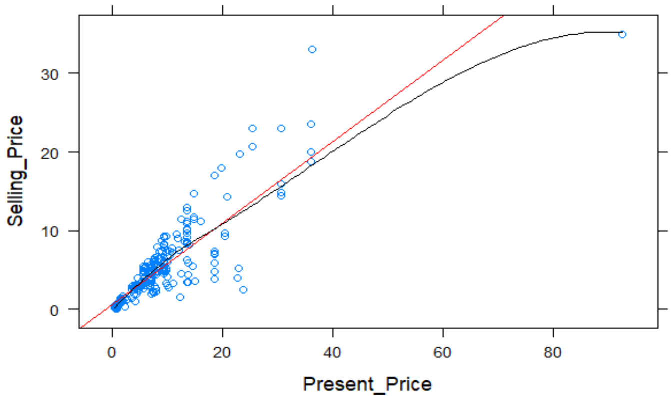
Name of the car

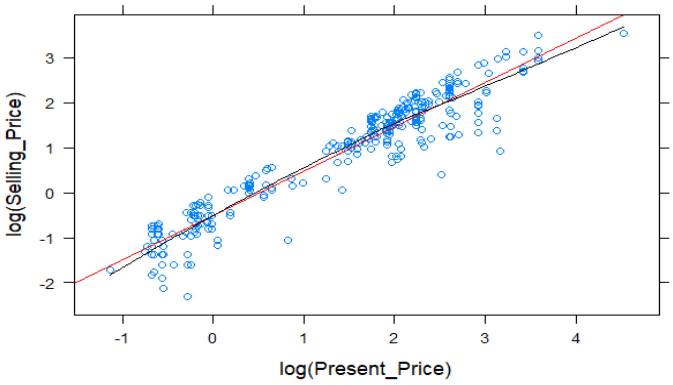
| | |
|---|---|
| city | 9% |
| corolla altis | 5% |
| Other (259) | 86% |

| | | |
|---|---|---|
| Valid 🟩 | 301 | 100% |
| Mismatched 🟧 | 0 | 0% |
| Missing 🟥 | 0 | 0% |
| Unique | 98 | |
| Most Common | city | 9% |

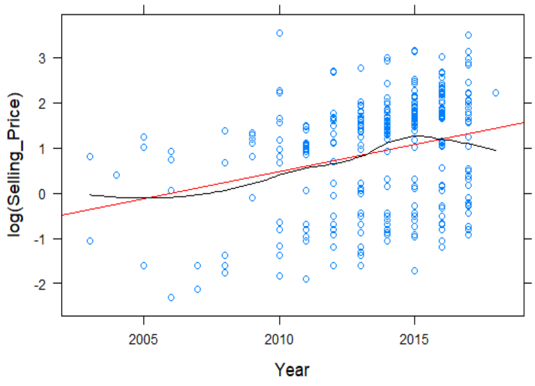
Now check for relationship between different variables-

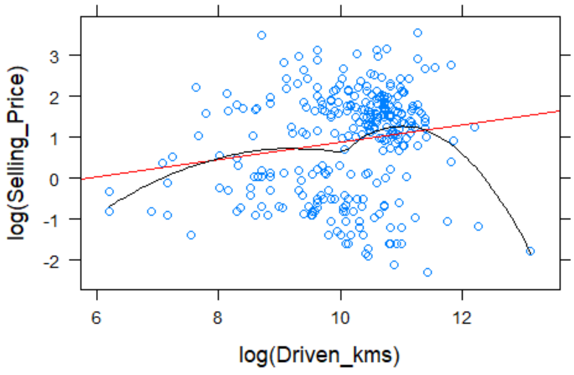
- **Selling_Price vs Present_Price** : We can see there are linear relationship between these two variables, but due to skewness most points in the plot are overlapped in a small region. If we take log-transformation of the variables to remove +ve skewness , observe the linear relationship is now more clear.



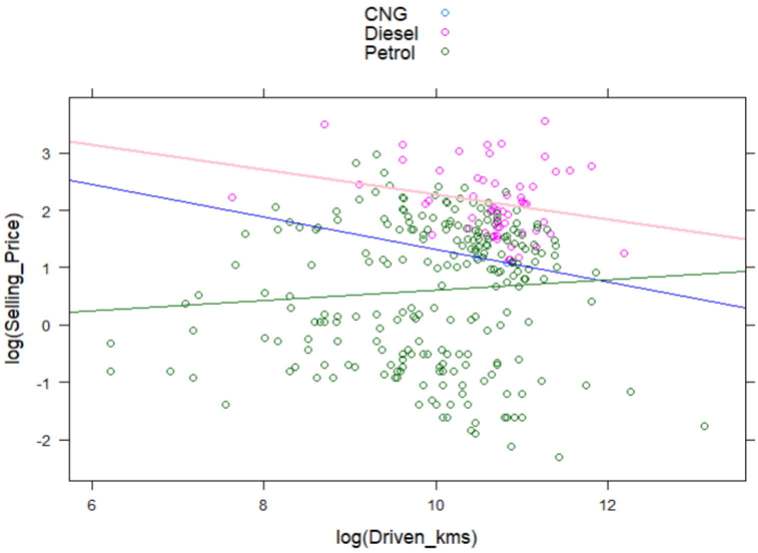


- **Selling_Price vs Year** : It is reasonable that older cars will get lower resell price.



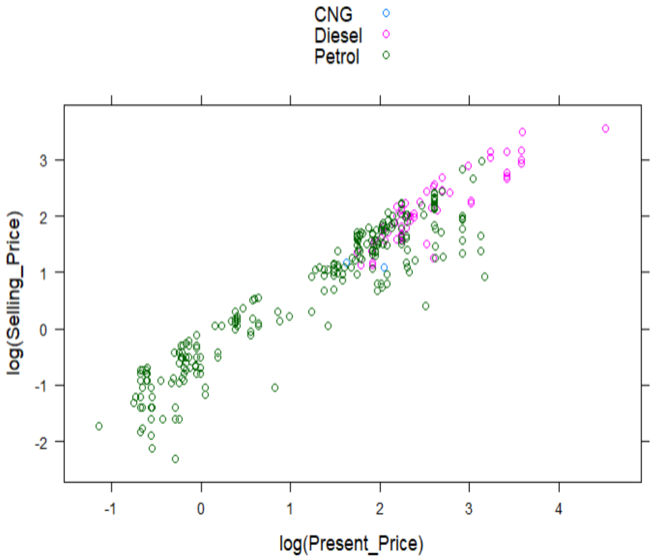


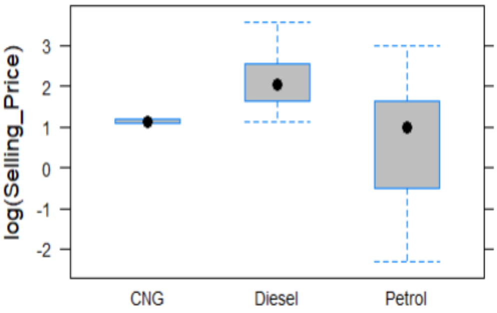
- **Selling_Price vs Driven_kms** : Both are positive skewed. If we plot them after log-transformation, we notice the least-square fitted line has a positive slope, while one intuitively suspect that it should be negative (a car that is driven more, get lower price). May be it is because there are hidden factors in overall plot.

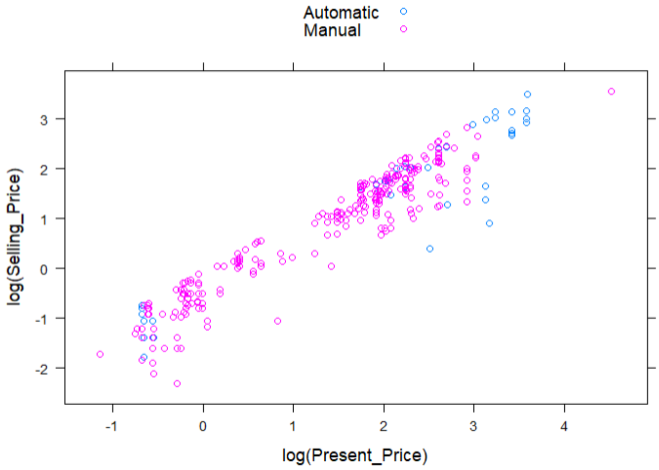


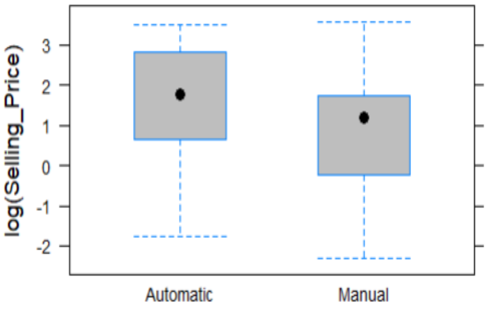
For example if we group the data with respect to Fuel_Type , for CNG and diesel the slope becomes negative.

- **Selling_Price vs Fuel_Type** : Cars will diesel type have higher price than petrol type on average. Very few obs from CNG.

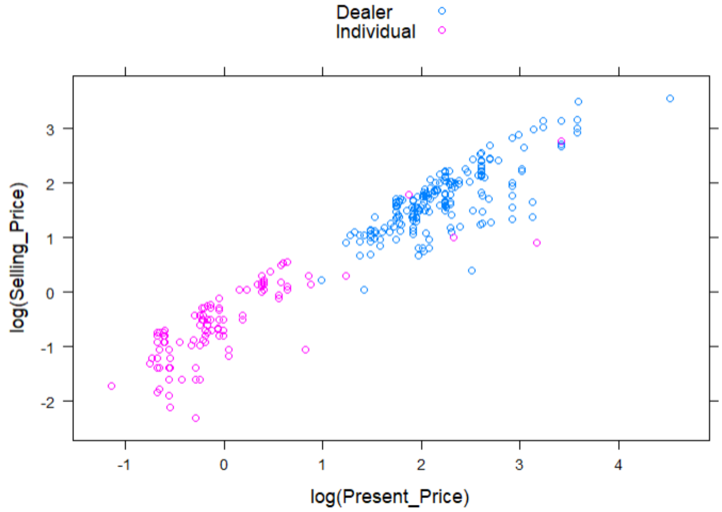


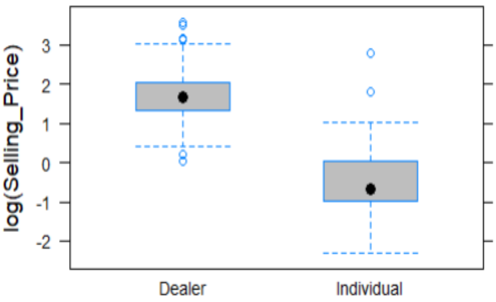






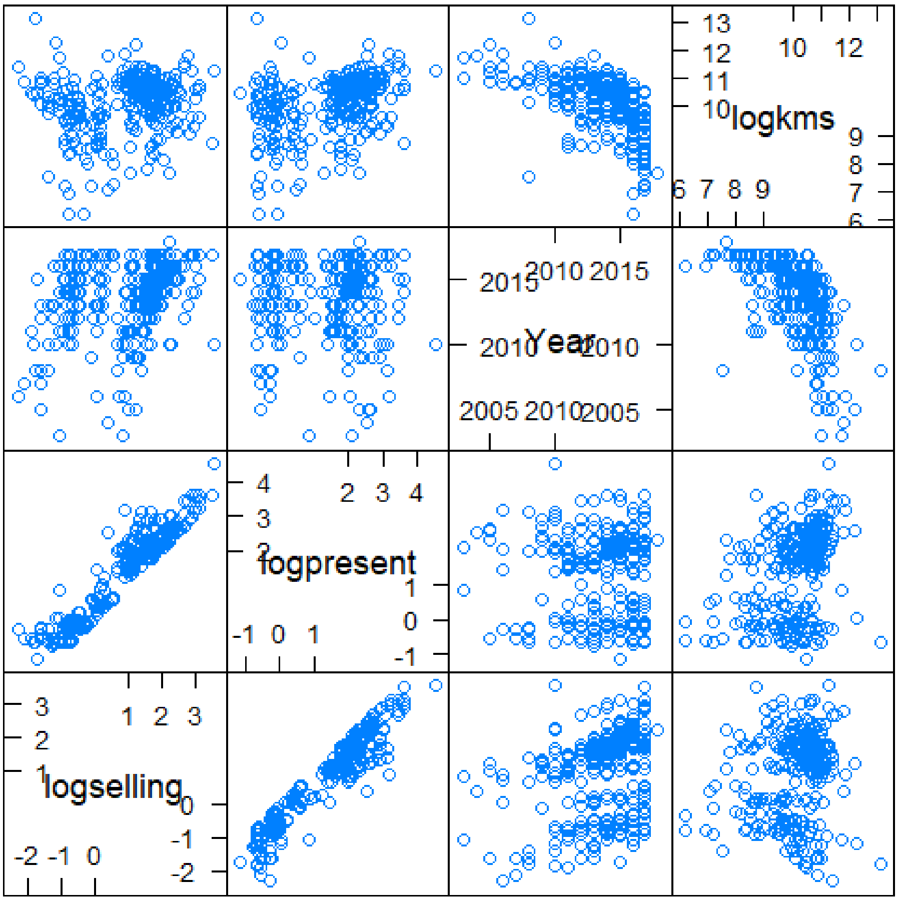
- **Selling_Price vs Transmission** : Cars with automatic transmission are likely to get higher price than cars with manual transmission.
- **Selling_Price vs Selling_Type** : Cars that are sold through dealer are likely to get higher price than that are not.





## Correlation Structure:

The scatter-plot matrix of numerical variables are given by-

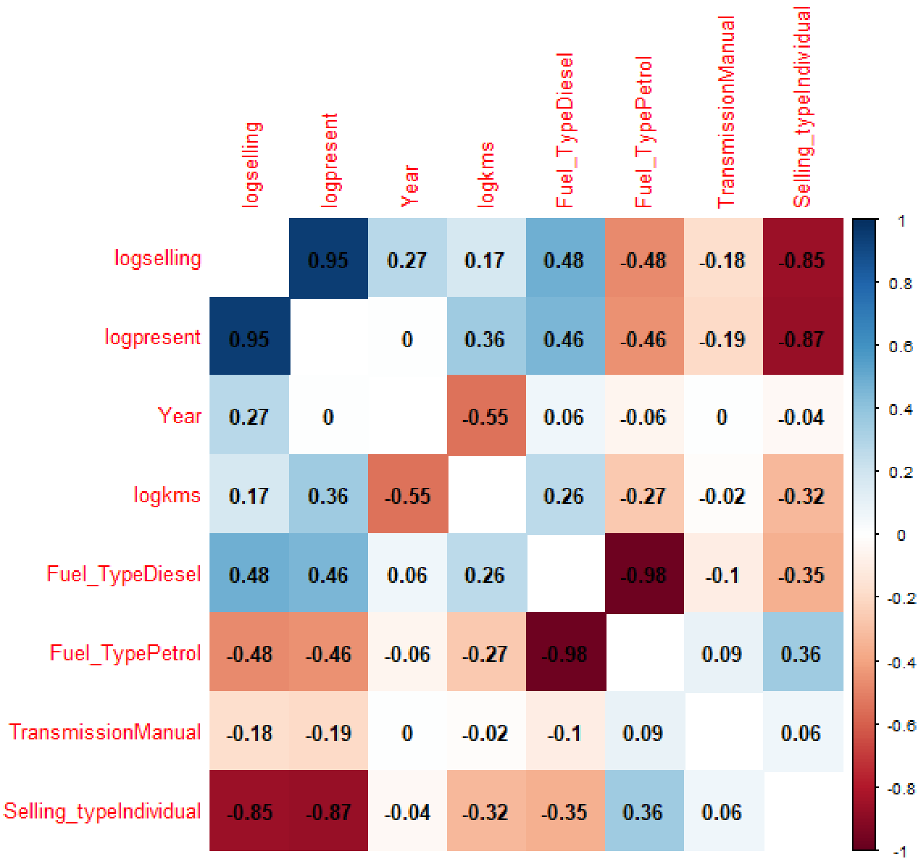


Scatter Plot Matrix

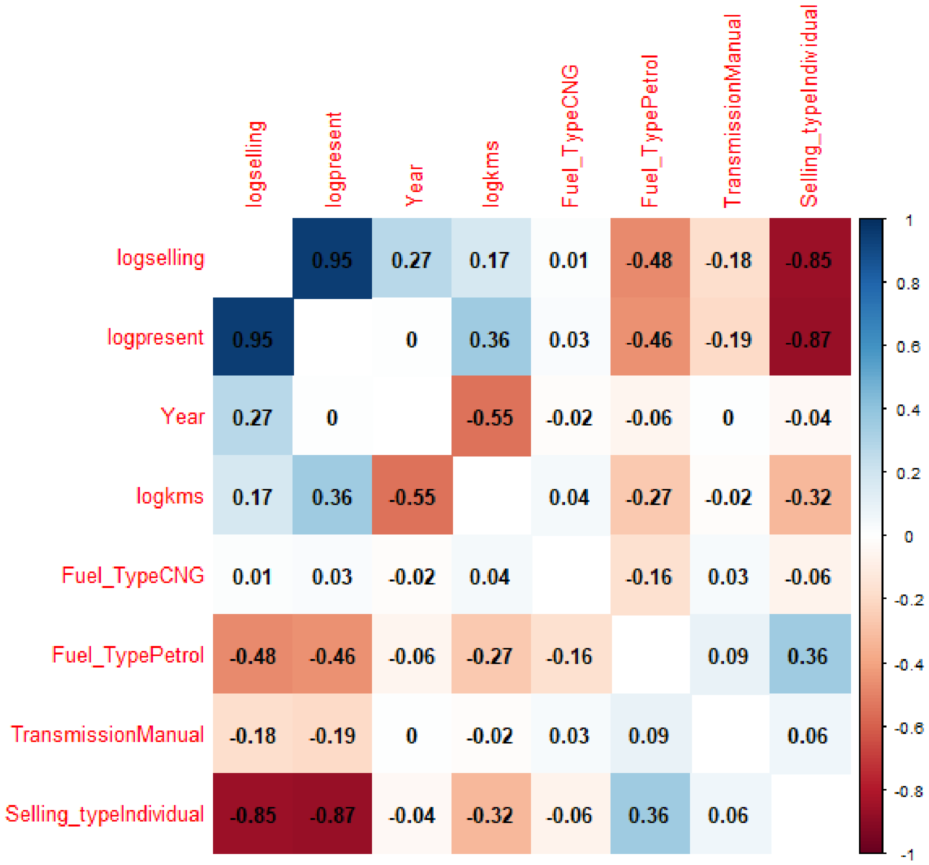
Here log(Selling_Price) and log(Present_Price) are highly positively correlated. Also we can encode categorical variables using dummy variables and compute correlation matrix. [ Note: we use k-1 dummy for a variable with k levels, to avoid exact linear relationship between dummy variables.
 Still dummy variables corresponding to Fuel_TypePetrol & Fuel_TypeDiesel are highly negatively correlated, since there are very few observation from CNG. So

*Indicator(Petrol)+ Indicator(Diesel)+Indicator(CNG) = 1*

*Indicator(CNG) = 0, Indicator(Petrol)+ Indicator(Diesel) = 1* ,for most observations. But taking dummy variables corresponding to Fuel_typeCNG & Fuel_TypePetrol avoid this problem.]



| | logselling | logpresent | Year | logkms | Fuel_TypeDiesel | Fuel_TypePetrol | TransmissionManual | Selling_typeIndividual |
|---|---|---|---|---|---|---|---|---|
| logselling | | 0.95 | 0.27 | 0.17 | 0.48 | -0.48 | -0.18 | -0.85 |
| logpresent | 0.95 | | 0 | 0.36 | 0.46 | -0.46 | -0.19 | -0.87 |
| Year | 0.27 | 0 | | -0.55 | 0.06 | -0.06 | 0 | -0.04 |
| logkms | 0.17 | 0.36 | -0.55 | | 0.26 | -0.27 | -0.02 | -0.32 |
| Fuel_TypeDiesel | 0.48 | 0.46 | 0.06 | 0.26 | | -0.98 | -0.1 | -0.35 |
| Fuel_TypePetrol | -0.48 | -0.46 | -0.06 | -0.27 | -0.98 | | 0.09 | 0.36 |
| TransmissionManual | -0.18 | -0.19 | 0 | -0.02 | -0.1 | 0.09 | | 0.06 |
| Selling_typeIndividual | -0.85 | -0.87 | -0.04 | -0.32 | -0.35 | 0.36 | 0.06 | |



| | logselling | logpresent | Year | logkms | Fuel_TypeCNG | Fuel_TypePetrol | TransmissionManual | Selling_typeIndividual |
|---|---|---|---|---|---|---|---|---|
| logselling | | 0.95 | 0.27 | 0.17 | 0.01 | -0.48 | -0.18 | -0.85 |
| logpresent | 0.95 | | 0 | 0.36 | 0.03 | -0.46 | -0.19 | -0.87 |
| Year | 0.27 | 0 | | -0.55 | -0.02 | -0.06 | 0 | -0.04 |
| logkms | 0.17 | 0.36 | -0.55 | | 0.04 | -0.27 | -0.02 | -0.32 |
| Fuel_TypeCNG | 0.01 | 0.03 | -0.02 | 0.04 | | -0.16 | 0.03 | -0.06 |
| Fuel_TypePetrol | -0.48 | -0.46 | -0.06 | -0.27 | -0.16 | | 0.09 | 0.36 |
| TransmissionManual | -0.18 | -0.19 | 0 | -0.02 | 0.03 | 0.09 | | 0.06 |
| Selling_typeIndividual | -0.85 | -0.87 | -0.04 | -0.32 | -0.06 | 0.36 | 0.06 | |

## Regression Analysis :

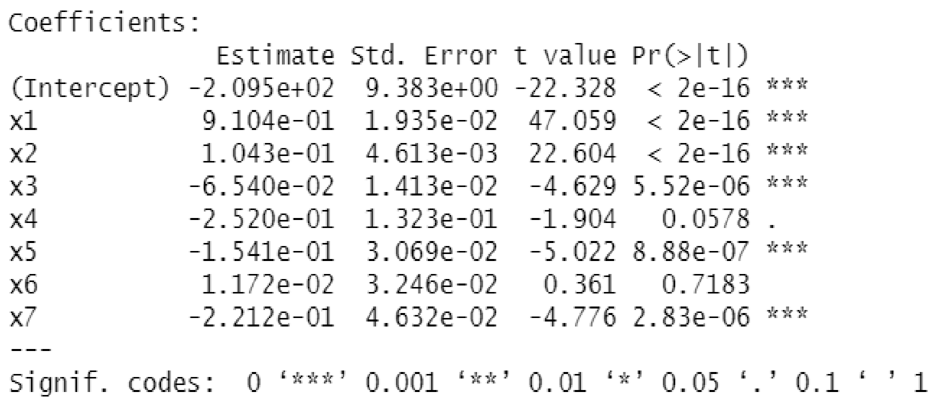
Let denote, Y= log(Selling_Price)
$X_1$= log(Present_Price)
$X_2$= Year
$X_3$= log(Driven_kms)
$X_4$= Fuel_TypeCNG
$X_5$= Fuel_TypePetrol
$X_6$= TransmissionManual
$X_7$= Selling_typeIndividual

### Full Model :

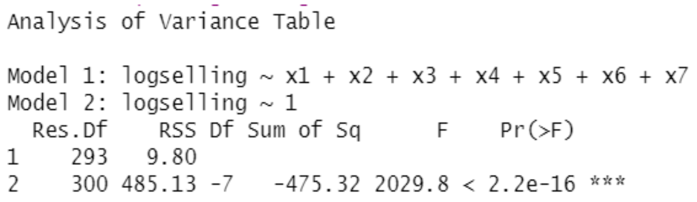
Based on our discussion in EDA, start with the following model

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_7 X_7 + \varepsilon$$

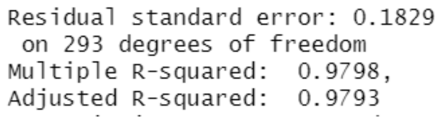
where $\alpha$, $\beta_1$, $\beta_2$,..., $\beta_7$ unknown parameter, $\varepsilon$ random error. Using least square method we get coefficients are significantly different from zero except $\beta_4$, $\beta_6$. This linear model differs significantly from "intercept only" model. It explains about 98% of total variability in response variable.



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.095e+02  9.383e+00 -22.328  < 2e-16 ***
x1           9.104e-01  1.935e-02  47.059  < 2e-16 ***
x2           1.043e-01  4.613e-03  22.604  < 2e-16 ***
x3          -6.540e-02  1.413e-02  -4.629 5.52e-06 ***
x4          -2.520e-01  1.323e-01  -1.904   0.0578 .  
x5          -1.541e-01  3.069e-02  -5.022 8.88e-07 ***
x6           1.172e-02  3.246e-02   0.361   0.7183    
x7          -2.212e-01  4.632e-02  -4.776 2.83e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



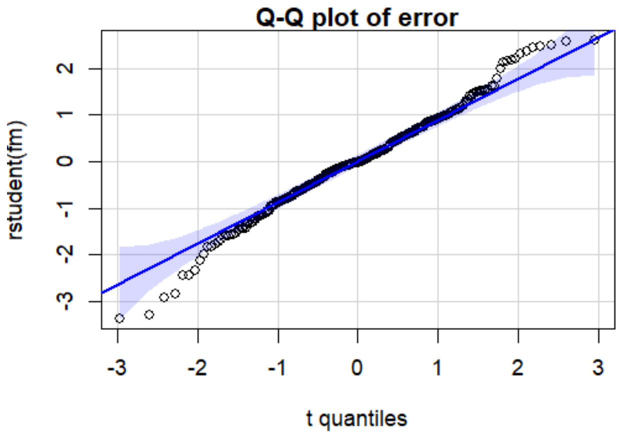
```
Analysis of Variance Table

Model 1: logselling ~ x1 + x2 + x3 + x4 + x5 + x6 + x7
Model 2: logselling ~ 1
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1    293   9.80                                  
2    300 485.13 -7   -475.32 2029.8 < 2.2e-16 ***
```

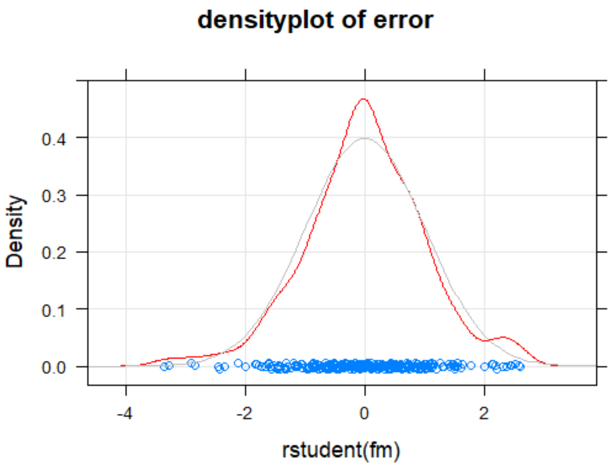


```
Residual standard error: 0.1829
 on 293 degrees of freedom
Multiple R-squared:  0.9798,
Adjusted R-squared:  0.9793
```

### Residual Analysis :

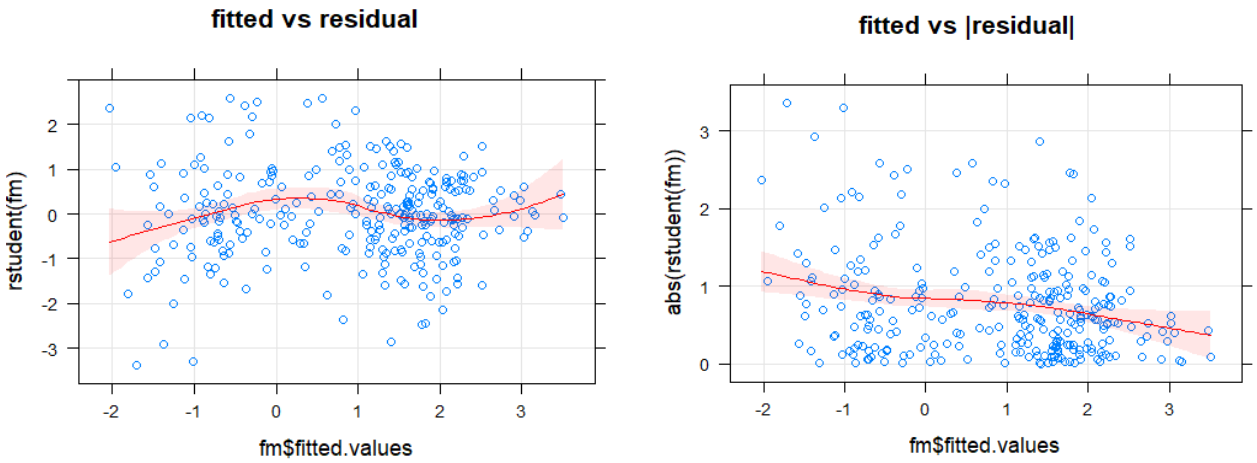
- From Q-Q plot of studentized residuals, mostly points are inside 95% confidence band, except a few at two ends. So distribution of $\varepsilon$ may produce extreme values, unlike normal distribution.





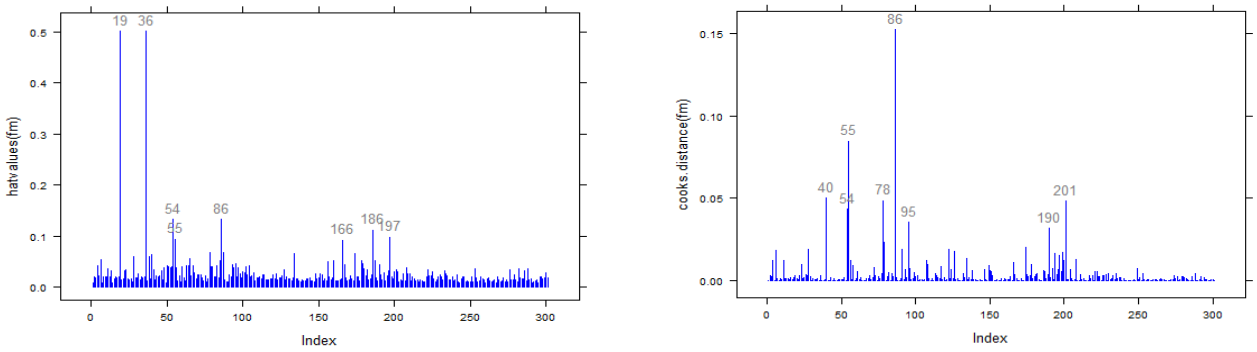
But the density plot suggests the distribution is atleast symmetric and unimodal.

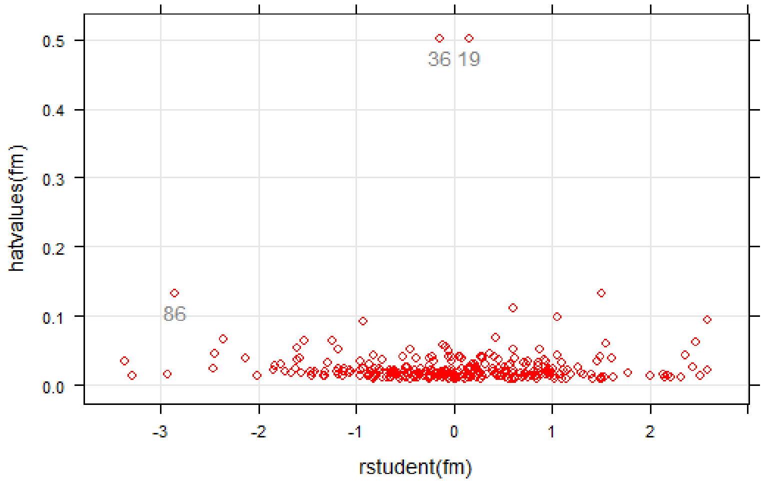
- If we plot studentized residuals against fitted values , the LOWESS line approximately equal to zero line But there is an inward funnel shape.
  In fitted value vs |residual| the LOWESS curve goes downward.



So, Homoscedasticity assumption does not holds.

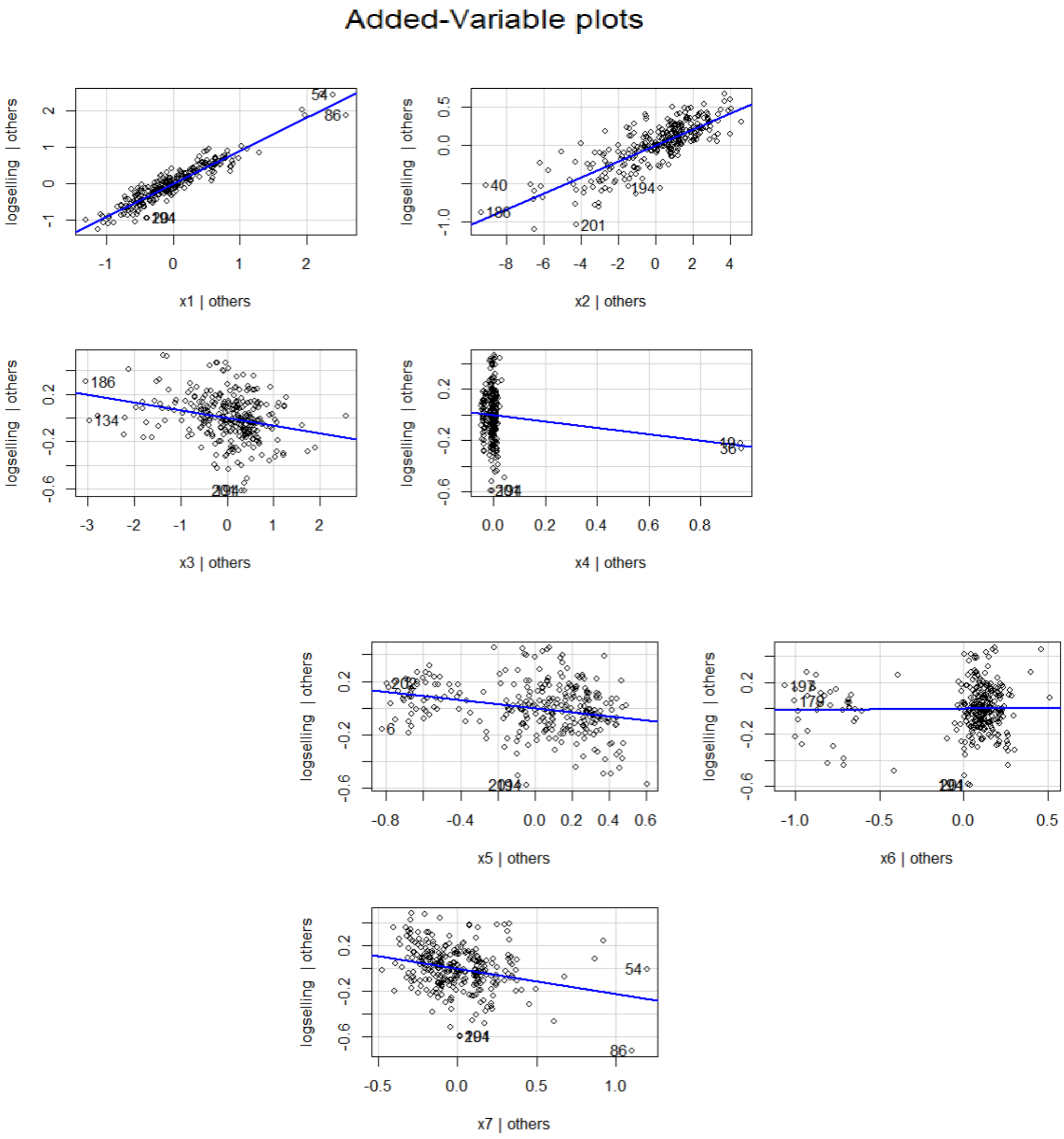
- If we plot the hat values , we see two observations, namely 19th & 36th are high leverage points. But, from the plot of cook's distances we get the most influential observation is 86th one.

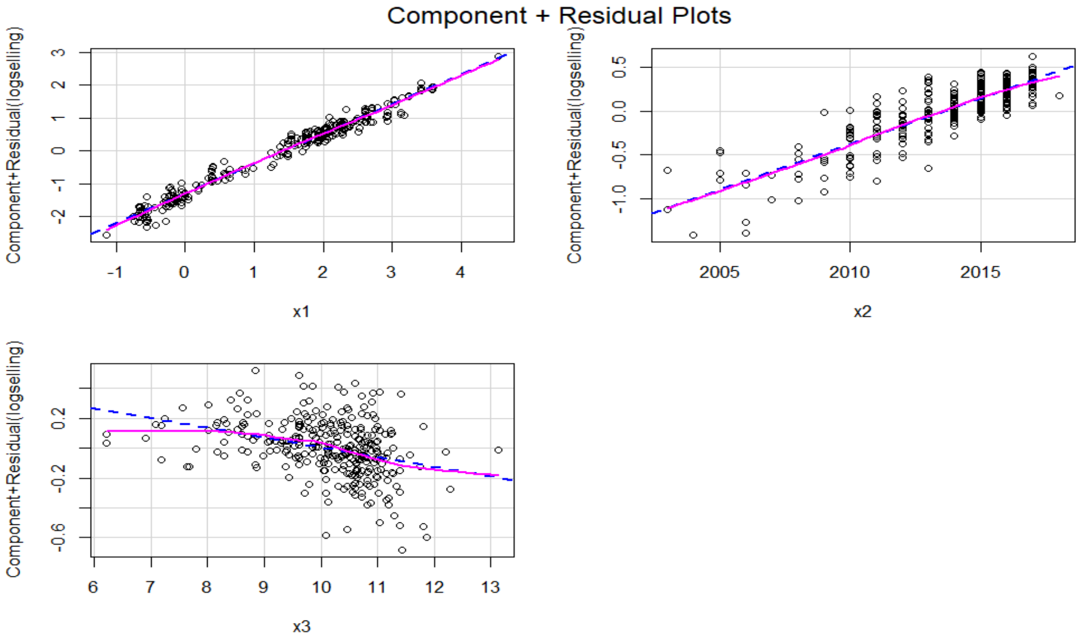




Observations 19th & 36th are not influential , as they corresponds to high leverages but low residuals. 86th observation corresponds to high residual and moderate leverage.

From 'added variable plots' we can check various unusual observations and their effects. For example- in added variable plot of $X_2$ , 186th observation is at most distance from the centroid of x-axis, it tilts the regression line by pulling to itself, in plot of $X_1$ or $X_7$ the pulling of 54th & 86th observation try to balance each other.

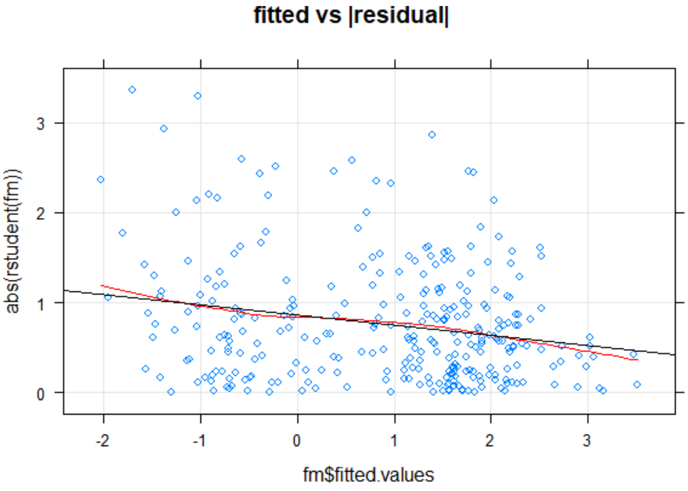




- ‘component + residual plots’ of numeric variables suggests the linearity assumption holds.

## ♦ Correction for Hetero-scedasticity :

It is important to detect and correct non constant error variance , otherwise least square estimates will still be unbiased, but they won't have minimum variance property. We know, under normality assumption Expectation of absolute residual is proportional to corresponding error standard deviation, i.e. $\mathrm{E}(|residual|) \propto \sigma$ , where $\varepsilon \sim N(0,\sigma^2)$ .

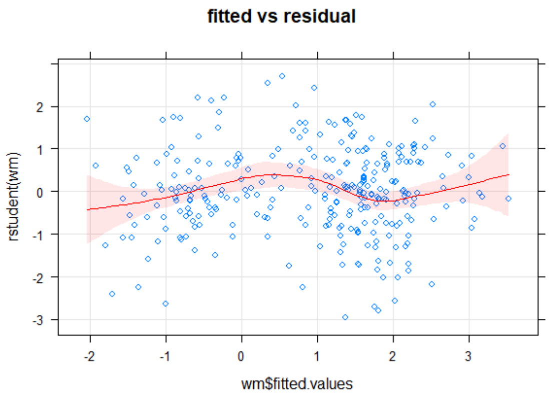


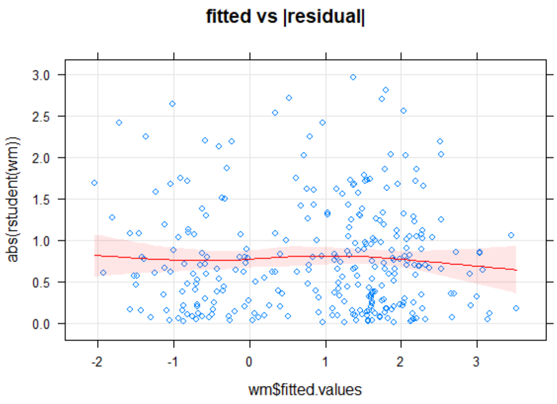
Now we can approximate $\mathrm{E}(|residual|)$ by the regression line of "|residual| vs fitted full model" , and thus can obtain a weighted least square ,where weight is proportional to reciprocal of $\hat{\sigma}^2$ (more the error variance ,less the weightage). Corresponding estimated coefficients are given by –

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.162e+02  9.549e+00 -22.637  < 2e-16 ***
x1           8.990e-01  1.781e-02  50.469  < 2e-16 ***
x2           1.076e-01  4.693e-03  22.921  < 2e-16 ***
x3          -5.948e-02  1.422e-02  -4.185 3.78e-05 ***
x4          -2.576e-01  1.248e-01  -2.064   0.0399 *  
x5          -1.658e-01  2.580e-02  -6.424 5.35e-10 ***
x6          -5.019e-03  2.891e-02  -0.174   0.8623    
x7          -2.265e-01  4.310e-02  -5.254 2.87e-07 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

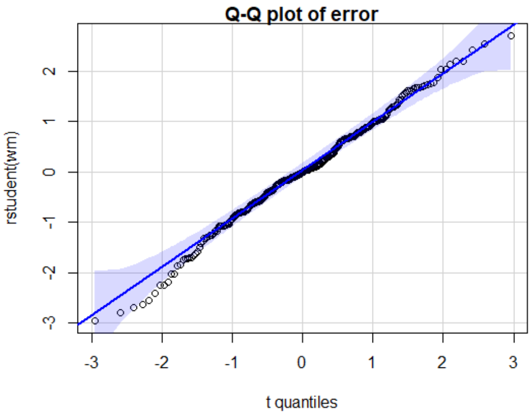
```
Residual standard error: 0.2358
 on 293 degrees of freedom
Multiple R-squared:  0.979,
Adjusted R-squared:  0.9785
```

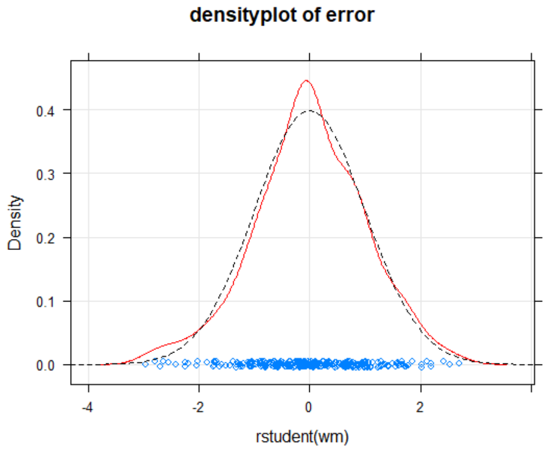
This model can explain about 98% of total variability in response variable





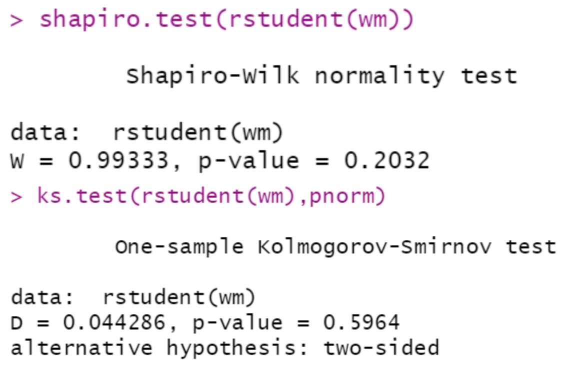
Here "fitted vs residual" plot looks same as earlier ,but "fitted vs |residual|" plot almost parallel to horizontal axis, which indicates non-constant error variance is corrected.



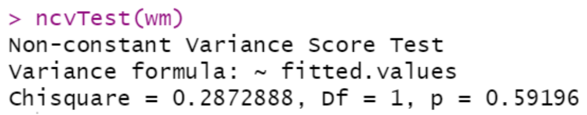


“Q-Q plot” and “density plot” of errors show that normality assumption is satisfied.

**Formal tests:**

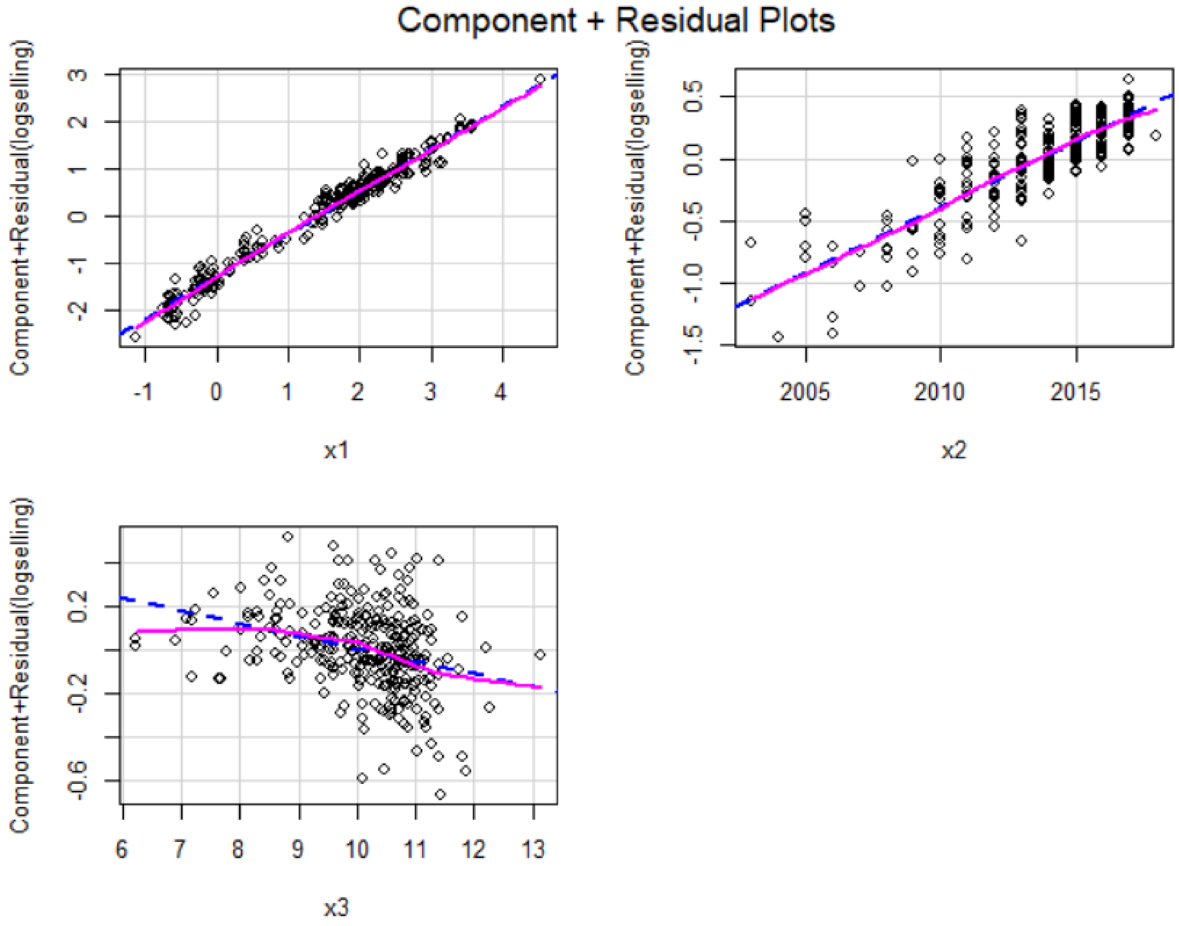


```
> shapiro.test(rstudent(wm))

        Shapiro-Wilk normality test

data:  rstudent(wm)
W = 0.99333, p-value = 0.2032
> ks.test(rstudent(wm),pnorm)

        One-sample Kolmogorov-Smirnov test

data:  rstudent(wm)
D = 0.044286, p-value = 0.5964
alternative hypothesis: two-sided
```

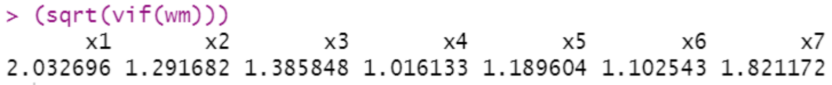


```
> ncvTest(wm)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.2872888, Df = 1, p = 0.59196
```

- “Component + Residual plot" indicates assumption of linearity is not violated. But may be including polnomial of $X_3$ improves the fit  further.



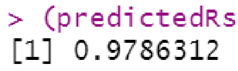
- There is no severe issue of multicollinearity in our model.



```
> (sqrt(vif(wm)))
       x1       x2       x3       x4       x5       x6       x7 
2.032696 1.291682 1.385848 1.016133 1.189604 1.102543 1.821172 
```

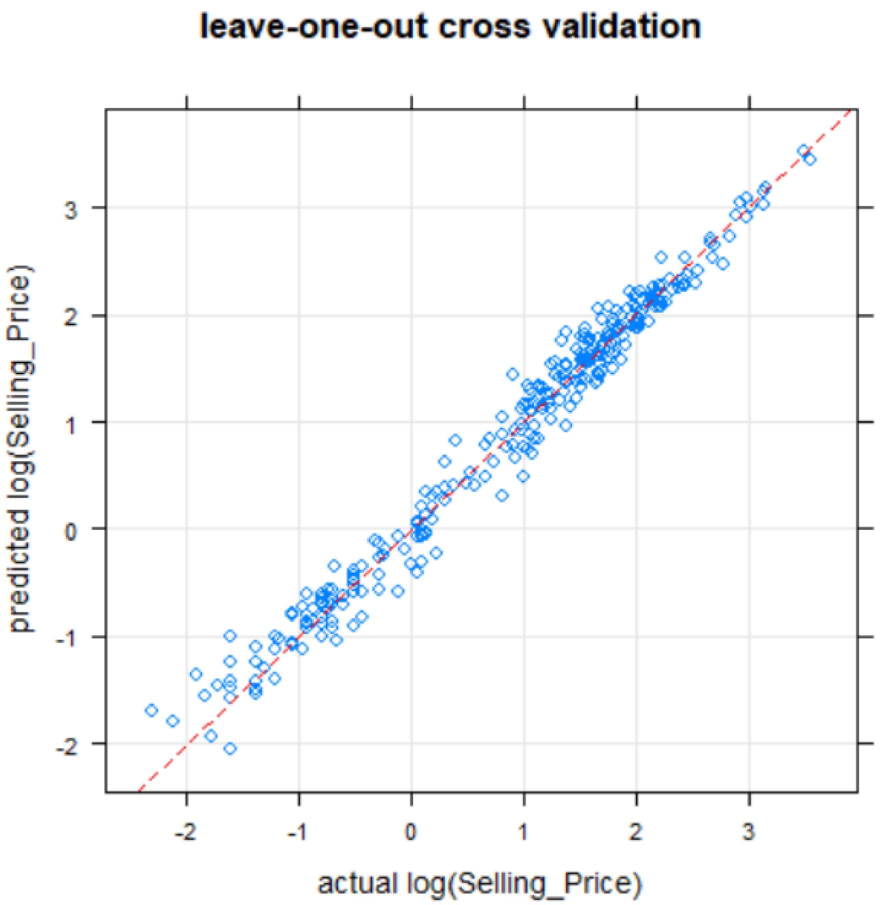
## ♦ Predictive Performance of the Model :

We will use ‘Leave one out Cross-validation’ to get an idea about predicting power of the weighted least square model. We get

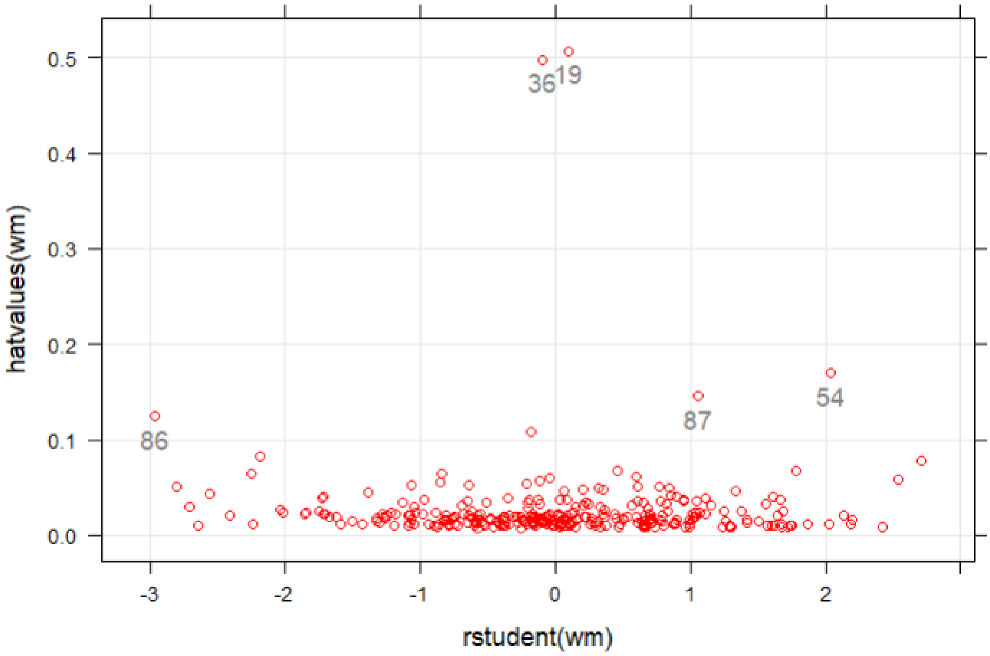


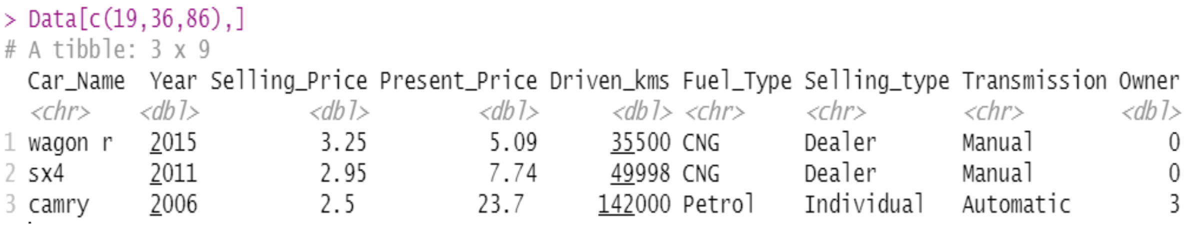
```
> (predictedRsq= 1- SE/ST)
[1] 0.9786312
```

,means this model can explain about 98% variability in new data.

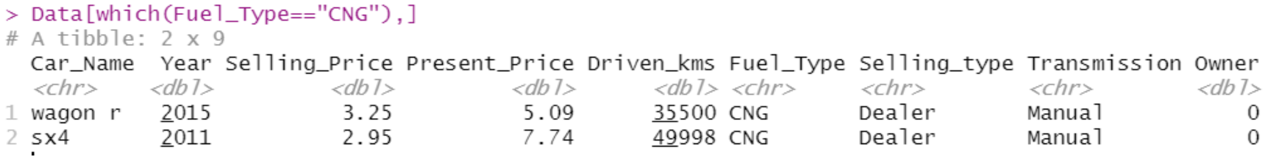


- Possible explanation for the most unusual observations :

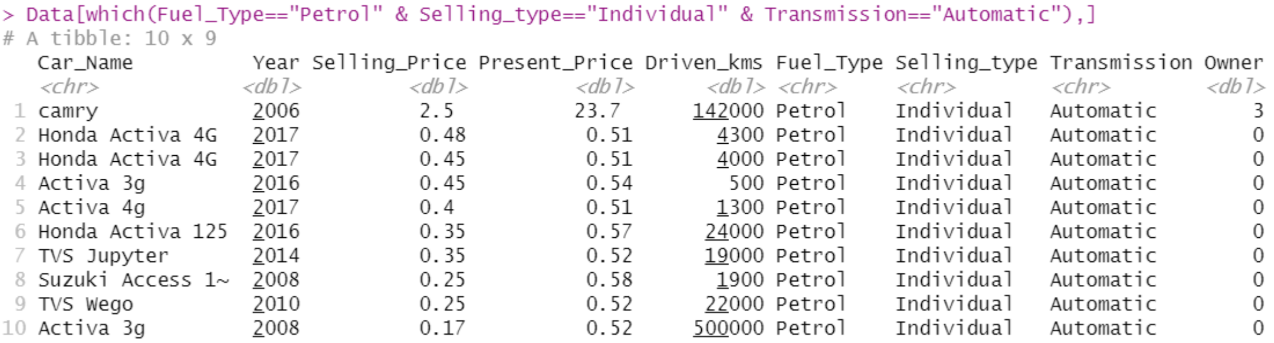



```
> Data[c(19,36,86),]
# A tibble: 3 x 9
  Car_Name  Year Selling_Price Present_Price Driven_kms Fuel_Type Selling_type Transmission Owner
  <chr>    <dbl>         <dbl>         <dbl>      <dbl> <chr>     <chr>        <chr>        <dbl>
1 wagon r   2015          3.25          5.09      35500 CNG       Dealer       Manual           0
2 sx4       2011          2.95          7.74      49998 CNG       Dealer       Manual           0
3 camry     2006          2.5          23.7      142000 Petrol    Individual   Automatic        3
```

The extreme high leverage of observations 19th and 36th is due to the fact that they are the only two observations with Fuel_Type = CNG among all 301observations, sending them far apart from the centroid of x-space.

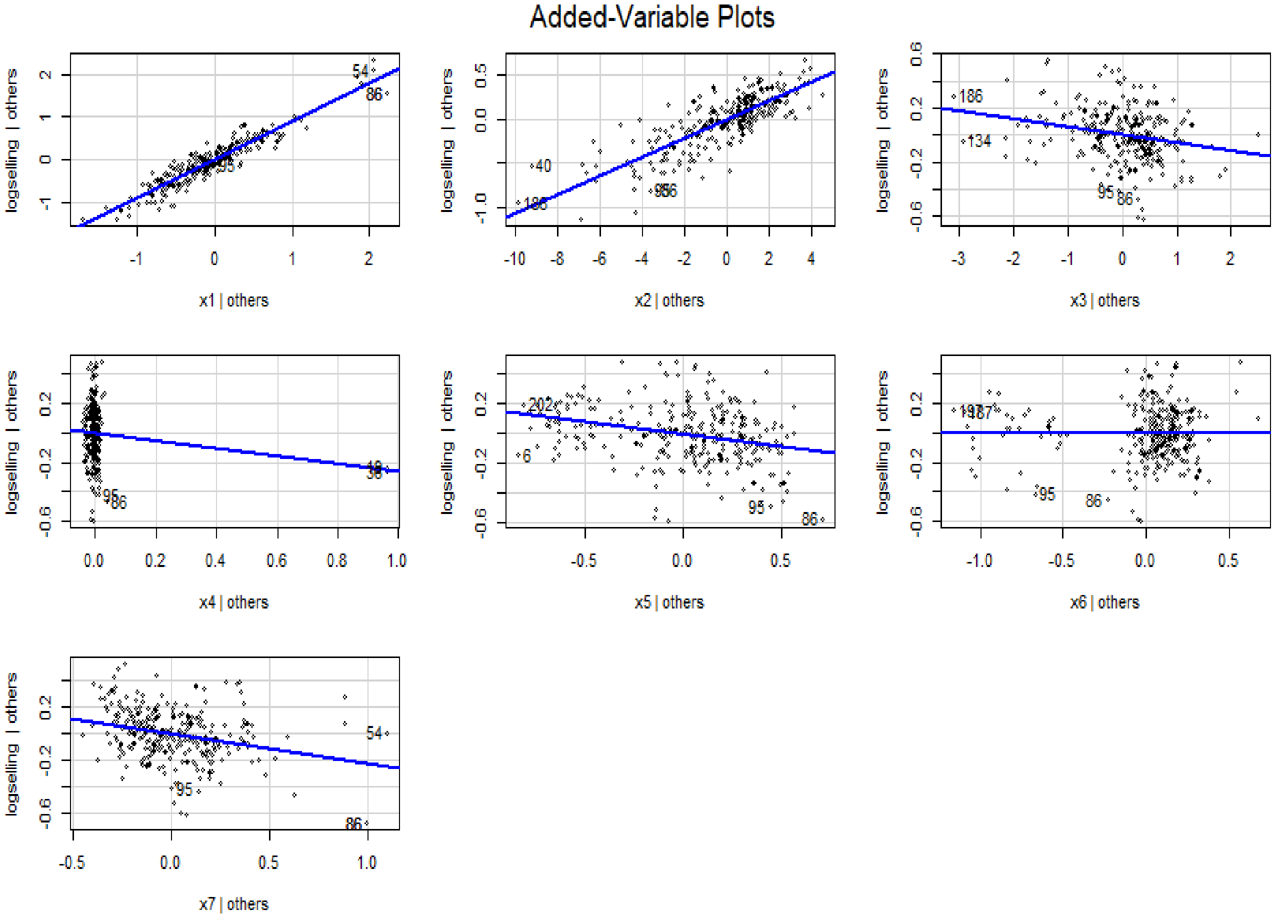


```
> Data[which(Fuel_Type=="CNG"),]
# A tibble: 2 x 9
  Car_Name  Year Selling_Price Present_Price Driven_kms Fuel_Type Selling_type Transmission Owner
  <chr>    <dbl>         <dbl>         <dbl>      <dbl> <chr>     <chr>        <chr>        <dbl>
1 wagon r   2015          3.25          5.09      35500 CNG       Dealer       Manual           0
2 sx4       2011          2.95          7.74      49998 CNG       Dealer       Manual           0
```



```
> Data[which(Fuel_Type=="Petrol" & Selling_type=="Individual" & Transmission=="Automatic"),]
# A tibble: 10 x 9
   Car_Name           Year Selling_Price Present_Price Driven_kms Fuel_Type Selling_type Transmission Owner
   <chr>             <dbl>         <dbl>         <dbl>      <dbl> <chr>     <chr>        <chr>        <dbl>
 1 camry              2006          2.5          23.7      142000 Petrol    Individual   Automatic        3
 2 Honda Activa 4G    2017          0.48          0.51       4300 Petrol    Individual   Automatic        0
 3 Honda Activa 4G    2017          0.45          0.51       4000 Petrol    Individual   Automatic        0
 4 Activa 3g          2016          0.45          0.54        500 Petrol    Individual   Automatic        0
 5 Activa 4g          2017          0.4           0.51       1300 Petrol    Individual   Automatic        0
 6 Honda Activa 125   2016          0.35          0.57      24000 Petrol    Individual   Automatic        0
 7 TVS Jupyter        2014          0.35          0.52      19000 Petrol    Individual   Automatic        0
 8 Suzuki Access 1~   2008          0.25          0.58       1900 Petrol    Individual   Automatic        0
 9 TVS Wego           2010          0.25          0.52      22000 Petrol    Individual   Automatic        0
10 Activa 3g          2008          0.17          0.52     500000 Petrol    Individual   Automatic        0
```

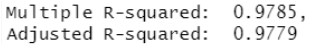
Among all the observations with Selling_type = Individual , Transmission = Automatic , Fuel_Type = Petrol , for the 86th observation Selling_Price and Present_Price differs very much from others. So, its presence or absence change the regression plane a lot, making 86th observation the one with highest cooks-distance.



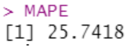
## Finding sparser model :

We have already seen that, in our model not all estimated coefficients are significantly different from zero . We can try to find model with fewer term , without losing fit and predictive performance very much. From now on, we split our data in two parts with Train_Set : Test_Set = 80 : 20 ratio. We will use the first one to estimate model coefficients , and last one to evaluate its predictive performance.

Based on these splitting , we fit our model $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_7 X_7 + \varepsilon$ by weighted least square and get –

On **Train_Set :** 

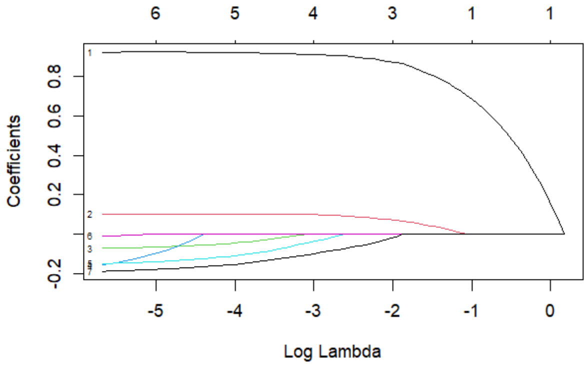
```
Multiple R-squared:  0.9785,
Adjusted R-squared:  0.9779
```

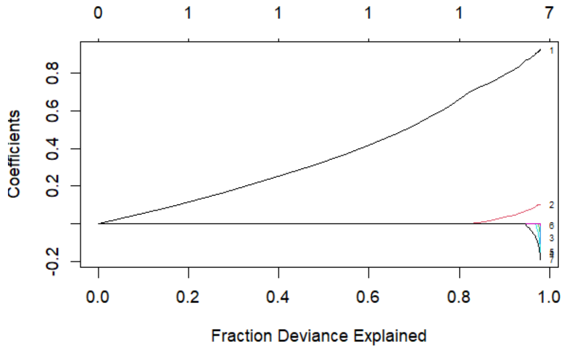
On **Test_Set :** 

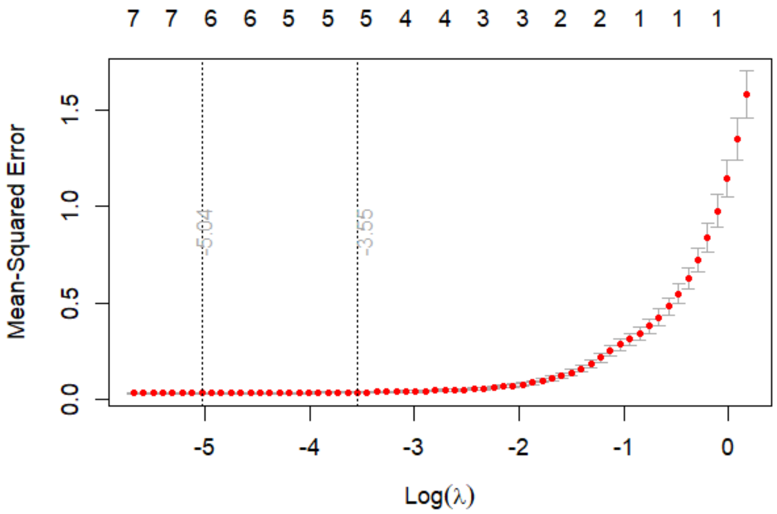
```
> MAPE
[1] 25.7418
```

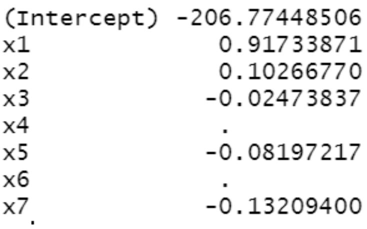
### Penalized Regression – LASSO :

Initially starting with 7 predictors, as we increase penalty , more and more coefficients estimated as zero, at a cost of decrease in explained fraction deviance. For optimum choice of penalty parameter, we take the maximum possible value ( implies maximum possible sparsitity), for which MSE is within one standard error of minimum MSE( implies best fitting). We get the estimated coefficients



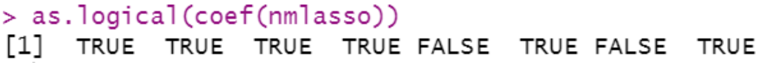






```
(Intercept) -206.77448506
x1             0.91733871
x2             0.10266770
x3            -0.02473837
x4             .
x5            -0.08197217
x6             .
x7            -0.13209400
```

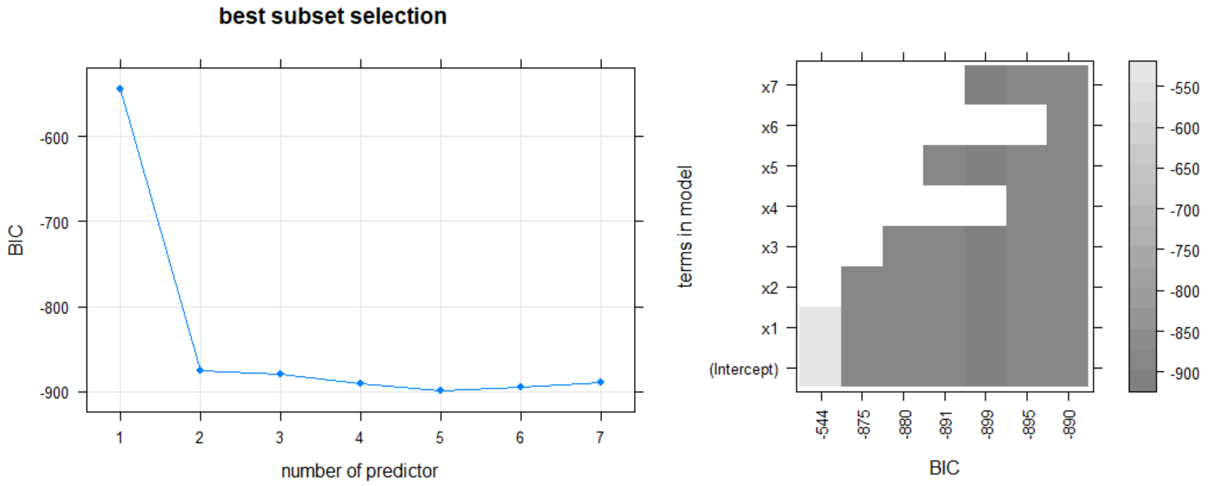
clearly, our selected predictors are $X_1$ , $X_2$ , $X_3$ , $X_5$ , $X_6$ .



```
> as.logical(coef(nmlasso))
[1]  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
```

### Best Subset Selection :

Since our total number of predictors is not large , we can search, through all possible subsets, for the one giving minimum BIC.

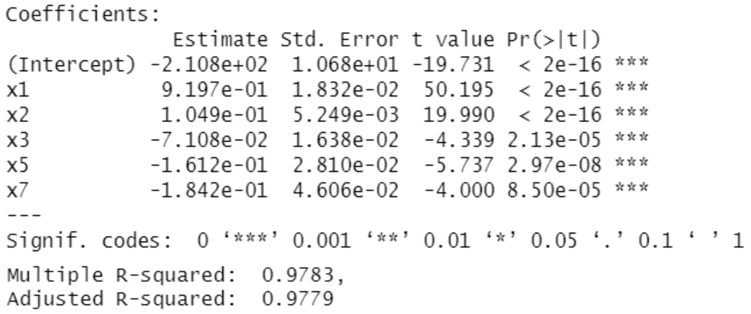


The selected subset here also has five predictors with intercept term, namely X$_1$ , X$_2$ , X$_3$ , X$_5$ , X$_6$ .

## Fitting with selected variables :

Now our model becomes $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_5 X_5 + \beta_7 X_7 + \varepsilon$

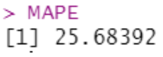
- On **Train_Set :**



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.108e+02  1.068e+01 -19.731  < 2e-16 ***
x1           9.197e-01  1.832e-02  50.195  < 2e-16 ***
x2           1.049e-01  5.249e-03  19.990  < 2e-16 ***
x3          -7.108e-02  1.638e-02  -4.339 2.13e-05 ***
x5          -1.612e-01  2.810e-02  -5.737 2.97e-08 ***
x7          -1.842e-01  4.606e-02  -4.000 8.50e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Multiple R-squared:  0.9783,
Adjusted R-squared:  0.9779
```

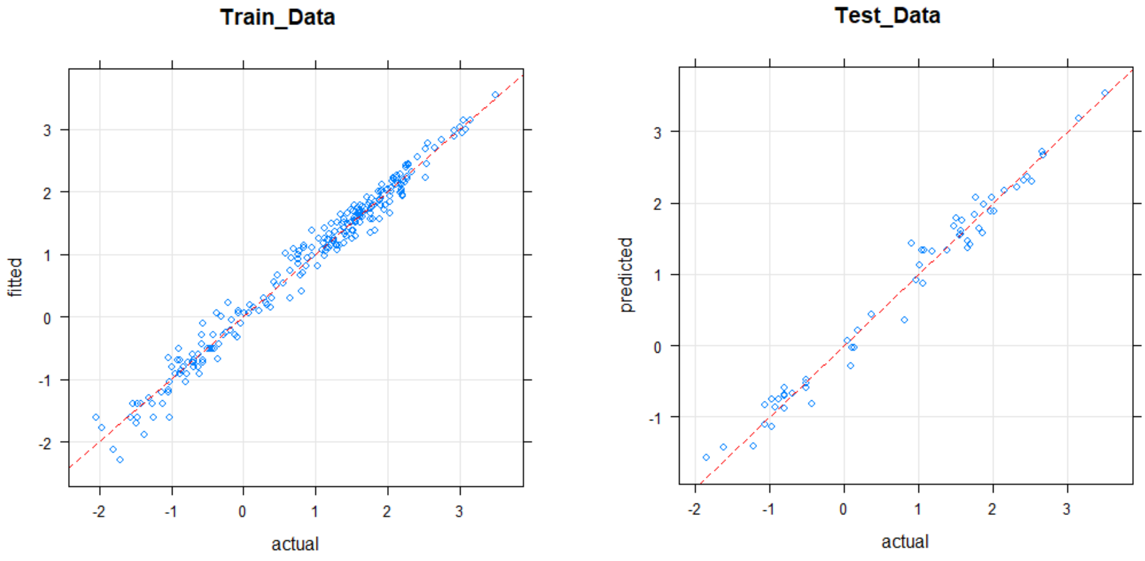
Here partial tests of all coefficients are rejected, that implies each predictor has significant contribution given other terms in the model. This fitted model can explain about 98% of total variation in response variable.

- On **Test_Set :**



```
> MAPE
[1] 25.68392
```

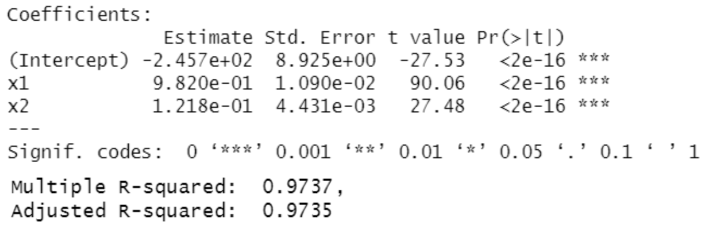
which means the average deviation between actual and forecasted values is about 25% of actual value in magnitude.



- But from the ‘number of predictor vs BIC’ plot in best subset selection we notice that-

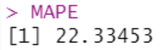
We can drop number of variables from five to two , without gaining much BIC. Best subsetof size two is $X_1$ ,$X_2$ . So we now try the model $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

- On **Train_Set :**



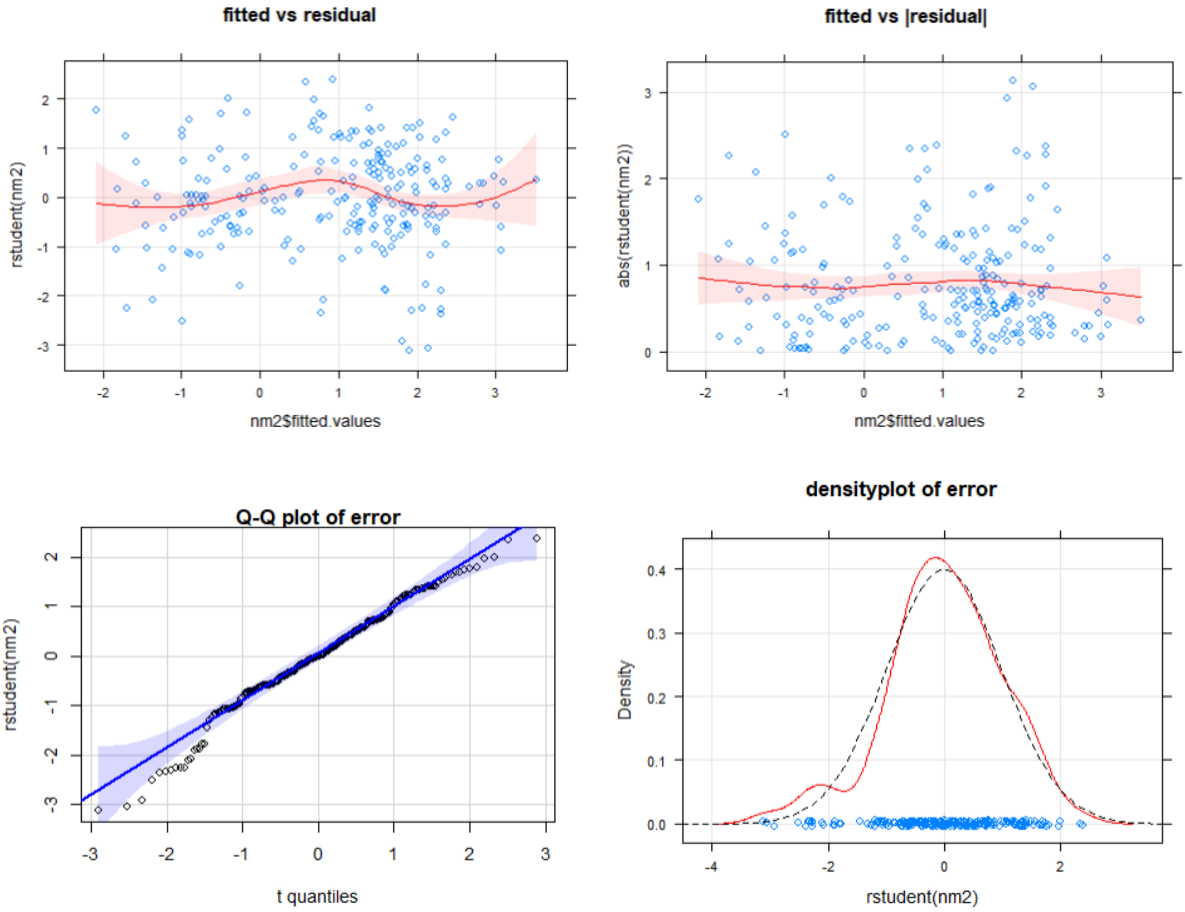
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.457e+02  8.925e+00  -27.53   <2e-16 ***
x1           9.820e-01  1.090e-02   90.06   <2e-16 ***
x2           1.218e-01  4.431e-03   27.48   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Multiple R-squared:  0.9737,
Adjusted R-squared:  0.9735
```

- On **Test_Set :**

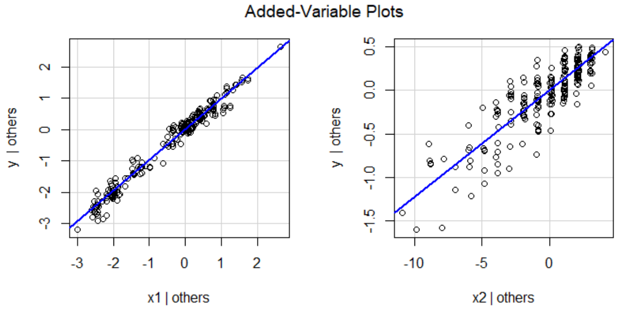


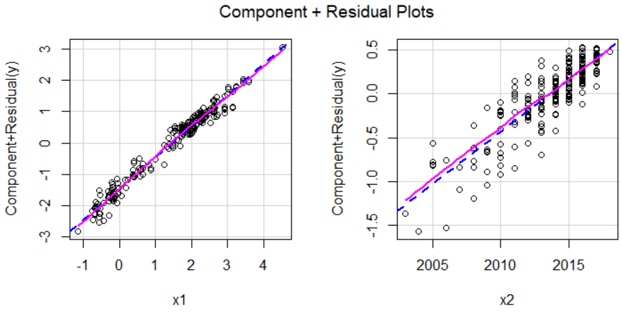
```
> MAPE
[1] 22.33453
```

- **Residual Plots :**

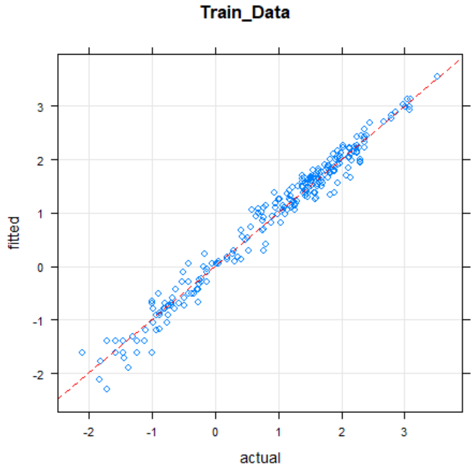


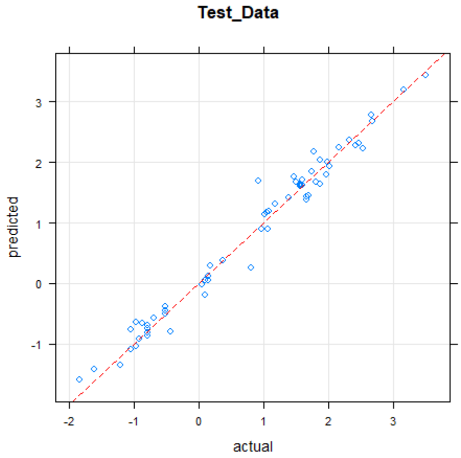
Here residual plots suggests no severe violation of linear regression assumptions , except the possibility that the error distribution may be slightly heavy tailed at left side.





- This fitted model can explain  about 97% of variability in response variable.
  On new data, average deviation between actual and fitted value is 22% of the actual value in magnitude.

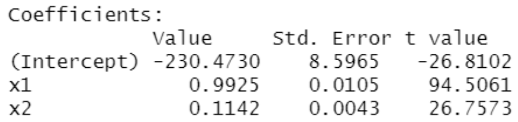




## ♦ Robust Regression :

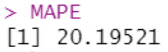
We can also try Robust Regression method -

- On **Train_Set :**

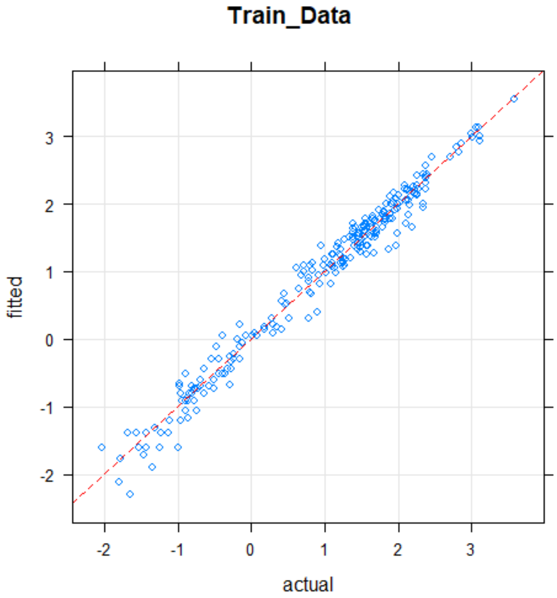


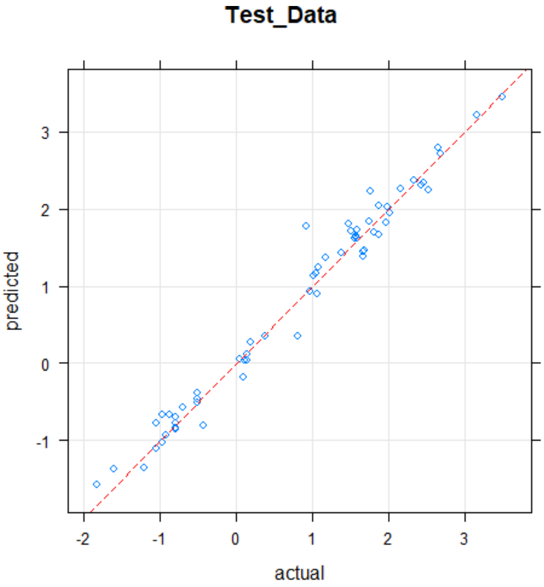
```
 Coefficients:
             Value     Std. Error t value 
(Intercept) -230.4730    8.5965   -26.8102
x1             0.9925    0.0105    94.5061
x2             0.1142    0.0043    26.7573
```

- On **Test_Set :**



```
> MAPE
[1] 20.19521
```





Here the predictive performance further improved slightly.

## ♦ Prediction Model Chosen :

| **SUMMARY** | $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_7 X_7 + \varepsilon$ | $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_5 X_5 + \beta_7 X_7 + \varepsilon$ | $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ |
|---|---|---|---|
| $\widehat{\alpha}$ | -210.450 | -210.718 | -245.687 |
| $\widehat{\beta_1}$ | 0.911 | 0.920 | 0.982 |
| $\widehat{\beta_2}$ | 0.105 | 0.105 | 0.122 |
| $\widehat{\beta_3}$ | -0.071 | -0.071 | - |
| $\widehat{\beta_4}$ | -0.228 | - | - |
| $\widehat{\beta_5}$ | -0.167 | -0.161 | - |
| $\widehat{\beta_6}$ | -0.025 | - | - |
| $\widehat{\beta_7}$ | -0.199 | -0.184 | - |
| $R^2$ | 0.979 | 0.978 | 0.974 |
| $no.of\ parameter$ | 8 | 6 | 3 |
| $R^2_{adj}$ | 0.978 | 0.978 | 0.974 |
| $MAPE$ | 25.74 | 25.68 | 22.33 |

*USE THIS ONE*

**Note :** By dropping number of variable to two , decrease in $R^2$ (that is loss in explained variability) is very little. But MAPE value (that is error in predicting new data) also decreased ,instead of increasing. So our earlier models were overfitted.

- A 95% prediction interval of $Y$ , for new data $X_1$=$x_{01}$ , $X_2$=$x_{02}$ ,..., $X_7$=$x_{07}$ is given by-

$$\left[\ \widehat{y_0} - 1.97\sqrt{\widehat{\sigma^2}(1 + x_0' M x_0)}\ ,\ \widehat{y_0} + 1.97\sqrt{\widehat{\sigma^2}(1 + x_0' M x_0)}\ \right]$$

where $x_0'$ = ( 1 , $x_{01}$ , $x_{02}$ ,..., $x_{07}$ )

$\widehat{y_0}$ = -245.687 + 0.982 $x_{01}$ + 0.122 $x_{02}$ → predicted value of $Y$ at $x_0$

$\widehat{\sigma^2}$ = 0.039 → estimated error variance

$$M = \begin{pmatrix} 2024.417 & 0.017 & -1.005 \\ 0.017 & 0.003 & 0.000 \\ -1.005 & 0.000 & 0.000 \end{pmatrix}$$ → $(X'X)^{-1}$ , $X$ is model matrix

- Here $\widehat{y_0}$ gives the predicted value of log(Selling_Price) at $x_0$ .
  So a reasonable prediction for Selling_Price is $e^{\widehat{y_0}}$

## Conclusion :

Starting with many variables , finally it turns out that not all of them important. It is reasonable that, information about how old the car is , carries information about its driven kms. Similarly , model type , fuel type etc all determines the present price of car. So Having information about present price of same car model and year of purchase of the used car is enough to make good prediction of its resale value.