Linear Regression

Here **mtcars** dataset is using for this topic, which is available in the R. Let us determine the factors on which the mileage of car depends using multiple linear regression.To know more about the dataset use the code `?mtcars` .
Here *dplyr* package is using for easy manipulation of dataset.

```
library(dplyr)
```

Now import dataset using below code.

```
data("mtcars")
df = mtcars
str(df) # or use glimpse(df)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Select only those variables which are continuous in nature.

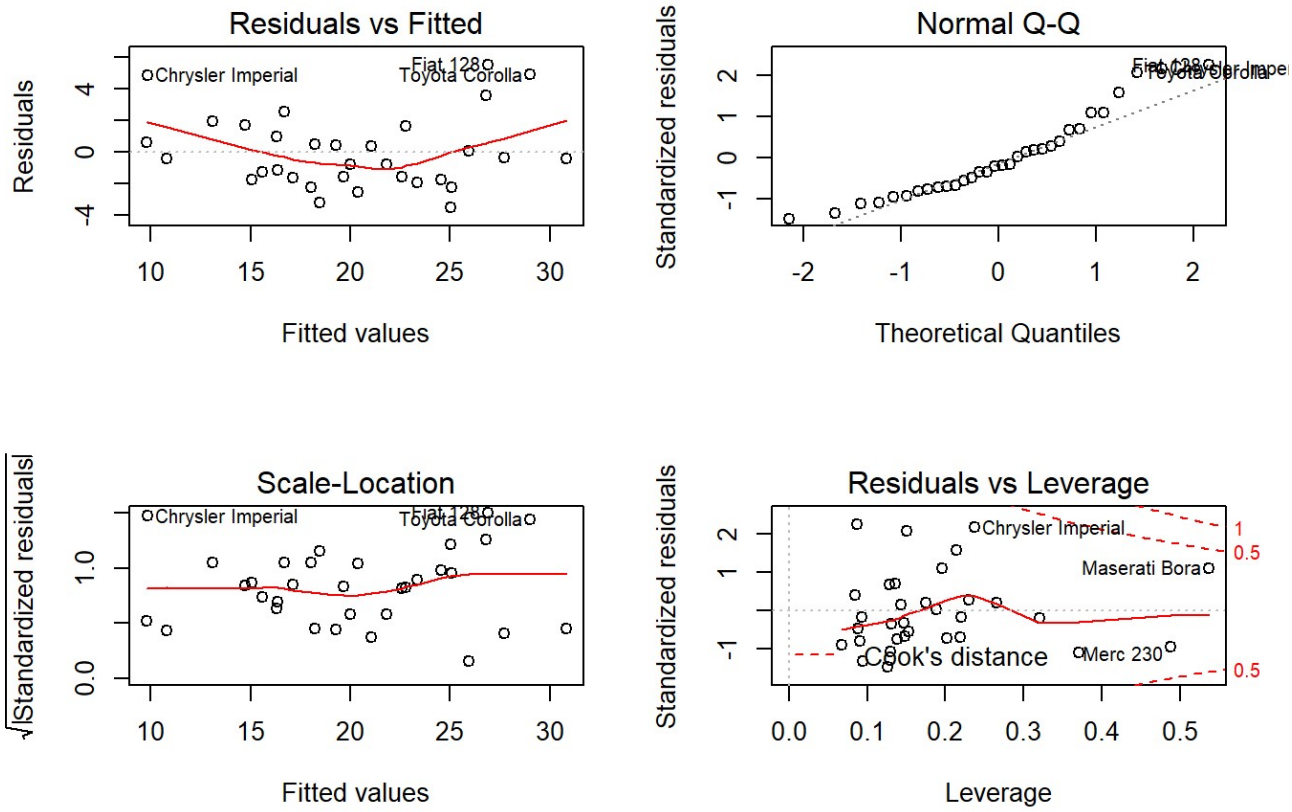
```
df1 <- select(df,-c(cyl,vs,am,gear,carb))
```

Fit a multiple linear regression model for the above data using *mpg* as response variable and all other variables as *predictor* variables.

```
mlr1 <-lm(mpg~.,data = df1)
summary(mlr1)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = df1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5404 -1.6701 -0.4264  1.1320  5.4996
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.53357   10.96423   1.508  0.14362
## disp         0.00872    0.01119   0.779  0.44281
## hp          -0.02060    0.01528  -1.348  0.18936
## drat         2.01578    1.30946   1.539  0.13579
## wt          -4.38546    1.24343  -3.527  0.00158 **
## qsec         0.64015    0.45934   1.394  0.17523
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.558 on 26 degrees of freedom
## Multiple R-squared:  0.8489, Adjusted R-squared:  0.8199
## F-statistic: 29.22 on 5 and 26 DF,  p-value: 6.892e-10
```

```
par(mfrow=c(2,2))
plot(mlr1)
```




```
dev.off() # to reset grephic parameter to default
```

Except *wt* all others variables have p-value more than 0.05. In Residual vs fitted graph data points are not evenly distributed around zero line. Therefore the model is not a good fit.
In the model there may be presence of heteroscedacity and multicolinearity.Use **Breusch-Pagan test** and **NCV Test** for cheking heteroscadacity in the model.

```
lmtest::bptest(mlr1)  # Breusch-Pagan test
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mlr1
## BP = 2.5157, df = 5, p-value = 0.7741
```

```
car::ncvTest(mlr1)    #NCV Test
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.255517, Df = 1, p = 0.2625
```

Here the p-value for both the test is more than 0.05, therefore there is no heteroscedacity in the model. **VIF** is used to determine the presence of multicolinearity in the data.

```
car::vif(mlr1)
```

```
##     disp       hp     drat       wt     qsec
## 9.110869 5.201833 2.322343 7.012686 3.191939
```

from the description of dataset we came to know that *qsec* and *disp* are the function of someother variables, it is better to drop those variables from the model

```
mlr2 <-update(mlr1,~.-qsec-disp)
summary(mlr2)
```

```
##
## Call:
## lm(formula = mpg ~ hp + drat + wt, data = df1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3598 -1.8374 -0.5099  0.9681  5.7078
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.394934   6.156303   4.775 5.13e-05 ***
## hp          -0.032230   0.008925  -3.611 0.001178 **
## drat         1.615049   1.226983   1.316 0.198755
## wt          -3.227954   0.796398  -4.053 0.000364 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.561 on 28 degrees of freedom
## Multiple R-squared:  0.8369, Adjusted R-squared:  0.8194
## F-statistic: 47.88 on 3 and 28 DF,  p-value: 3.768e-11
```

Above result shows *drat* is insignificant in the model because it has p-value more than 0.05. Now the updated model is given by,

```
mlr3 <- update(mlr2,~.-drat)
summary(mlr3)
```

```
##
## Call:
## lm(formula = mpg ~ hp + wt, data = df1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.941 -1.600 -0.182  1.050  5.854
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
## hp          -0.03177    0.00903  -3.519  0.00145 **
## wt          -3.87783    0.63273  -6.129 1.12e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.593 on 29 degrees of freedom
## Multiple R-squared:  0.8268, Adjusted R-squared:  0.8148
## F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

The model is given by $mpg = 37.22727 - 0.03177 * hp - 3.87783 * wt$
The occuracy of the model is 81.48%.