Prediction of Resale Value of Used Cars

Ramit Nandi , roll- MD2211

12-Nov-2022

Dataset Description

- Here our Data is collected fromhttps://www.kaggle.com/datasets/vijayaadithyanvg/car-price-predictionused-cars

```
dim(Data)
```

```
## [1] 301   9
```

```
summary(is.na(Data))
```

```
##   Car_Name            Year          Selling_Price   Present_Price
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:301       FALSE:301       FALSE:301       FALSE:301
##  Driven_kms      Fuel_Type       Selling_type    Transmission
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:301       FALSE:301       FALSE:301       FALSE:301
##    Owner
##  Mode :logical
##  FALSE:301
```

```
head(Data[,1:5])
```

```
## # A tibble: 6 x 5
##   Car_Name       Year Selling_Price Present_Price Driven_kms
##   <chr>         <dbl>         <dbl>         <dbl>      <dbl>
## 1 ritz           2014          3.35          5.59      27000
## 2 sx4            2013          4.75          9.54      43000
## 3 ciaz           2017          7.25          9.85       6900
## 4 wagon r        2011          2.85          4.15       5200
## 5 swift          2014          4.6           6.87      42450
## 6 vitara brezza  2018          9.25          9.83       2071
```

```
head(Data[,c(1,6:9)])
```

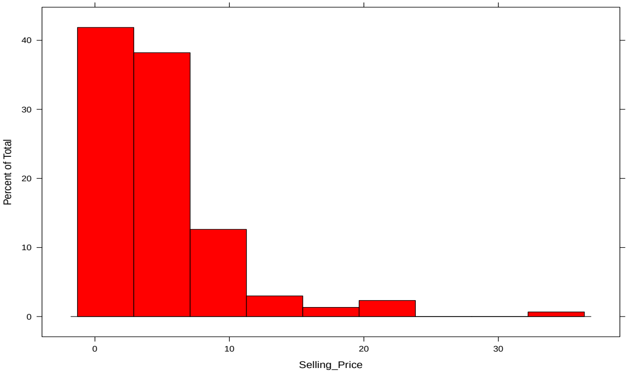
```
## # A tibble: 6 x 5
##   Car_Name      Fuel_Type Selling_type Transmission Owner
##   <chr>         <chr>     <chr>        <chr>        <dbl>
## 1 ritz          Petrol    Dealer       Manual           0
## 2 sx4           Diesel    Dealer       Manual           0
## 3 ciaz          Petrol    Dealer       Manual           0
## 4 wagon r       Petrol    Dealer       Manual           0
## 5 swift         Diesel    Dealer       Manual           0
## 6 vitara brezza Diesel    Dealer       Manual           0
```

EDA

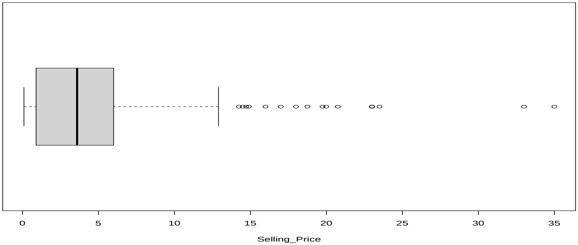
Selling_Price : the response

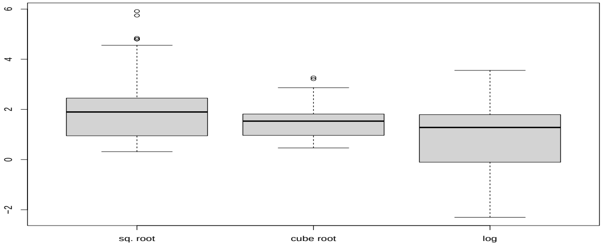
```
summary(Selling_Price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.100   0.900   3.600   4.661   6.000  35.000
```

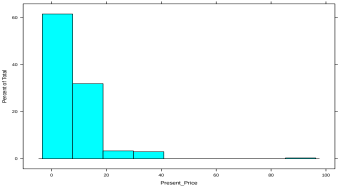


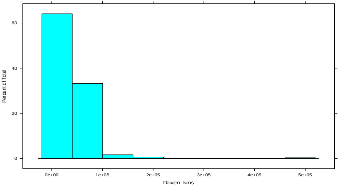
Selling_Price

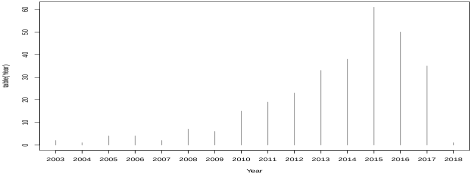




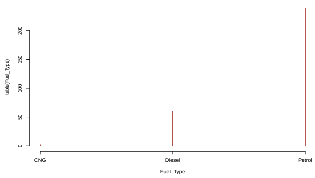
possible Numerical predictors

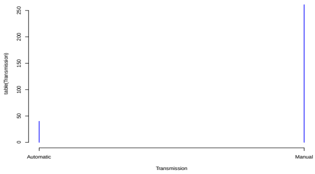


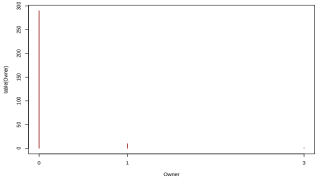




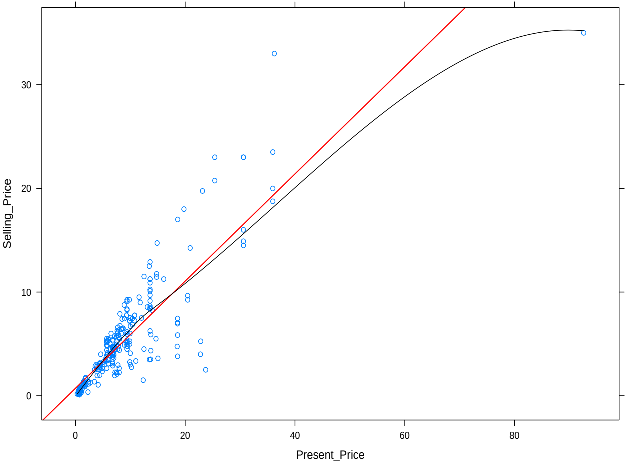
possible Categorical predictors







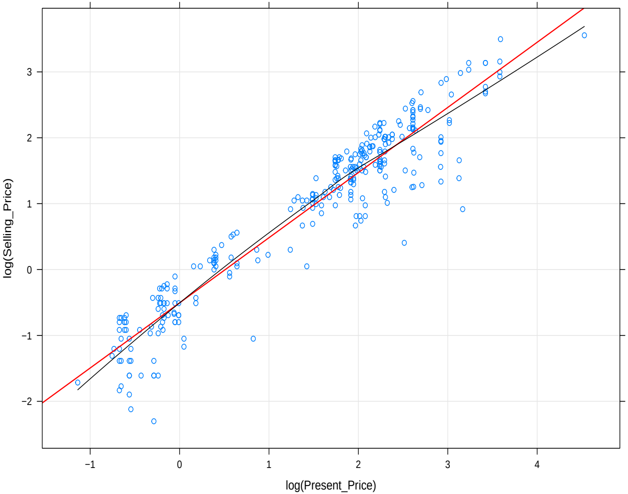
Selling_Price vs Present_Price



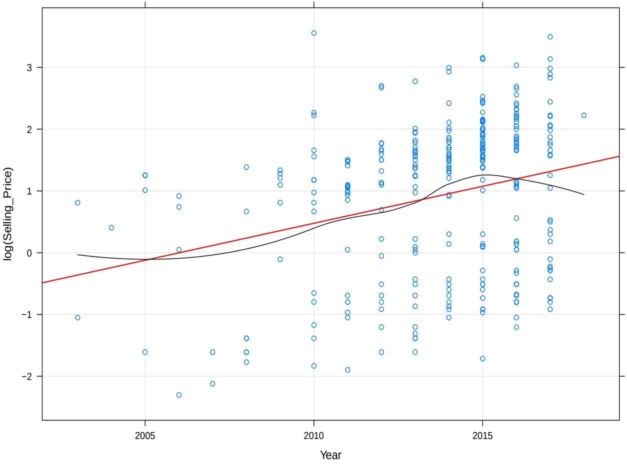
- points are overlapped in a small region

Selling_Price vs Present_Price

- linear relationship is now more clear

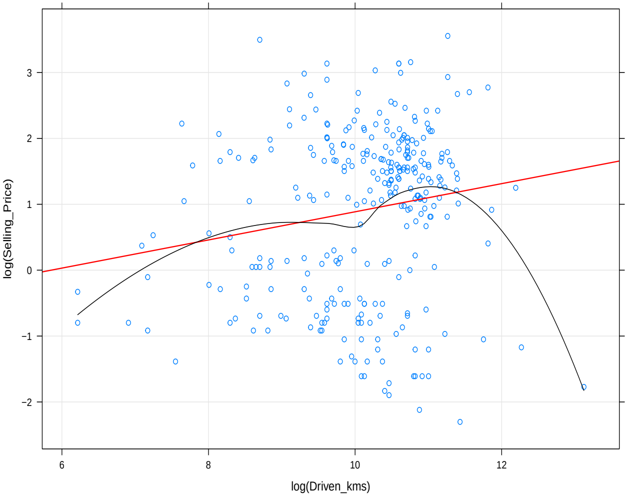


Selling_Price vs Year

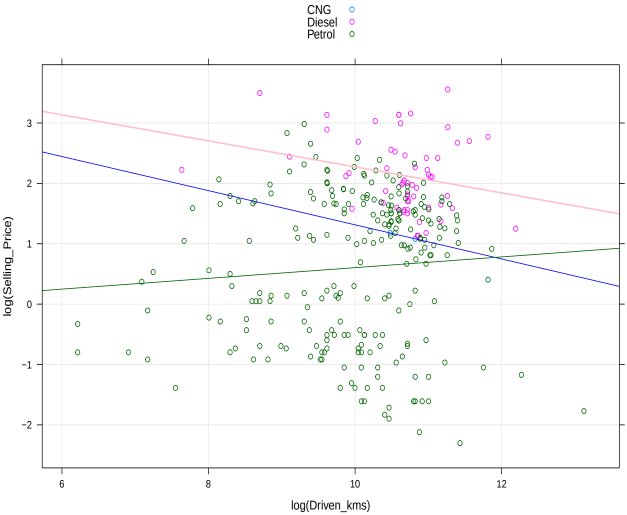


- older car gets lower resale value

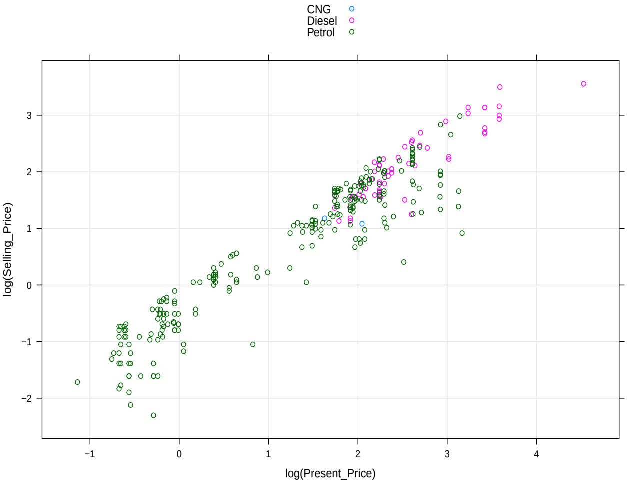
Selling_Price vs Driven_kms



- a car that is driven more, should get lower resale value
- here we get positive slope between log(Selling_Price) and log(Driven_kms)
- may be because there are hidden grouping variables

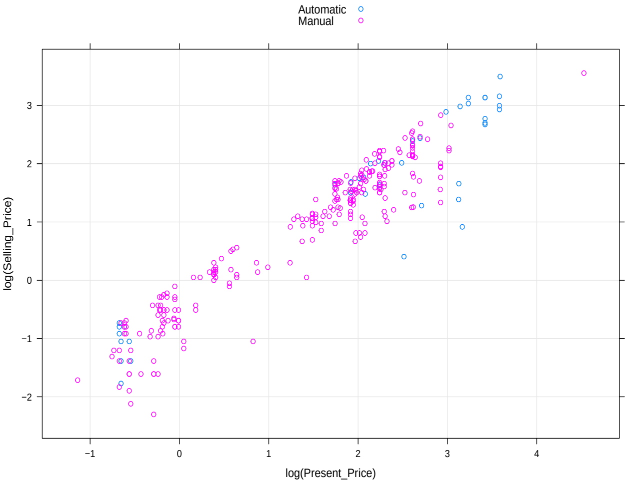

CNG
Diesel
Petrol
log(Selling_Price)
log(Driven_kms)
3
2
1
0
−1
−2
6
8
10
12

Selling_Price vs Fuel_Type



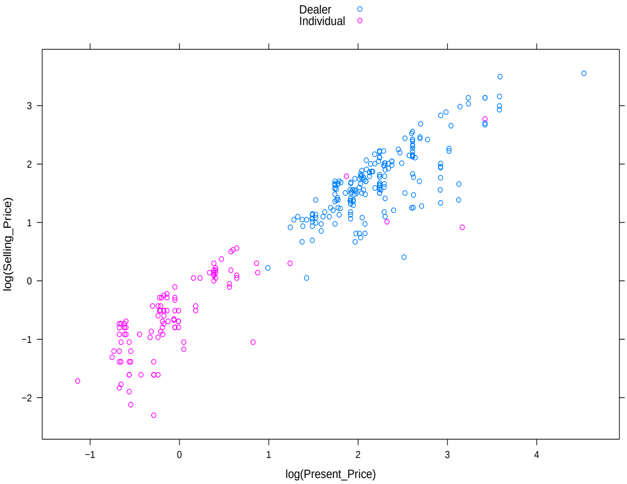
- Diesel cars get higher resale value than petrol

Selling_Price vs Transmission



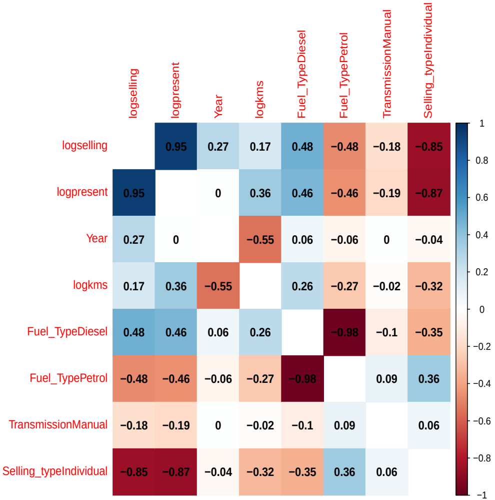
- cars with automatic transmission get higher value

Selling_Price vs Selling_Type



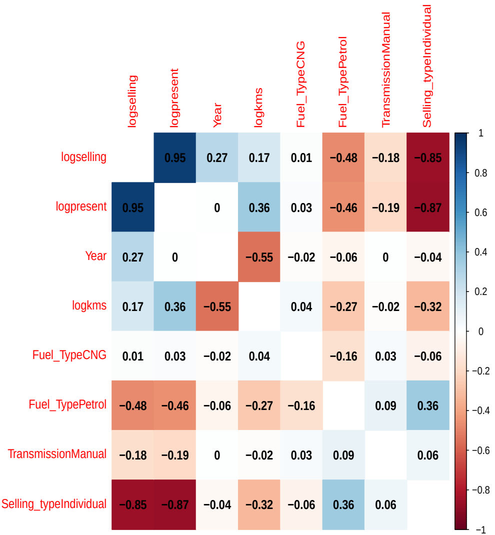
- cars sold through dealer , get higher price

Correlation Structure

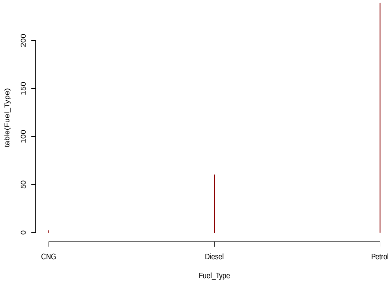


- notice the correlation between two fuel types

after altering the choice of Fuel_Type



why such behavior in correlation matrix?



REGRESSION

REGRESSION

starting with the Full Model

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \varepsilon$$

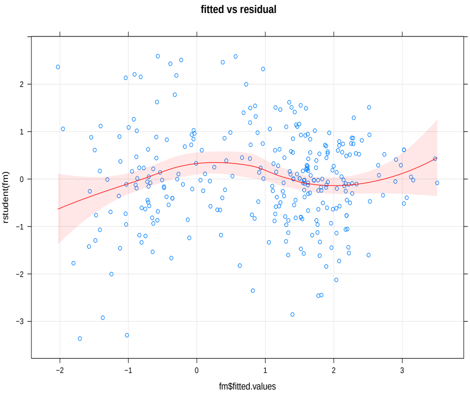
- Y = log(Selling_Price)
- X_1= log(Present_Price)
- X_2= Year
- X_3= log(Driven_kms)
- X_4= Fuel_TypeCNG
- X_5= Fuel_TypePetrol
- X_6= TransmissionManual
- X_7= Selling_typeIndividual

OLS

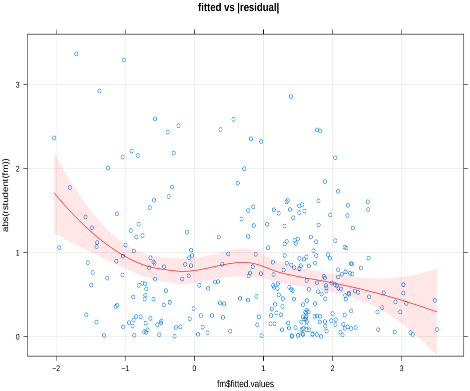
summary of OLS fit

```
##
## Call:
## lm(formula = logselling ~ x1 + x2 + x3 + x4 + x5 + x6 + x7)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59411 -0.10607 -0.00375  0.11020  0.46426
##
## 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
##
## Residual standard error: 0.1829 on 293 degrees of freedom
## Multiple R-squared:  0.9798, Adjusted R-squared:  0.9793
## F-statistic:  2030 on 7 and 293 DF,  p-value: < 2.2e-16
```

Residual Plots

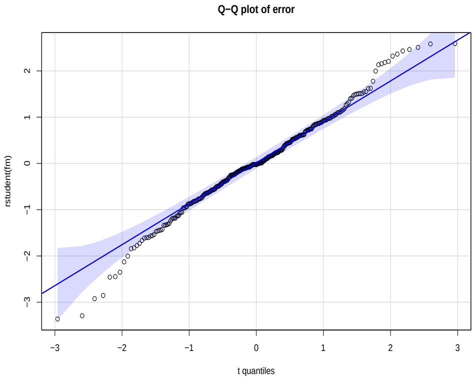


Residual Plots

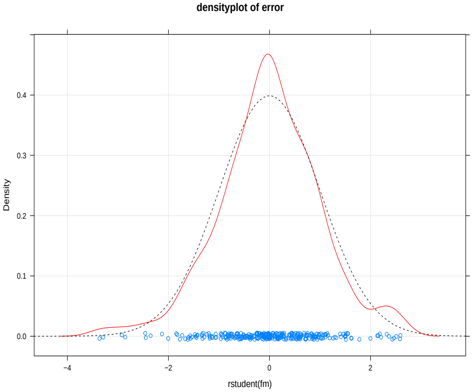


- non constant error variance

Residual Plots

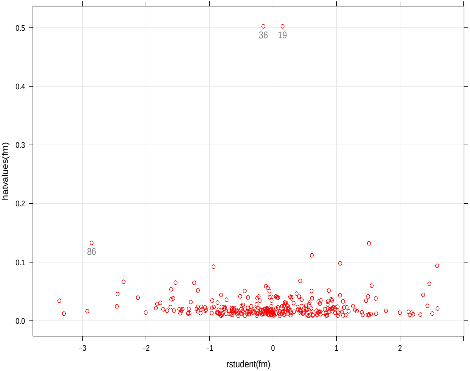


Residual Plots

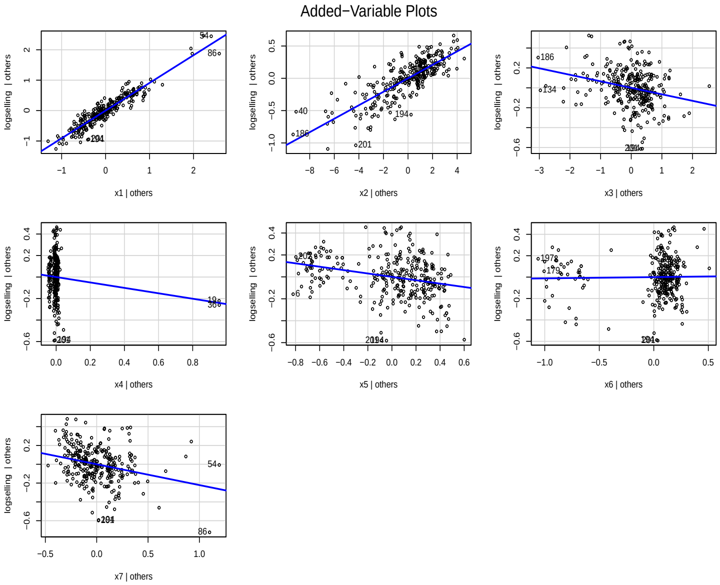


- not much deviation from normality

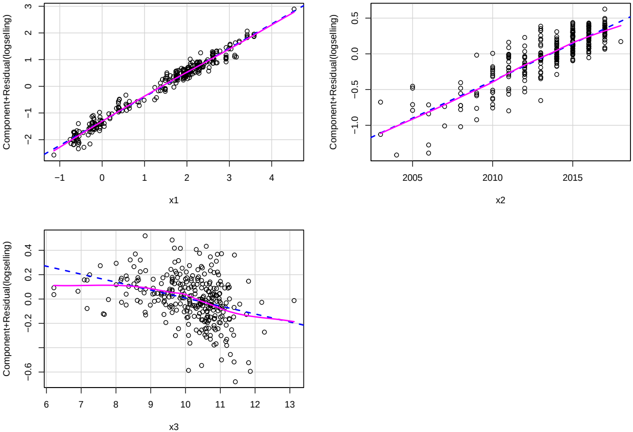
Unusual Observation



Added Variable Plots



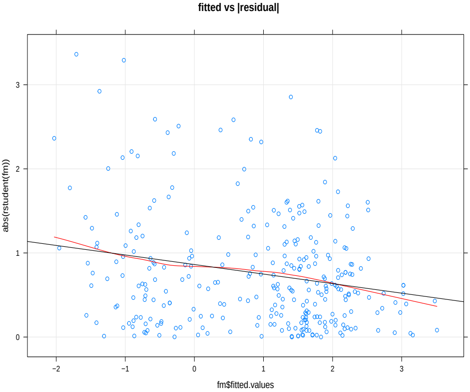
Component + Residual Plots



- linearity assumption holds

WLS : correcrtion for heteroscedasticity

estimating weights

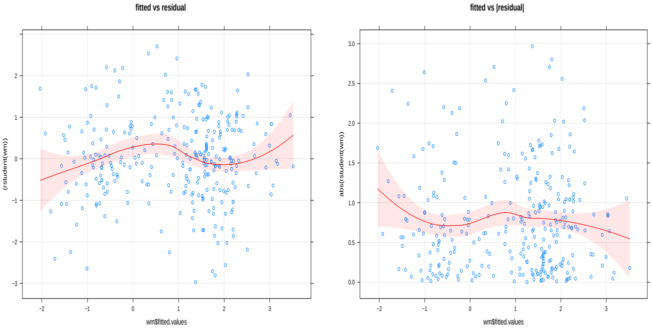


- estimate σ by least square line

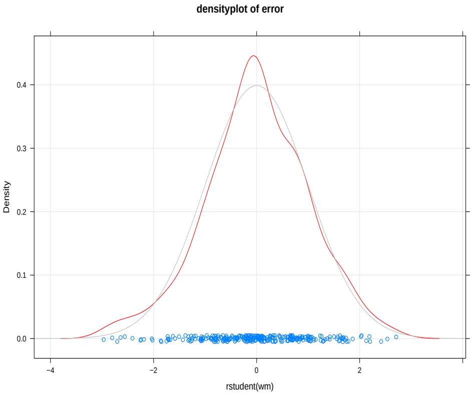
summary of WLS fit

```
##
## Call:
## lm(formula = logselling ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, weights = wt)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64605 -0.14314 -0.00272  0.15934  0.60673
##
## 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
## F-statistic:  1950 on 7 and 293 DF,  p-value: < 2.2e-16
```

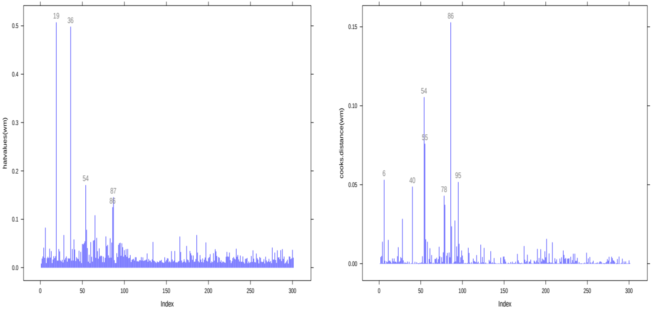
Residual Plots : heteroscedasticity rectified



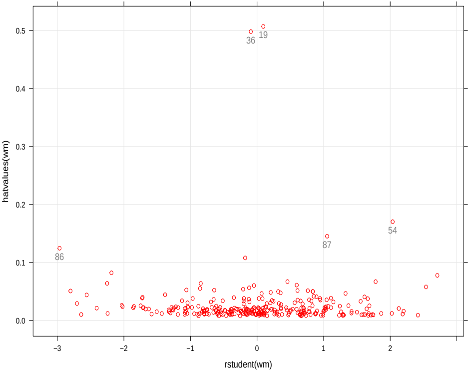
Density plot of errors



The most unusual observations



The most unusual observations



- observations 19th and 36th have extreme leverages, low residuals
- observation 86th have moderate leverage , high residual

The most unsual observations : possible explanation

```
Data[c(19,36,86),-8]
```

```
## # A tibble: 3 x 8
##   Car_Name  Year Selling_Price Present_Price Driven_kms Fuel_Type Selling_type
##   <chr>    <dbl>         <dbl>         <dbl>      <dbl> <chr>     <chr>
## 1 wagon r   2015          3.25          5.09      35500 CNG       Dealer
## 2 sx4       2011          2.95          7.74      49998 CNG       Dealer
## 3 camry     2006          2.5          23.7      142000 Petrol    Individual
## # ... with 1 more variable: Owner <dbl>
```

```
which(Fuel_Type=="CNG")
```

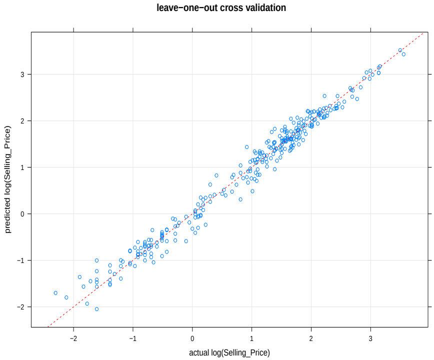
```
## [1] 19 36
```

```
Data[which(Fuel_Type=="Petrol" & Selling_type=="Individual" & Transmission=="Automatic"),1:5]
```

```
## # A tibble: 10 x 5
##    Car_Name           Year Selling_Price Present_Price Driven_kms
##    <chr>             <dbl>         <dbl>         <dbl>      <dbl>
##  1 camry              2006          2.5          23.7      142000
##  2 Honda Activa 4G    2017          0.48          0.51       4300
##  3 Honda Activa 4G    2017          0.45          0.51       4000
##  4 Activa 3g          2016          0.45          0.54        500
##  5 Activa 4g          2017          0.4           0.51       1300
##  6 Honda Activa 125   2016          0.35          0.57      24000
##  7 TVS Jupyter        2014          0.35          0.52      19000
##  8 Suzuki Access 125  2008          0.25          0.58       1900
##  9 TVS Wego           2010          0.25          0.52      22000
## 10 Activa 3g          2008          0.17          0.52     500000
```

R^2_{pred}

```
## [1] 0.9786312
```

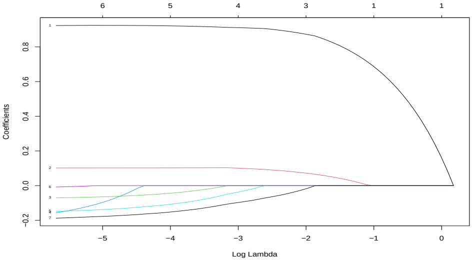


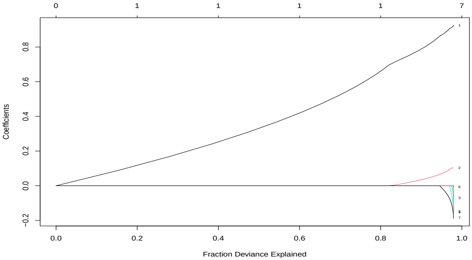
Finding sparser model, if any

- split the data in Train_Set:Test_Set = 80:20 for further calculations

LASSO

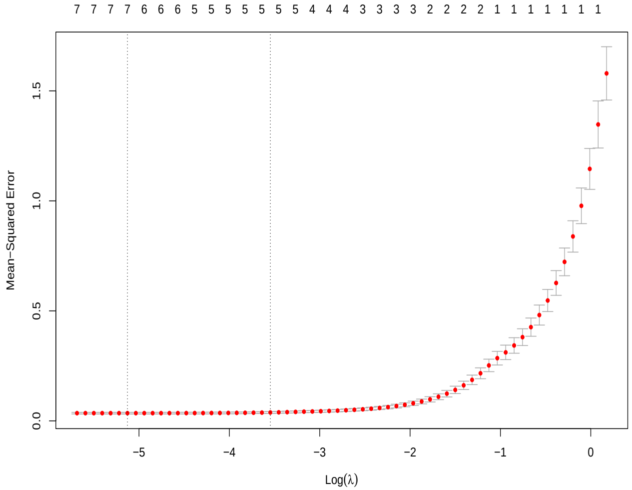
LASSO for various penalty parameter





- as penalty increases , more coefficients are estimated as zero , at a cost of decrease in explained variability
- we take the max possible penalty (i.e. max sparsity) ,for which MSE is within 1 standard error of the minimum MSE (i.e. best fitting)

optimum penalty



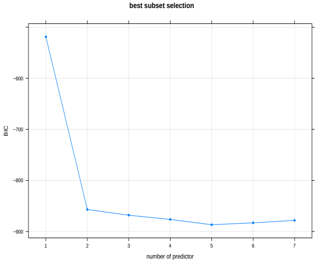
optimum LASSO model

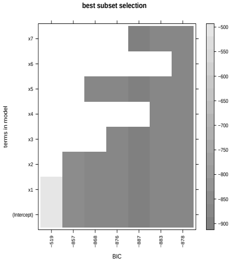
```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                         s0
## (Intercept) -206.77448506
## x1             0.91733871
## x2             0.10266770
## x3            -0.02473837
## x4             .
## x5            -0.08197217
## x6             .
## x7            -0.13209400
```

- so selected predictors are X_1, X_2, X_3, X_5, X_7

Best Subset Selection

Best Subset Selection





- so selected predictors are X_1, X_2, X_3, X_5, X_7

New Model with selected predictors

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

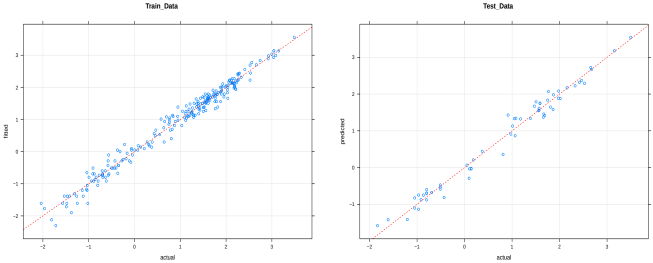
summary of fit

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x5 + x7, data = training_dataset,
##     weights = wt2)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65902 -0.13191  0.00052  0.15932  0.58104
##
## 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
##
## Residual standard error: 0.2344 on 234 degrees of freedom
## Multiple R-squared:  0.9783, Adjusted R-squared:  0.9779
## F-statistic:  2113 on 5 and 234 DF,  p-value: < 2.2e-16
```

```
MAPE
```

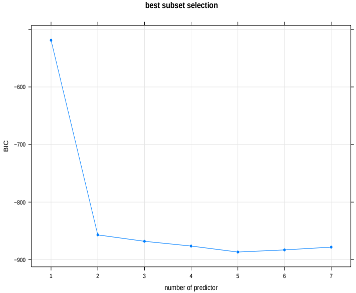
```
## [1] 25.68392
```

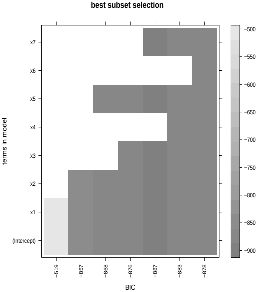
performance on Train Set and Test Set



Can we reduce further ?

notice this plot again





- there is not much increase in BIC , when no. of predictors dropped to two from five
- so we can try the model with the best subset of size two , X_1 & X_2

Reduced Model

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

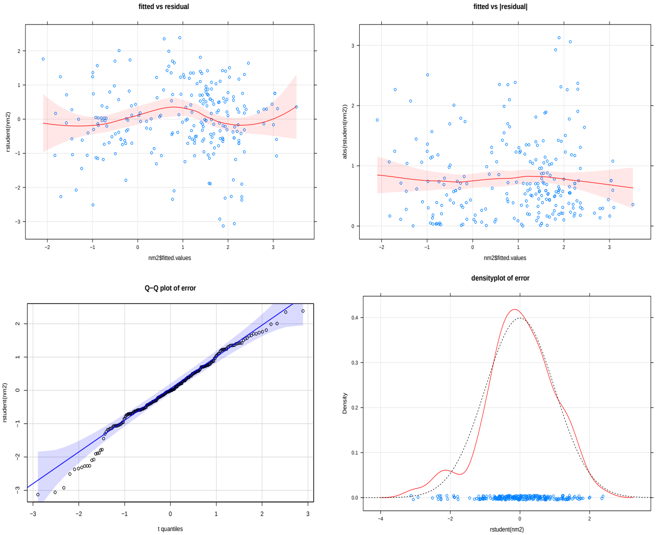
summary of fit

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = training_dataset, weights = wt2)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7737 -0.1498 -0.0019  0.1790  0.6043
##
## 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
##
## Residual standard error: 0.2565 on 237 degrees of freedom
## Multiple R-squared:  0.9737, Adjusted R-squared:  0.9735
## F-statistic:  4391 on 2 and 237 DF,  p-value: < 2.2e-16
```

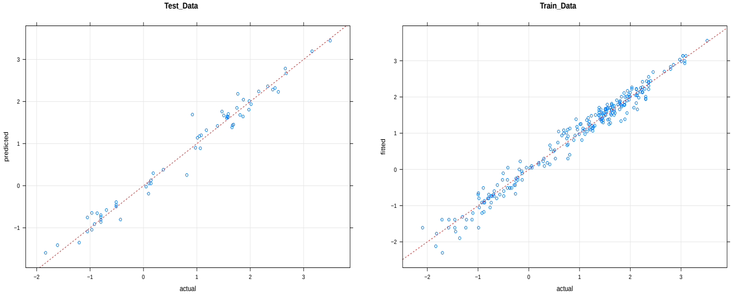
```
MAPE
```

```
## [1] 22.33453
```

Residual Plots : no severe violation of assumptions



performance



- no apparent difference from the earlier ones

- The ‘Principle of Parsimony’ suggests we should use this simpler model
- By dropping number of predictors MAPE value has decreased. Earlier models were overfitted

95% prediction interval

A 95% prediction interval of Y, for new data $X_1=x_{01}, X_2=x_{02}, \dots, X_7=x_{07}$ 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 $x_0' = (1, x_{01}, x_{02}, \dots, x_{07})$

$\widehat{y_0} = -245.687 + 0.982\, x_{01} + 0.122\, x_{02}$ → 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

Conclusion

It is enough to collect information about current price of a same car model and how old the used car is , to make reasonable prediction about its resale value

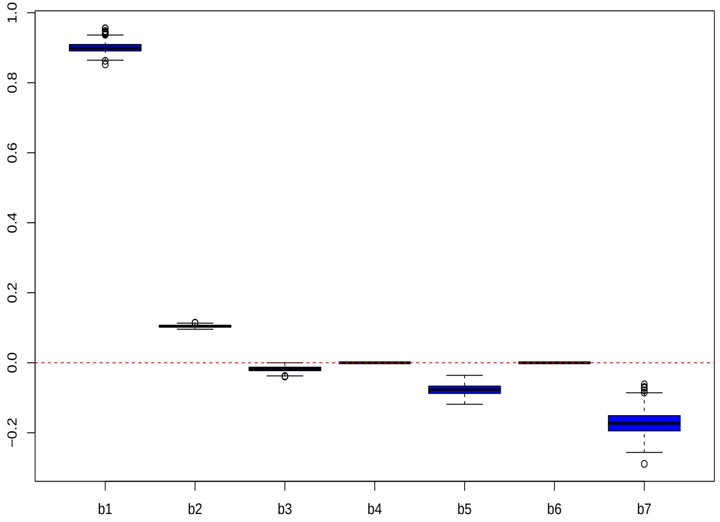
SUMMARY

SUMMARY	$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_7 X_7 + \varepsilon$	$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_5 X_5 + \beta_7 X_7 + \varepsilon$	$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
$\hat{\alpha}$	-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

Appendix

simulation for LASSO coefficients

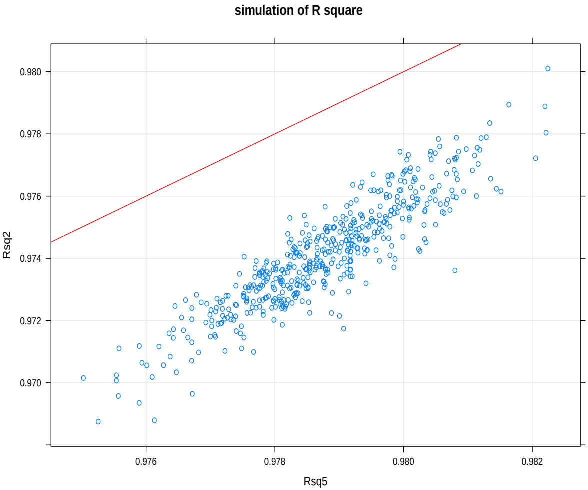

1.0
0.8
0.6
0.4
0.2
0.0
−0.2
b1
b2
b3
b4
b5
b6
b7

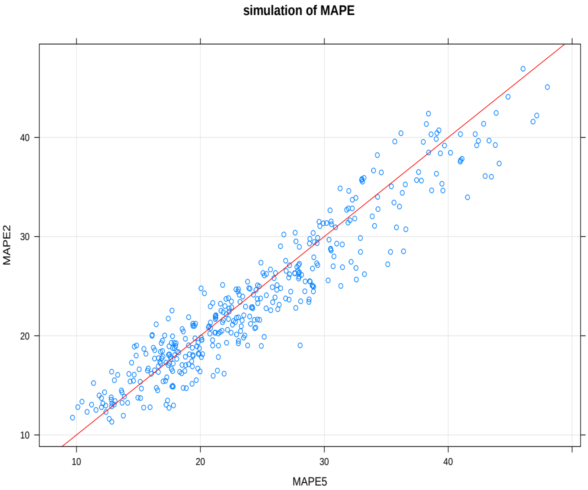
comparison

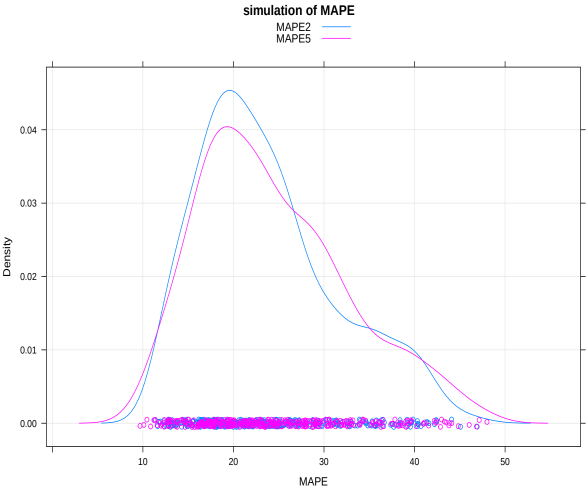
$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_5 X_5 + \beta_7 X_7 + \varepsilon$$

vs

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$


simulation of R square
Rsq2
Rsq5
0.970
0.972
0.974
0.976
0.978
0.980
0.976
0.978
0.980
0.982


simulation of MAPE
MAPE2
MAPE5
10
20
30
40
10
20
30
40


simulation of MAPE
MAPE2
MAPE5
Density
0.00
0.01
0.02
0.03
0.04
10
20
30
40
50
MAPE

-THANK YOU-