**PROJECT 2- Create a Multiple Linear Regression Model for General Motors(GM) Data set?**

library(readxl)

GM\_MOTOR <- read\_excel("C:/Users/Anjana/Desktop/GM\_MOTOR.xls")

View(GM\_MOTOR)

fit<-lm(Price~.,data=GM\_MOTOR)

> summary(fit)

Call:

lm(formula = Price ~ ., data = GM\_MOTOR)

Residuals:

Min 1Q Median 3Q Max

-8420.7 -1743.5 -150.6 1315.7 26563.5

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.612e+04 1.815e+03 14.392 < 2e-16 \*\*\*

Mileage -2.058e-01 1.857e-02 -11.084 < 2e-16 \*\*\*

MakeChevrolet -1.706e+04 7.247e+02 -23.538 < 2e-16 \*\*\*

MakePontiac -1.851e+04 7.005e+02 -26.423 < 2e-16 \*\*\*

Cylinder -2.220e+03 5.013e+02 -4.430 1.17e-05 \*\*\*

Liter 7.691e+03 5.693e+02 13.509 < 2e-16 \*\*\*

Cruise 1.024e+02 4.007e+02 0.256 0.798

Sound 2.279e+02 3.877e+02 0.588 0.557

Leather 2.472e+02 4.198e+02 0.589 0.556

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3430 on 491 degrees of freedom

Multiple R-squared: 0.8823, Adjusted R-squared: 0.8803

F-statistic: 459.9 on 8 and 491 DF, p-value: < 2.2e-16

**Output Interpretation** - From this output, we have determined that the intercept is 26120 and the coefficient of Milage is -0.2058, for MakeChevrolet -17060,For MakePontjac -18510,For Cylinder -2220,For Liter 7691, For Cruise 102.4, For Sound 227.9 and for Leather is 247.2.

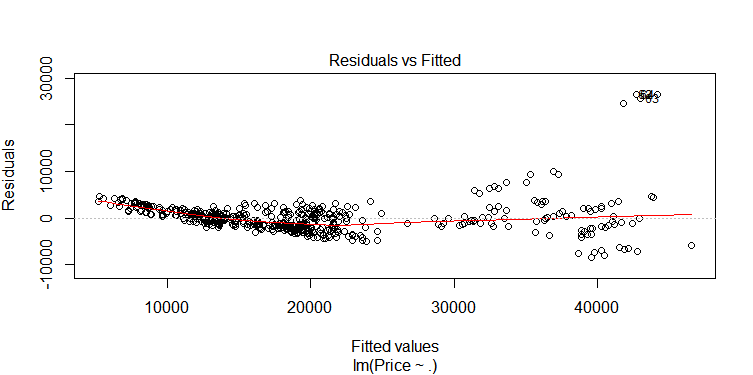
model<- 26120 + (-0.2058 \* Milage)+(-17060 \* MakeChevrolet)+(-18510 \* MakePontaic) +(-2220 \* Cylinder)+(7691\* Liter) +(102.4 \*Cruise) +(227.9 \* Sound) +(247.2 \* Leather)

Here for predicted price, the Milage, MakeChevrolet, MakePontiac, Cylinder and Liter are quite more significant variables.

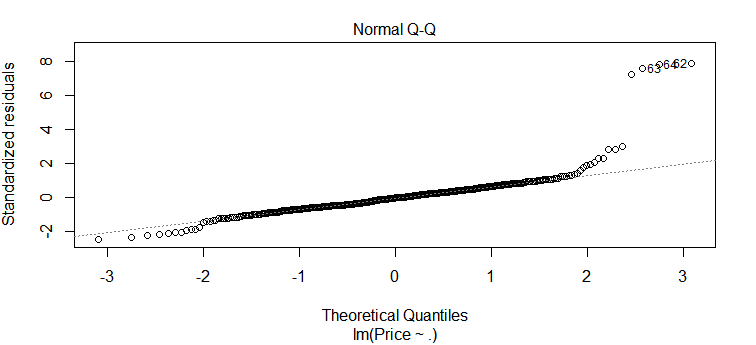
This equation tells us that the predicted retail price of used GM Motor will decrease by 0.2058 for every one percent increase in the Milage, will decrease by 17060 for every one percent increase in MakeChevrolet, will decrease by 18510 for every one percent increase in MakePontiac, will decrease by 2220 for every one percent increase in Cylinder and will increase by 7691 for every one percent increase in the Liter, will increase by 102.4 for every one percent increase in Cruise, will increase by 227.9 for every one percent increase in Sound and will increase by247.2 for every one percent increase in Leather.

The Adjusted R-squared value is 0.8803 which means that the model can explain about 88% of the variance of the Price variable.

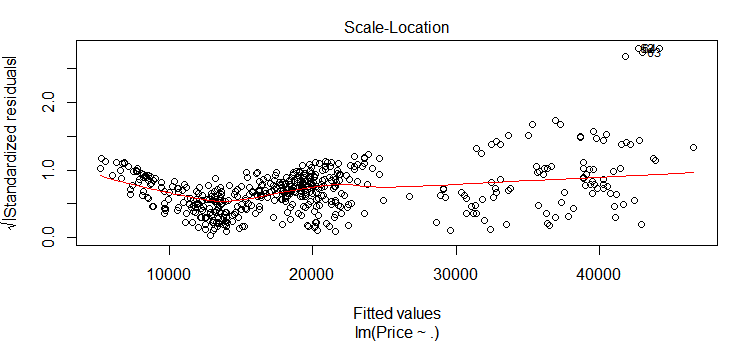
plot(fit)



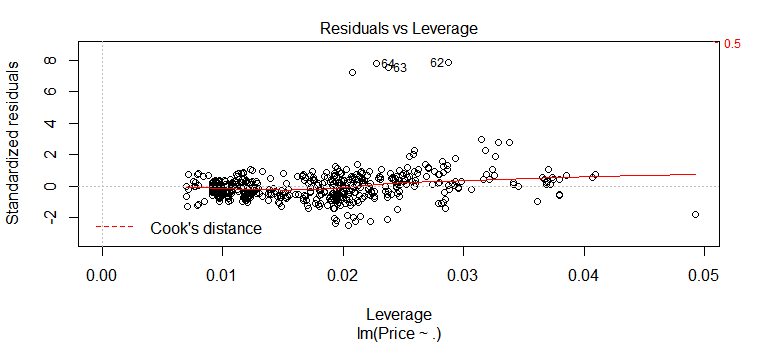
A plot of residuals versus fitted values - The residuals are the part of the dependent variable that the model couldn't explain. They are calculated by subtracting the predicted value from the actual value of the dependent variable. Under the usual assumptions for the linear regression model, we don't expect the variability of the residuals to change over the range of the dependent variable, so there shouldn't be any discernable pattern to this plot. Outliers in the plot are showing.



A normal quantile-quantile plot of the standardized residuals -We assume that the errors of the model follows a normal distribution. Thus, we expect a normal quantile-quantile plot of the residuals to follow a straight line. Deviations from a straight line could mean that the errors which do not follow a normal distribution. Here you can a few outliers.



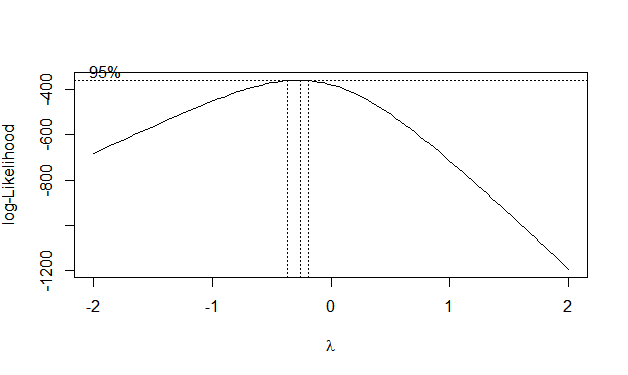
A scale-location plot - This plot is similar to the residuals versus fitted values plot, but it uses the square root of the standardized residuals.



Residuals Vs Leverage - Cook's distance tries to identify points which have more influence than other points. Generally these are points that are distant from other points in the data, either for the dependent variable or one or more independent variables. Each observation is represented as a line whose height is indicative of the value of Cook's distance for that observation.

Here we have to find out those influential cases whose impact is not much in regression model.

boxcox(fit,lambda=seq(-2,2,1/10),plotit=TRUE)



 Boxcox transformation tries to maximize a normal-distribution based likelihood function, which assumes constant variance. Here you can see at 95% inflaction point, the best exponent lies between 0 to -1 or -2.

> fit2<- lm(1/sqrt(Price)~.,data=GM\_MOTOR)

> summary(fit2)

Call:

lm(formula = 1/sqrt(Price) ~ ., data = GM\_MOTOR)

Residuals:

Min 1Q Median 3Q Max

-1.095e-03 -2.379e-04 -1.108e-05 2.401e-04 1.191e-03

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.742e-03 1.858e-04 41.666 < 2e-16 \*\*\*

Mileage 3.209e-08 1.901e-09 16.878 < 2e-16 \*\*\*

MakeChevrolet 2.020e-03 7.420e-05 27.230 < 2e-16 \*\*\*

MakePontiac 1.944e-03 7.173e-05 27.103 < 2e-16 \*\*\*

Cylinder 3.191e-04 5.133e-05 6.218 1.08e-09 \*\*\*

Liter -1.267e-03 5.829e-05 -21.733 < 2e-16 \*\*\*

Cruise -9.171e-05 4.102e-05 -2.236 0.02583 \*

Sound -1.132e-04 3.969e-05 -2.852 0.00452 \*\*

Leather -6.280e-05 4.299e-05 -1.461 0.14468

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0003512 on 491 degrees of freedom

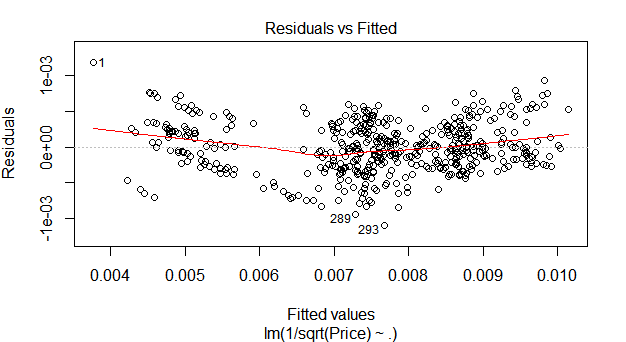
Multiple R-squared: 0.9426, Adjusted R-squared: 0.9417

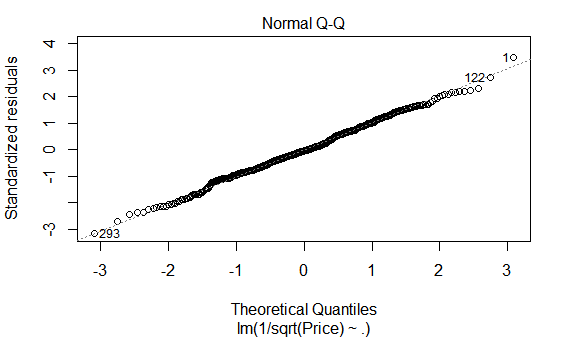
F-statistic: 1009 on 8 and 491 DF, p-value: < 2.2e-16

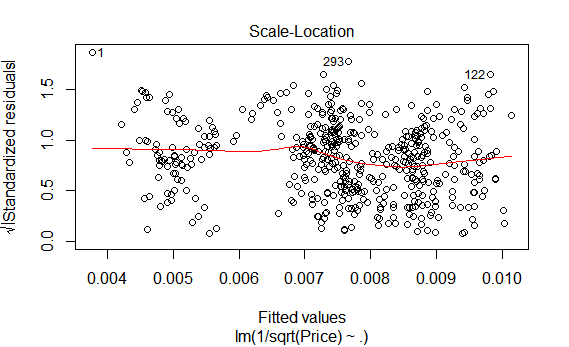
**Output Interpretation** - From this output, we have determined that the intercept is 0.007742.

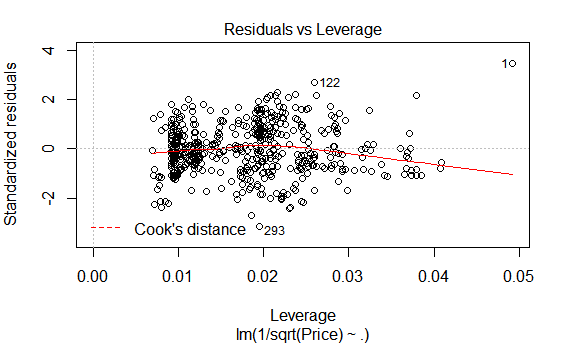
The Adjusted R-squared value is 0.9417 which means that the model can explain about 94% of the variance of the Price variable.

plot(fit2)









> fit3<- lm(log(Price)~.,data=GM\_MOTOR)

> summary(fit3)

Call:

lm(formula = log(Price) ~ ., data = GM\_MOTOR)

Residuals:

Min 1Q Median 3Q Max

-0.30978 -0.06502 -0.00189 0.06477 0.41267

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 9.862e+00 5.092e-02 193.666 < 2e-16 \*\*\*

Mileage -8.881e-06 5.210e-07 -17.045 < 2e-16 \*\*\*

MakeChevrolet -6.346e-01 2.033e-02 -31.205 < 2e-16 \*\*\*

MakePontiac -6.422e-01 1.966e-02 -32.671 < 2e-16 \*\*\*

Cylinder -9.199e-02 1.407e-02 -6.540 1.55e-10 \*\*\*

Liter 3.525e-01 1.598e-02 22.062 < 2e-16 \*\*\*

Cruise 1.933e-02 1.124e-02 1.719 0.0863 .

Sound 1.999e-02 1.088e-02 1.838 0.0667 .

Leather 1.436e-02 1.178e-02 1.219 0.2235

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09626 on 491 degrees of freedom

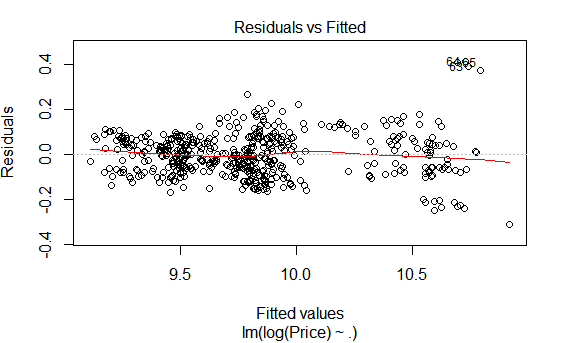
Multiple R-squared: 0.9471, Adjusted R-squared: 0.9462

F-statistic: 1098 on 8 and 491 DF, p-value: < 2.2e-16

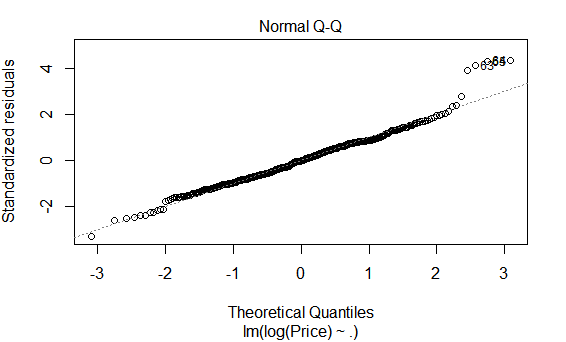
**Output Interpretation** - From this output, we have determined that the intercept is 9.862.

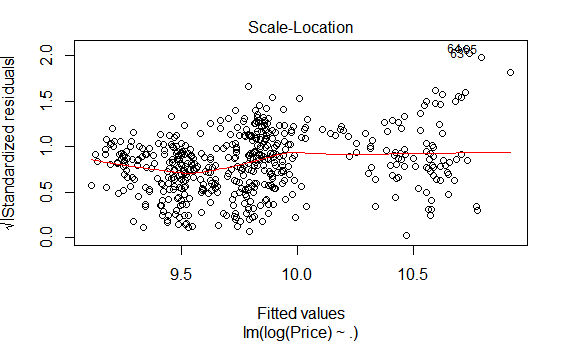
The Adjusted R-squared value is 0.9462 which means that the model can explain about 94.5% of the variance of the Price variable.

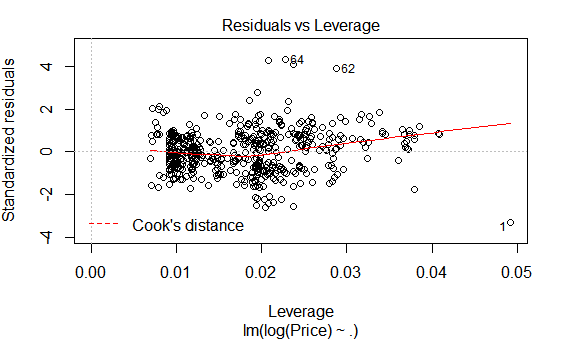
plot(fit3)



Here quite a straight line is shown.







**Comparison between “1/sqrt(Price)” and “log(Price) Model (fit2 Vs fit3)**

**Model with “1/sqrt(Price)” ie fit2 Model:**

1. Residuals Vs fitted Plot is not random, there are some curvature.

2. Normal QQ Plot is good, assumption is not violated.

**Model with “log(Price)” ie fit3 model:**

1. Residuals Vs fitted Plot is better than the fit2 model,shows some randomness.

2. Normal QQ Plot is OK.

From the above observations, we can opt for log(Price) Model or fit3 model.

**Checking the collinearity**

vif(fit)

Mileage MakeChevrolet MakePontiac Cylinder Liter

1.0115 5.1407 3.6927 21.8500 18.4440

Cruise Sound Leather

1.3929 1.2036 1.1893

Here it shows that Cylinder and Liter have collinearity relationship.

**Selection of Variables**

AIC is a goodness of fit measure that favours smaller residual error in the model, but penalises for including further predictors and helps avoiding overfitting.

stepAIC(fit3,direction="both")

Start: AIC=-2331.76

log(Price) ~ Mileage + Make + Cylinder + Liter + Cruise + Sound +

Leather

Df Sum of Sq RSS AIC

- Leather 1 0.0138 4.5636 -2332.2

<none> 4.5498 -2331.8

- Cruise 1 0.0274 4.5772 -2330.8

- Sound 1 0.0313 4.5811 -2330.3

- Cylinder 1 0.3963 4.9462 -2292.0

- Mileage 1 2.6922 7.2421 -2101.3

- Liter 1 4.5102 9.0600 -1989.4

- Make 2 10.3639 14.9138 -1742.2

Step: AIC=-2332.25

log(Price) ~ Mileage + Make + Cylinder + Liter + Cruise + Sound

Df Sum of Sq RSS AIC

<none> 4.5636 -2332.2

- Cruise 1 0.0217 4.5853 -2331.9

+ Leather 1 0.0138 4.5498 -2331.8

- Sound 1 0.0443 4.6079 -2329.4

- Cylinder 1 0.4189 4.9825 -2290.3

- Mileage 1 2.6929 7.2565 -2102.4

- Liter 1 4.6827 9.2464 -1981.2

- Make 2 11.3889 15.9525 -1710.5

Call:

lm(formula = log(Price) ~ Mileage + Make + Cylinder + Liter +

Cruise + Sound, data = GM\_MOTOR)

Coefficients:

(Intercept) Mileage MakeChevrolet MakePontiac Cylinder

9.880e+00 -8.882e-06 -6.401e-01 -6.489e-01 -9.396e-02

Liter Cruise Sound

3.553e-01 1.695e-02 2.311e-02

According to this procedure, the best model is the one that includes variable Cruise, Sound, Cylinder, Mileage, Liter and Make.

model<- 9.880+ (-0.000008882 \* Milage)+(-0.6401 \* MakeChevrolet)+(-0.6489 \* MakePontaic) +(-0.0939\* Cylinder)+(0.3553\* Liter) +(0.01695 \* Cruise) +(0.02311 \* Sound).