Regression Models Course Project

Supposing I work for *Motor Trend*, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

- **"Is an automatic or manual transmission better for MPG"**
- **"Quantify the MPG difference between automatic and manual transmissions"**

1. Loading Data

We load the dataset

```
data(mtcars)
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Motor Trend Car Road Tests

Description

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

Format

A data frame with 32 observations on 11 (numeric) variables.

[, 1]	mpg	Miles/(US) gallon
[, 2]	cyl	Number of cylinders
[, 3]	disp	Displacement (cu.in.)
[, 4]	hp	Gross horsepower
[, 5]	drat	Rear axle ratio
[, 6]	wt	Weight (1000 lbs)
[, 7]	qsec	1/4 mile time
[, 8]	vs	Engine (0 = V-shaped, 1 = straight)
[, 9]	am	Transmission (0 = automatic, 1 = manual)
[,10]	gear	Number of forward gears
[,11]	carb	Number of carburetors

```
## Loading required package: carData
```

```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
```

```
##   +.gg   ggplot2
```

For convenience we can convert the variable “am” to a factor and add a more clear classification “Automatic” & “Manual”. and we can perform a friefly analysis of both variables

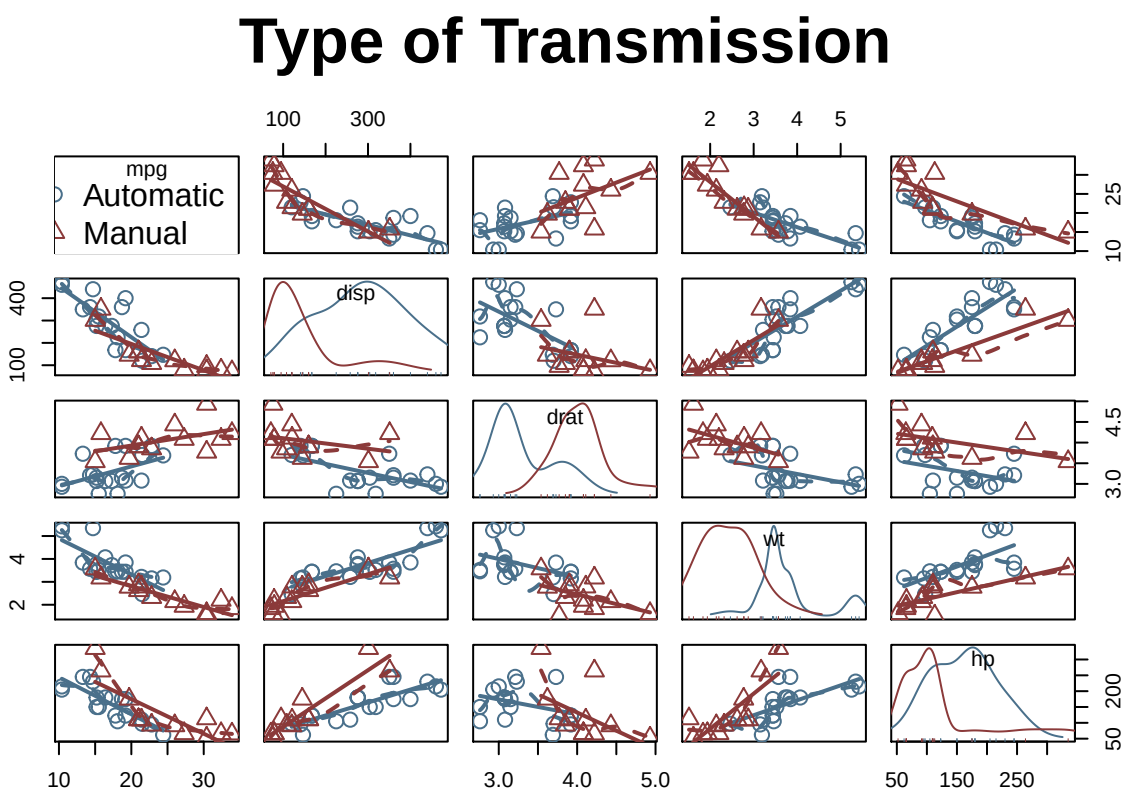
```
mtcars$am = as.factor(mtcars$am)
levels(mtcars$am) = c("Automatic", "Manual")
summary(mtcars$mpg); summary(mtcars$am)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.43   19.20   20.09   22.80   33.90
```

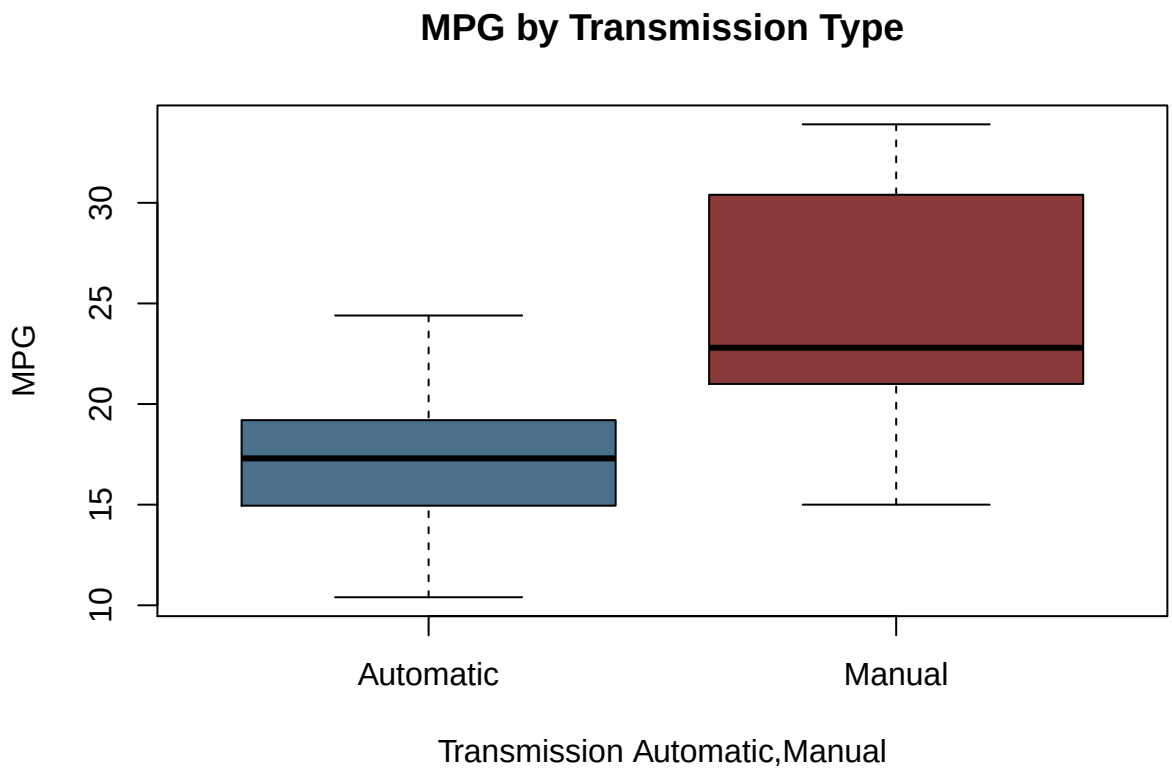
```
## Automatic    Manual
##        19        13
```

2.EDA

```
scatterplotMatrix(~mpg+disp+drat+wt+hp|am, data=mtcars, col = c("skyblue4", "indianred4"), main="Type o
```

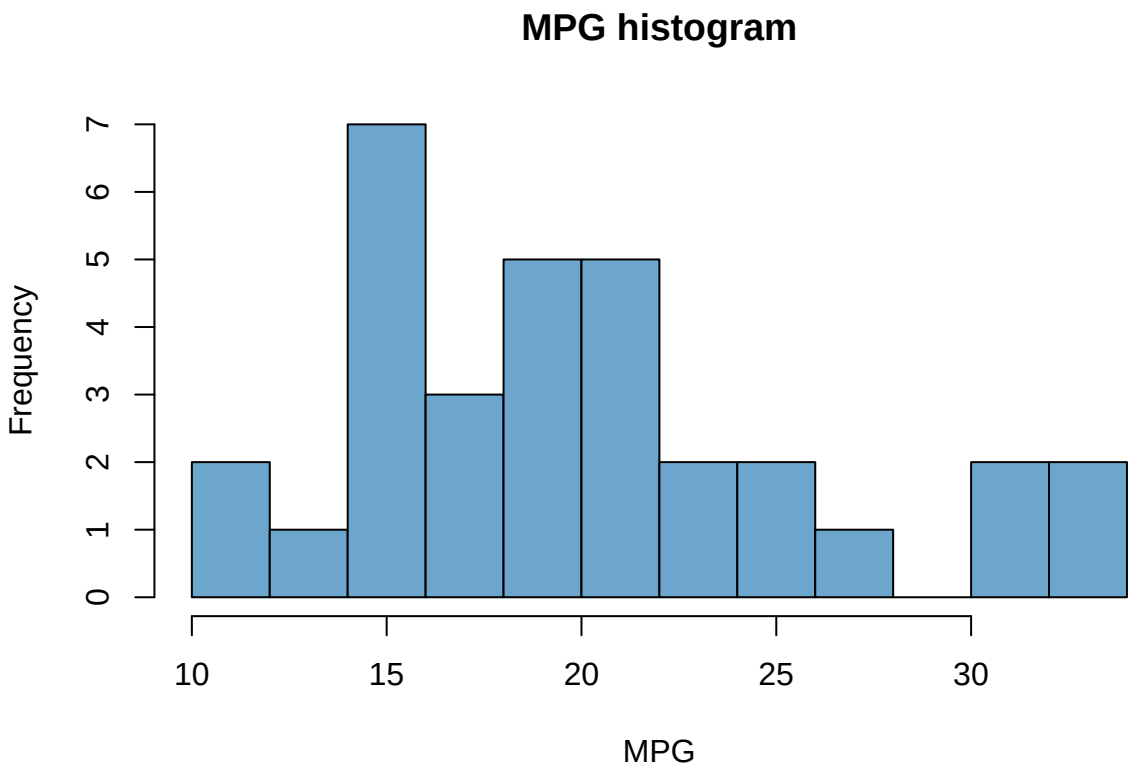



```
boxplot(mpg~am,data = mtcars,xlab = "Transmission Automatic,Manual", ylab = "MPG", main="MPG by Transmi
```

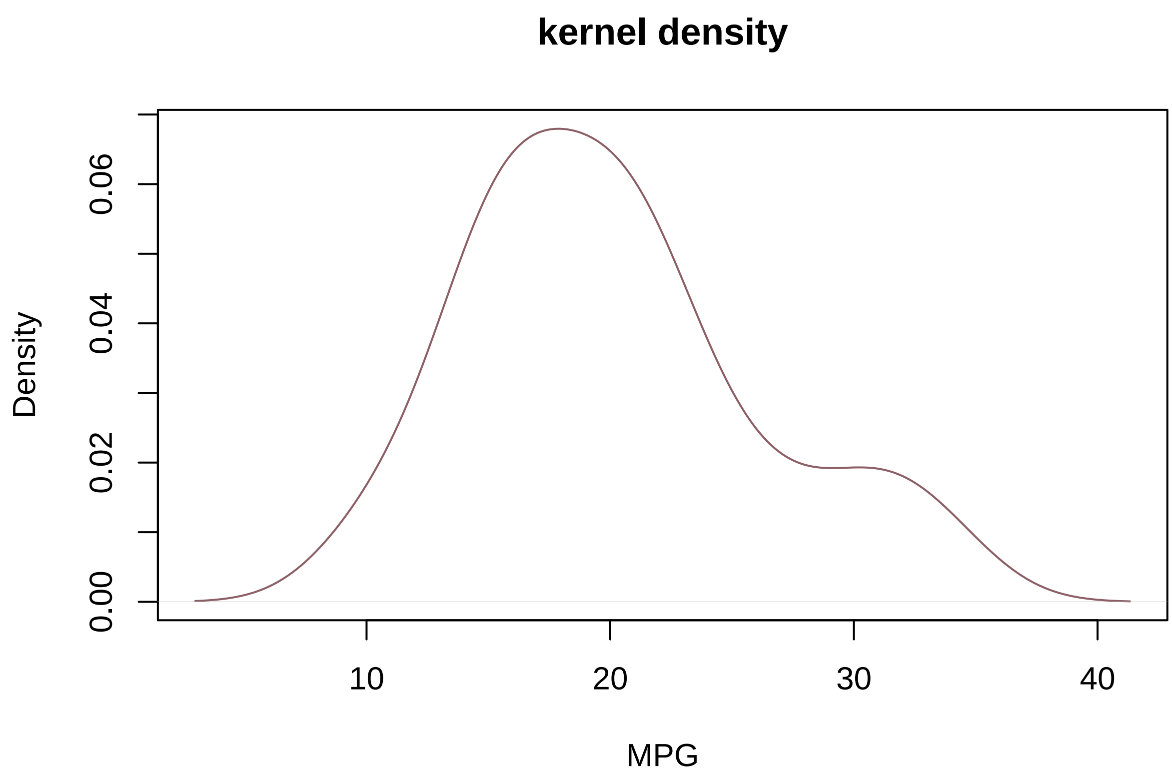


#3. t-test

```
hist(mtcars$mpg, breaks=10, xlab="MPG", main="MPG histogram", col = "skyblue3")
```



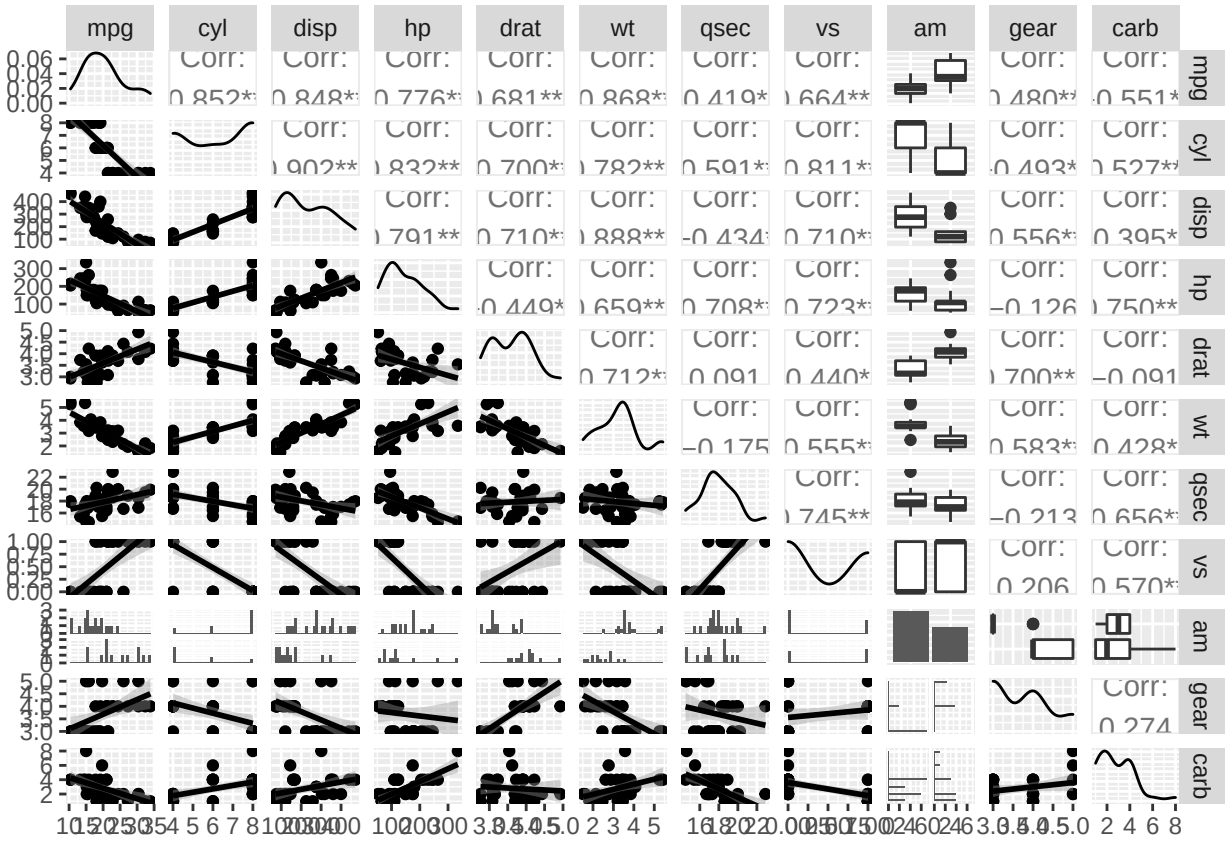

```
plot(density(mtcars$mpg), main="kernel density", xlab="MPG", col="lightpink4" )
```




```r
library(GGally)
library(ggplot2)

gp = ggpairs(mtcars, lower = list(continuous = "smooth"))
gp
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Interpretation : In this plot, we see many multi-collinearity, and it suggests that we should NOT use all the variables as predictor otherwise it will be overfitting.

4.Quantify the MPG difference between automatic and manual transmissions

Consider all the other varaibles as possible predictor and MPG as outcome. Use R step function to find out the best fit model

First, Glimpse at all relationship between each variable

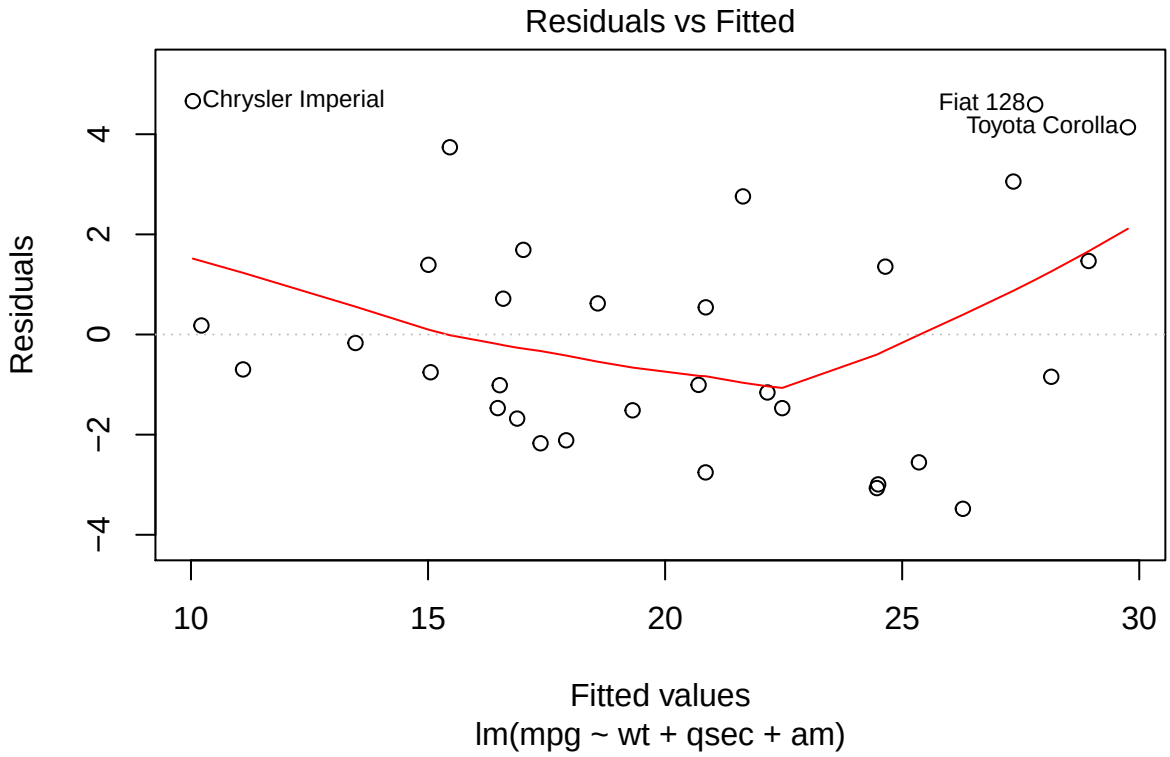
Finding best model

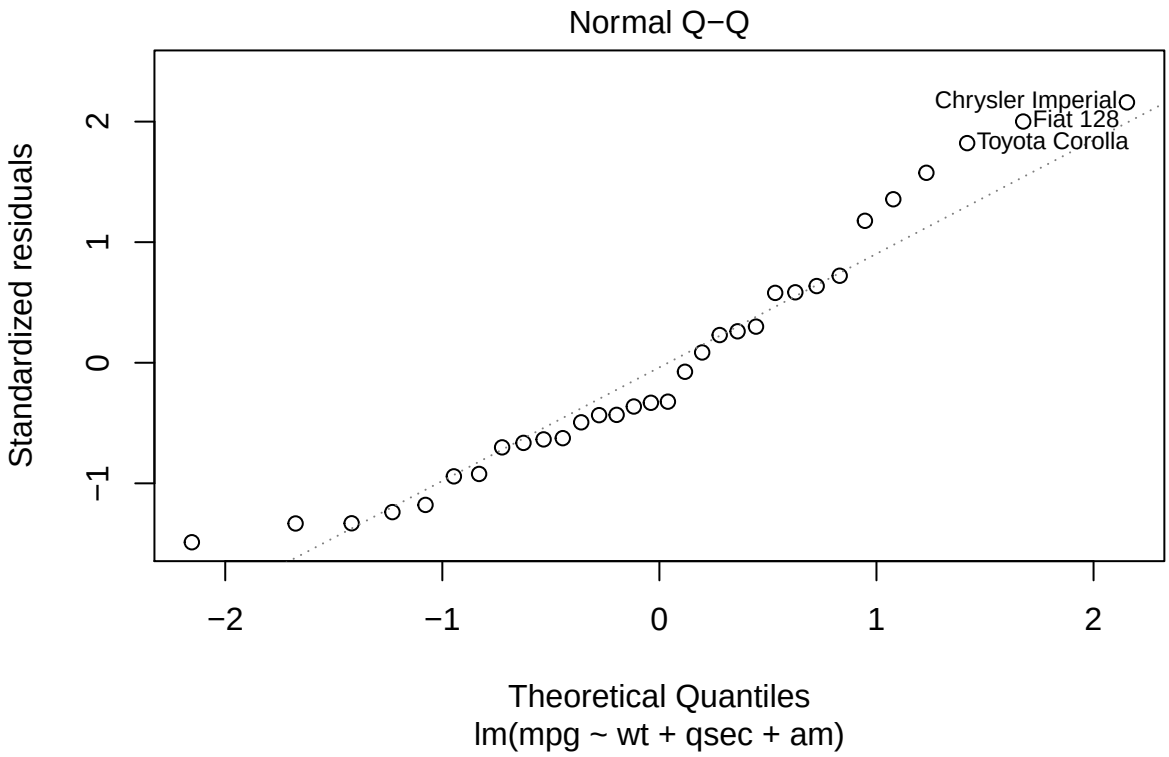
```
best_model<-step(lm(mpg ~ .,data = mtcars), trace=0)
summary(best_model)
```

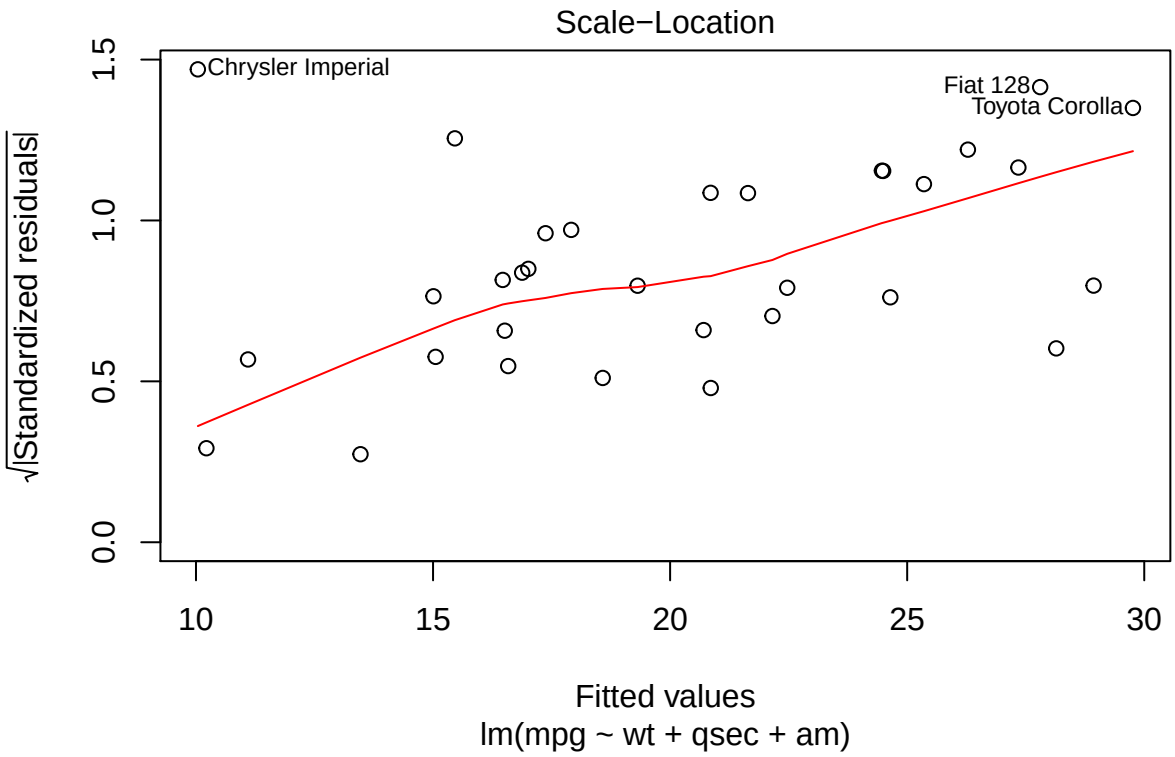
```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
```

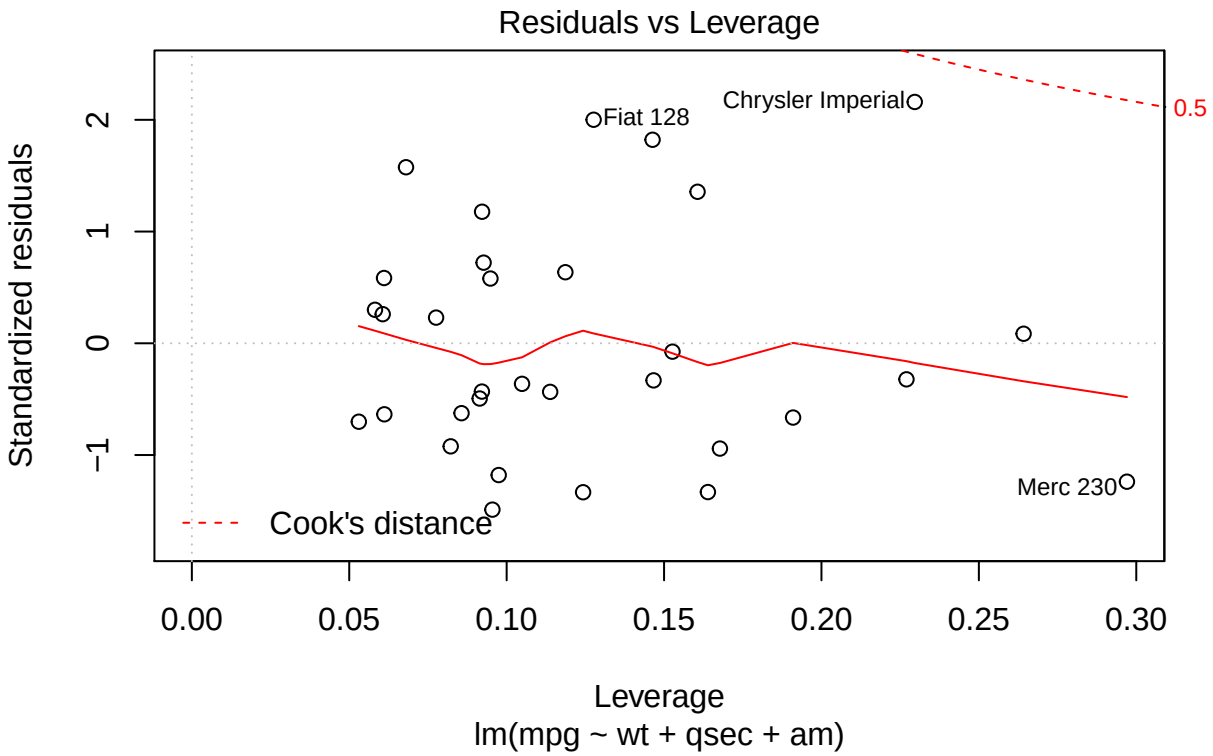
```
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## amManual      2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

```
#par(mfrow=c(2,2))
plot(best_model)
```




Normal Q−Q
Chrysler Imperial
Fiat 128
Toyota Corolla
Standardized residuals
−1
0
1
2
−2
−1
0
1
2
Theoretical Quantiles
lm(mpg ~ wt + qsec + am)


Scale−Location
Chrysler Imperial
Fiat 128
Toyota Corolla
√|Standardized residuals|
0.0
0.5
1.0
1.5
10
15
20
25
30
Fitted values
lm(mpg ~ wt + qsec + am)



We can conclude that the best model are with wt/qsec/am as predictor and the R-square is 84.97%, which is good fitting to mpg outcome.The mpg of manual cars is 2.9358 mpg better than that of automatic cars.