> summary(FE2010)

EngDispl NumCyl FE NumGears TransLockup TransCreeperGear

Min. :1.000 Min. : 2.000 Min. :17.50 Min. :1.000 Min. :0.0000 Min. :0.00000

1st Qu.:2.400 1st Qu.: 4.000 1st Qu.:29.09 1st Qu.:5.000 1st Qu.:0.0000 1st Qu.:0.00000

Median :3.500 Median : 6.000 Median :34.51 Median :6.000 Median :1.0000 Median :0.00000

Mean :3.507 Mean : 5.971 Mean :34.71 Mean :5.268 Mean :0.6802 Mean :0.04878

3rd Qu.:4.300 3rd Qu.: 8.000 3rd Qu.:39.20 3rd Qu.:6.000 3rd Qu.:1.0000 3rd Qu.:0.00000

Max. :8.400 Max. :16.000 Max. :69.64 Max. :8.000 Max. :1.0000 Max. :1.00000

IntakeValvePerCyl ExhaustValvesPerCyl VarValveTiming VarValveLift

Min. :0.000 Min. :0.000 Min. :0.0000 Min. :0.0000

1st Qu.:2.000 1st Qu.:2.000 1st Qu.:1.0000 1st Qu.:0.0000

Median :2.000 Median :2.000 Median :1.0000 Median :0.0000

Mean :1.862 Mean :1.837 Mean :0.8229 Mean :0.1671

3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:1.0000 3rd Qu.:0.0000

Max. :3.000 Max. :2.000 Max. :1.0000 Max. :1.0000

> str(FE2010)

'data.frame': 1107 obs. of 10 variables:

$ EngDispl : num 4.7 4.7 4.2 4.2 5.2 5.2 2 6 3 3 ...

$ NumCyl : int 8 8 8 8 10 10 4 12 6 6 ...

$ FE : num 28 25.6 26.8 25 24.8 ...

$ NumGears : int 6 6 6 6 6 6 6 6 6 6 ...

$ TransLockup : int 1 1 1 1 0 0 0 0 1 0 ...

$ TransCreeperGear : int 0 0 0 0 0 0 0 0 0 0 ...

$ IntakeValvePerCyl : int 2 2 2 2 2 2 2 2 2 2 ...

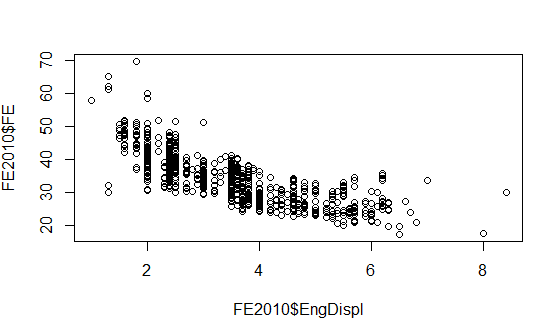
$ ExhaustValvesPerCyl: int 2 2 2 2 2 2 2 2 2 2 ...

$ VarValveTiming : int 1 1 1 1 1 1 1 1 1 1 ...

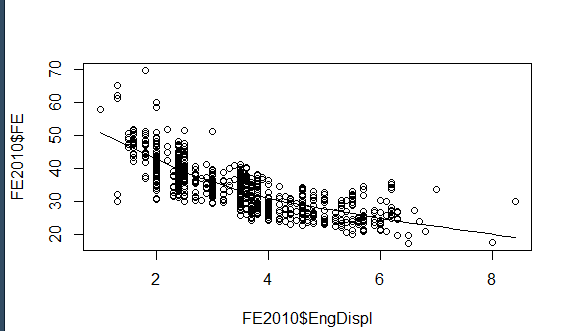
$ VarValveLift : int 0 0 0 0 0 0 0 0 1 1 ...

> library(ggplot2)

> plot(FE2010$EngDispl,FE2010$FE)



> scatter.smooth(x=FE2010$EngDispl,y=FE2010$FE, Main="EngDisp ~ FE")



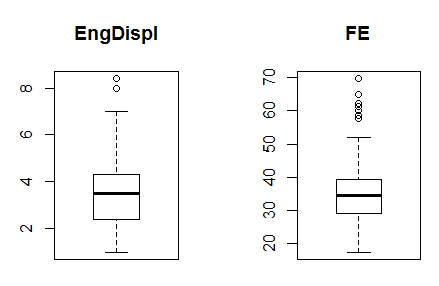
**In the above we can see there is an negative correlation between Eng Disp and FE.**

# using the Box plot to check for outliers

par(mfrow=c(1,2))# divid graph into 2 columns

boxplot(FE2010$EngDispl, main="EngDispl") #boxplot for Eng disp

boxplot(FE2010$FE, main="FE") #Boxplot for FE



# for skewness function: Density plot: To see the distribution of the predictor variable.

#Ideally, a close to normal distribution (a bell shaped curve),without being skewed to the left or right is preferred.

library(e1071)

par(mfrow=c(1, 2)) # divide graph area in 2 columns

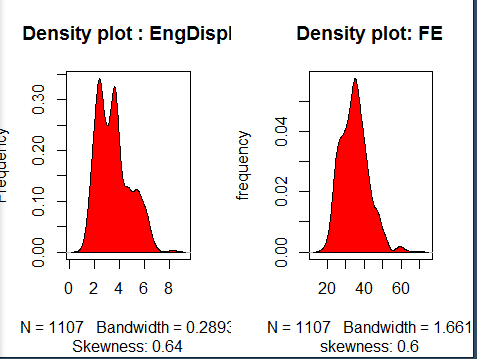
plot(density(FE2010$EngDispl), main="Density plot : EngDispl", ylab= "Frequency",

sub=paste("Skewness:", round(e1071::skewness(FE2010$EngDispl), 2)))

polygon(density(FE2010$EngDispl),col = "red")

plot(density(FE2010$FE), main= "Density plot: FE", ylab="frequency", sub=paste("skewness:", round(e1071::skewness(FE2010$FE), 2)))

polygon(density(FE2010$FE), col = "red")



#to check the correlation between EngDispl and EF

cor(FE2010$EngDispl,FE2010$FE)



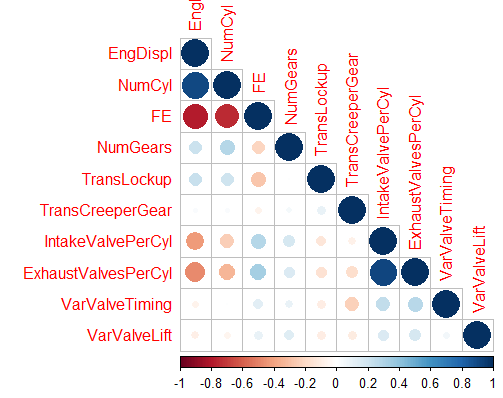
# Cor plot to visualize the relation

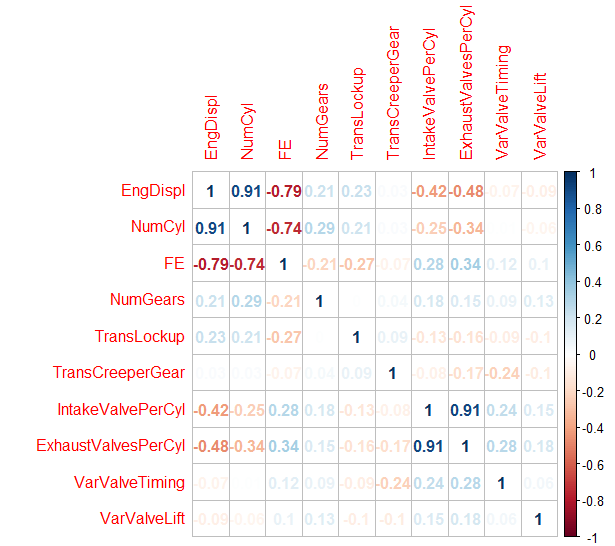
library(corrplot)

cr<-cor(FE2010)

corrplot(cr,type = "lower")

corrplot(cr,method = "number")





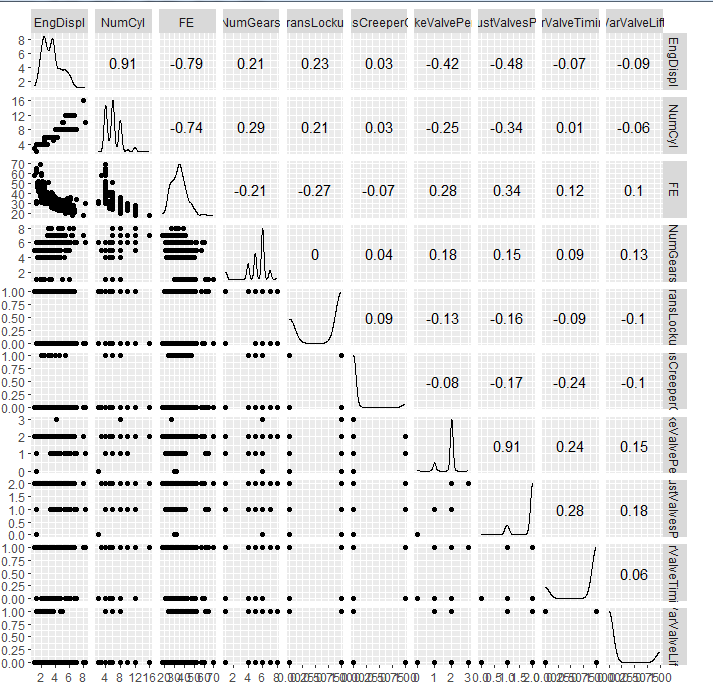
**In the above graph we can see only only EngDispl and Num cyl are negatively correlated with FE.**

install.packages("GGally")

library(GGally)

library(dplyr)

ggscatmat(FE2010,columns = 1:ncol(FE2010))

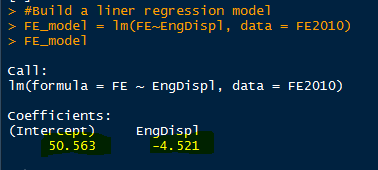


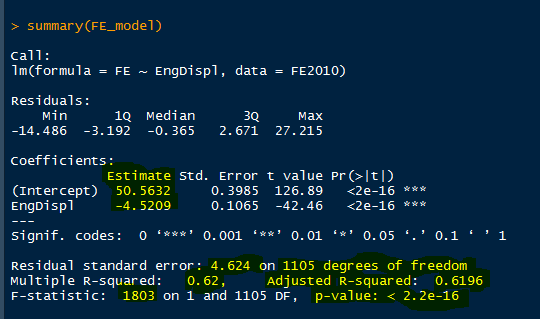
**From the graph we can say only EngDispl is the best predictor**

#Build a liner regression model

FE\_model = lm(FE~EngDispl, data = FE2010)

FE\_model

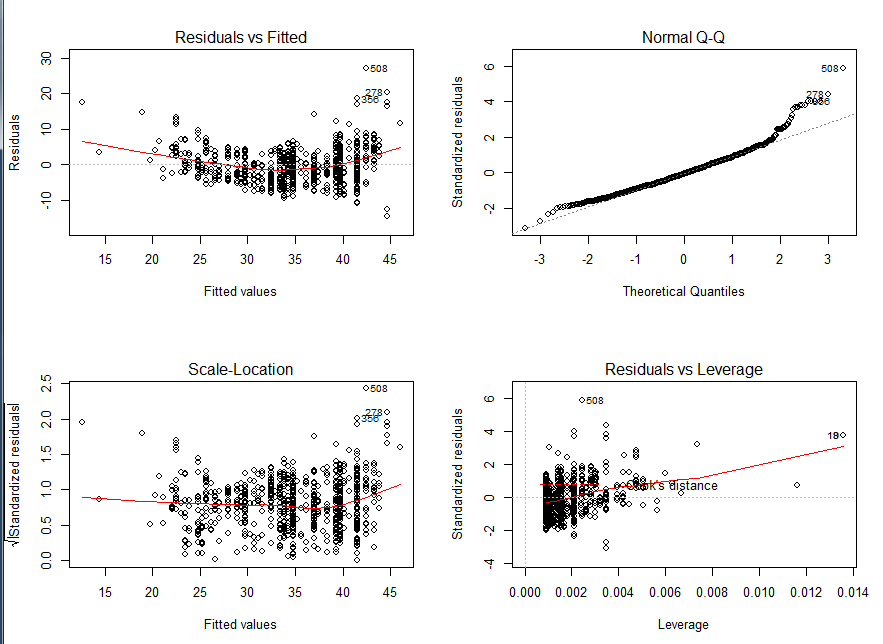




1. As we have negative co-relation we can say I unit of change in EngDispl will result in -4.5209 times change in FE.
2. As multiple R-Squared is 0.62, we can say only 62% of varience can be explained with EngDispl.
3. We cannot accept the null hypothesis as p-value is less than 0.05 and model is statically significant.
4. F-statistic

par(mfrow=c(2,2))

plot(FE\_model)



Normal Q-Q plot shows it has a outliers

> # Build Liner regression with all the varibles

> FE\_all\_mod<-lm(FE~., data = FE2010)

> FE\_all\_mod

Call:

lm(formula = FE ~ ., data = FE2010)

Coefficients:

(Intercept) EngDispl NumCyl NumGears

54.3472 -3.8610 -0.4888 -0.1725

TransLockup TransCreeperGear IntakeValvePerCyl ExhaustValvesPerCyl

-1.4450 -0.9138 -0.3737 -1.1105

VarValveTiming VarValveLift

1.6870 0.6235

> summary(FE\_all\_mod)

Call:

lm(formula = FE ~ ., data = FE2010)

Residuals:

Min 1Q Median 3Q Max

-17.1153 -2.7142 -0.3535 2.4191 25.6521

Coefficients:

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

(Intercept) 54.3472 1.0973 49.530 < 2e-16 \*\*\*

EngDispl -3.8610 0.2805 -13.765 < 2e-16 \*\*\*

NumCyl -0.4888 0.1845 -2.649 0.00819 \*\*

NumGears -0.1725 0.1065 -1.620 0.10555

TransLockup -1.4450 0.3000 -4.817 1.66e-06 \*\*\*

TransCreeperGear -0.9138 0.6681 -1.368 0.17167

IntakeValvePerCyl -0.3737 0.9892 -0.378 0.70566

ExhaustValvesPerCyl -1.1105 0.9598 -1.157 0.24752

VarValveTiming 1.6870 0.3796 4.444 9.71e-06 \*\*\*

VarValveLift 0.6235 0.3719 1.676 0.09393 .

---

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

Residual standard error: 4.489 on 1097 degrees of freedom

Multiple R-squared: 0.6445, Adjusted R-squared: 0.6415

F-statistic: 220.9 on 9 and 1097 DF, p-value: < 2.2e-16

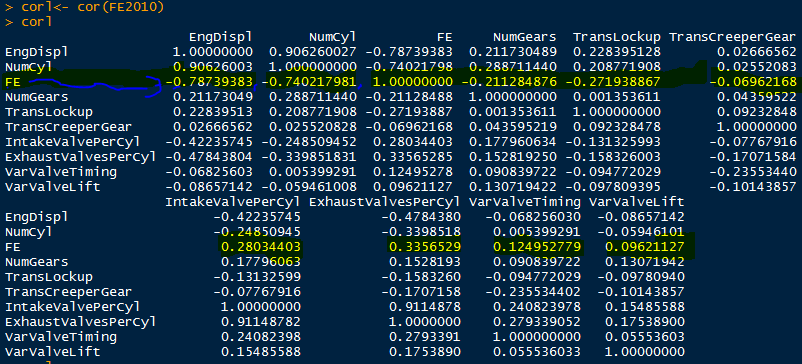
From the above we can say EngDispl, Numcyl , TransLockup and varvalveLift are the good predictors

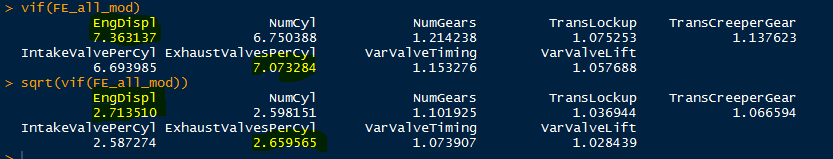
As P-value is less than 0.05, we can reject the null hypothesis and say Model is statically significant.

> # To check multi-collinarity Test

> library(car)

Loading required package: carData



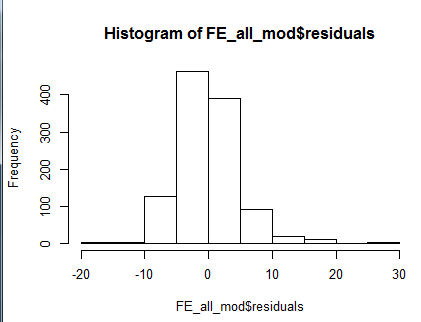


In the above correlation test we can see EngDispl has highest correlation against the predictor FE.

Though the Exhaustvalvespercyl is above 2.6 in VIF test , but scored less in the correlation test.

> #Normality test, Histogram of residuals

> hist(FE\_all\_mod$residuals)

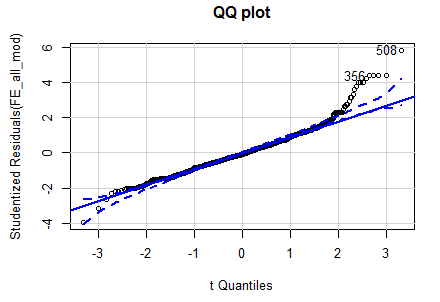


We can see it’s not perfectly bell shaped but near to the normally distributed.

we remove the outliers .

> qqPlot(FE\_all\_mod, main = "QQ plot")

[1] 356 508



> #Auto corelation test

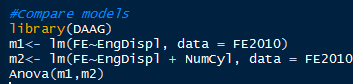
> durbinWatsonTest(FE\_all\_mod)

lag Autocorrelation D-W Statistic p-value

1 0.5404517 0.9185349 0

Alternative hypothesis: rho != 0

P value should be >0.05, as in the above we can see P-value is 0 which is <0.05, so we cannot accept the null hypothesis.



|  |
| --- |
| > Anova(m1,m2)  Anova Table (Type II tests)  Response: FE  Sum Sq Df F value Pr(>F)  EngDispl 38551 1 1820.2 < 2.2e-16 \*\*\*  Residuals 23382 1104  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
|  |
| |  | | --- | | > | |

As p value is less than 0.05 we can reject the null hypothesis and say the mean of both the model are different.

#Train and Test data

> library(caTools)

> set.seed(25)

> sample<-sample.split(FE2010,SplitRatio = 0.70)

> train1<-subset(FE2010,sample=TRUE)

> test<-subset(FE2010,sample=FALSE)

> train\_mod<-lm(FE~EngDispl,data = train1)

> pred\_FE2010<-predict(train\_mod,test)

> head(pred\_FE2010)

1 2 3 4 5 6

29.23577 29.23577 31.52359 31.52359 26.94796 26.94796

> summary(train\_mod)

Call:

lm(formula = FE ~ EngDispl, data = train1)

Residuals:

Min 1Q Median 3Q Max

-14.5929 -3.1719 -0.4117 2.6794 27.1353

Coefficients:

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

(Intercept) 50.7413 0.3979 127.53 <2e-16 \*\*\*

EngDispl -4.5756 0.1065 -42.97 <2e-16 \*\*\*

---

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

Residual standard error: 4.59 on 1105 degrees of freedom

Multiple R-squared: 0.6256, Adjusted R-squared: 0.6252

F-statistic: 1846 on 1 and 1105 DF, p-value: < 2.2e-16

> final<-cbind(test$FE,pred\_FE2010)

> actual\_pred<-data.frame(cbind(actuals=test$FE, Predictions=pred\_FE2010)) # make actuals\_predicteds dataframe.

> head(actual\_pred)

actuals Predictions

1 28.0198 29.23577

2 25.6094 29.23577

3 26.8000 31.52359

4 25.0451 31.52359

5 24.8000 26.94796

6 23.9000 26.94796

> # MAPE Calculation

> mape <- mean(abs((actual\_pred$Predictions - actual\_pred$actuals))/actual\_pred$actuals)

> mape

[1] 0.1033881

> Model\_rate = 100-mape

> Model\_rate

[1] 99.89661

From the model summary, the model p value and predictor’s p value are less than the significance level.

So you have a statistically significant model

> #==========================

> pred<-predict(FE\_model)

> F2010<-FE2010

> F2010$Predicted<-NA

> F2010$Predicted<-pred

> F2010$error<-FE\_model$residuals

> FE2011<-read.csv("FE2011.csv")

> pred2011<-predict(FE\_model,FE2011)

> head(pred2011)

1 2 3 4 5 6

23.74501 31.52359 31.52359 26.94796 26.94796 37.01435

> result<-cbind(FE2011$FE,pred2011)

> actual\_pred2011<-data.frame(cbind(actual=FE2011$FE,predictions=pred2011))

> head(actual\_pred2011)

actual predictions

1 22.9258 23.74501

2 26.7678 31.52359

3 24.3010 31.52359

4 24.3325 26.94796

5 23.0667 26.94796

6 32.8579 37.01435