# mtcars data analysis using regression models

*kalai*

*12/10/2019*

## Summary

mtcars dataset contains a collection of data compiled for 32 cars analyzed among 11 parameters. The following two questions are taken up for analysis in this work.

1. Is an automatic or manual transmission better for efficent fuel consumption
2. Quantifying the MPG difference between automatic and manual transmissions

First, exploratory data analysis is performed to analyse and understand the data. Then, regression models are used to determine the relationship and correlation among the measured parameters in the dataset. Regression analysis is done by including and excluding different parameters contained in the dataset.

The main objective is to infer how mpg is affected by manual and automatic transmission. An analysis is also performed to measure the influence of other parameters present in the dataset using regression analysis.

## Exploring the data

```r
library(datasets) #loading the data
data(mtcars)
```

```r
head(mtcars) #exploring the dataset
```

```
##                    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
```

```r
summary(mtcars)
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec             vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

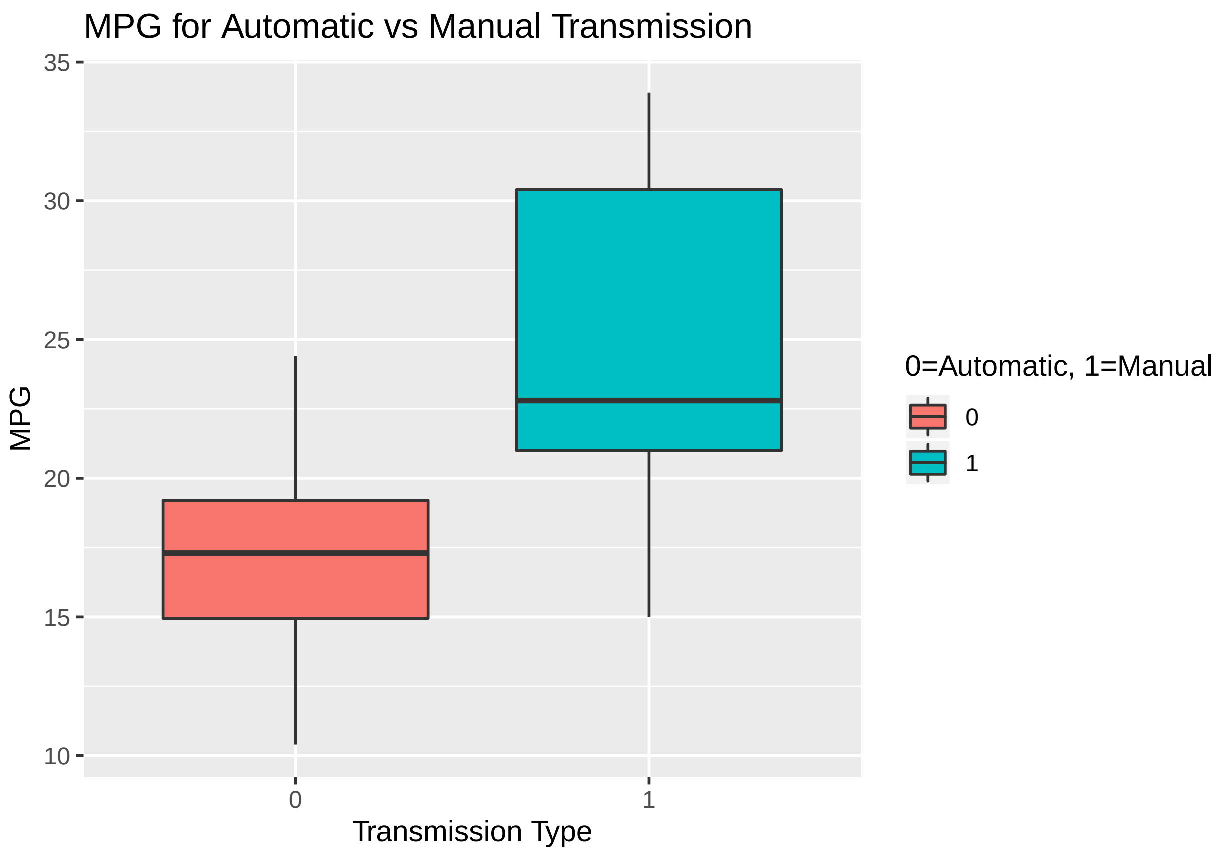
```r
str(mtcars)
```

```
## '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 ...
```

Let us visualize the mileage (mpg) obtained for automatic and manual transmission

```r
library(ggplot2)
```

```r
mtcars$am <- as.factor(mtcars$am)
h <- ggplot(mtcars, aes(x=am, y=mpg,fill=am)) + geom_boxplot()
h <- h+labs(title = "MPG for Automatic vs Manual Transmission")
h <- h + xlab("Transmission Type")
h <- h + ylab("MPG")
h <- h + labs(fill = "0=Automatic, 1=Manual")
h
```



It can be inferred from the plot that the mode of transmission has a significant impact on the mileage. It can also be seen that the median for automatic transmission is situated around the middle of the boxplot, whereas the median for manual transmission is situated way below the box, indicating larger dispersion of values.

Let us do a statistical analysis of the 'mpg' and 'am' column in the dataset

```
s = split(mtcars$mpg, mtcars$am)

sapply(s, mean) # finding the mean
```

```
##        0        1
## 17.14737 24.39231
```

```
sapply(s, sd) # standard deviation
```

```
##        0        1
## 3.833966 6.166504
```

To frame the hypothesis test, let's rearrange the required data

```
automatic <- mtcars[mtcars$am == "0",]
manual <- mtcars[mtcars$am == "1",]
```

### Performing t-test

```
t.test(automatic$mpg, manual$mpg)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  automatic$mpg and manual$mpg
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

From the result of the t-test it can be concluded that null hypothesis is rejected. It implies that mileage is significantly affected by the type of transmission.

## Regression analysis

To analyse how the transmission type impacts mileage, regression analysis is performed. Diferent variables of the datset are included and excluded to study the impact of those attributes over determining the result of the regression model.

A linear regression is performed by including all the variables in the dataset

```r
R_all <- lm(mpg ~., data = mtcars)
summary(R_all)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am1          2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

```r
RA <- lm(mpg ~ am, data = mtcars)
summary(RA)
```

```
##
## Call:
```

```
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am1            7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

```r
RS <- lm(mpg~am + cyl + hp + wt + disp, data = mtcars)
summary(RS)
```

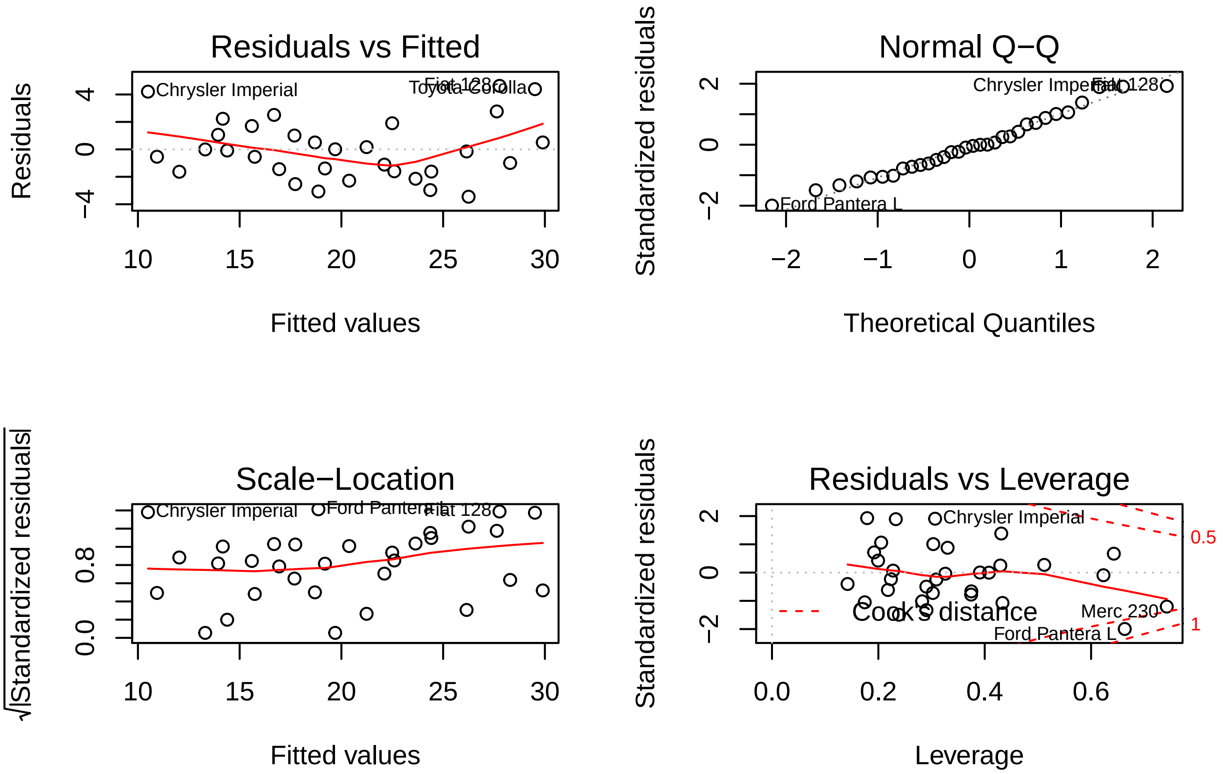
```
##
## Call:
## lm(formula = mpg ~ am + cyl + hp + wt + disp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5952 -1.5864 -0.7157  1.2821  5.5725
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.20280    3.66910  10.412 9.08e-11 ***
## am1          1.55649    1.44054   1.080  0.28984
## cyl         -1.10638    0.67636  -1.636  0.11393
## hp          -0.02796    0.01392  -2.008  0.05510 .
## wt          -3.30262    1.13364  -2.913  0.00726 **
## disp         0.01226    0.01171   1.047  0.30472
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.505 on 26 degrees of freedom
## Multiple R-squared:  0.8551, Adjusted R-squared:  0.8273
## F-statistic:  30.7 on 5 and 26 DF,  p-value: 4.029e-10
```

```r
anova(RA, RS)
```

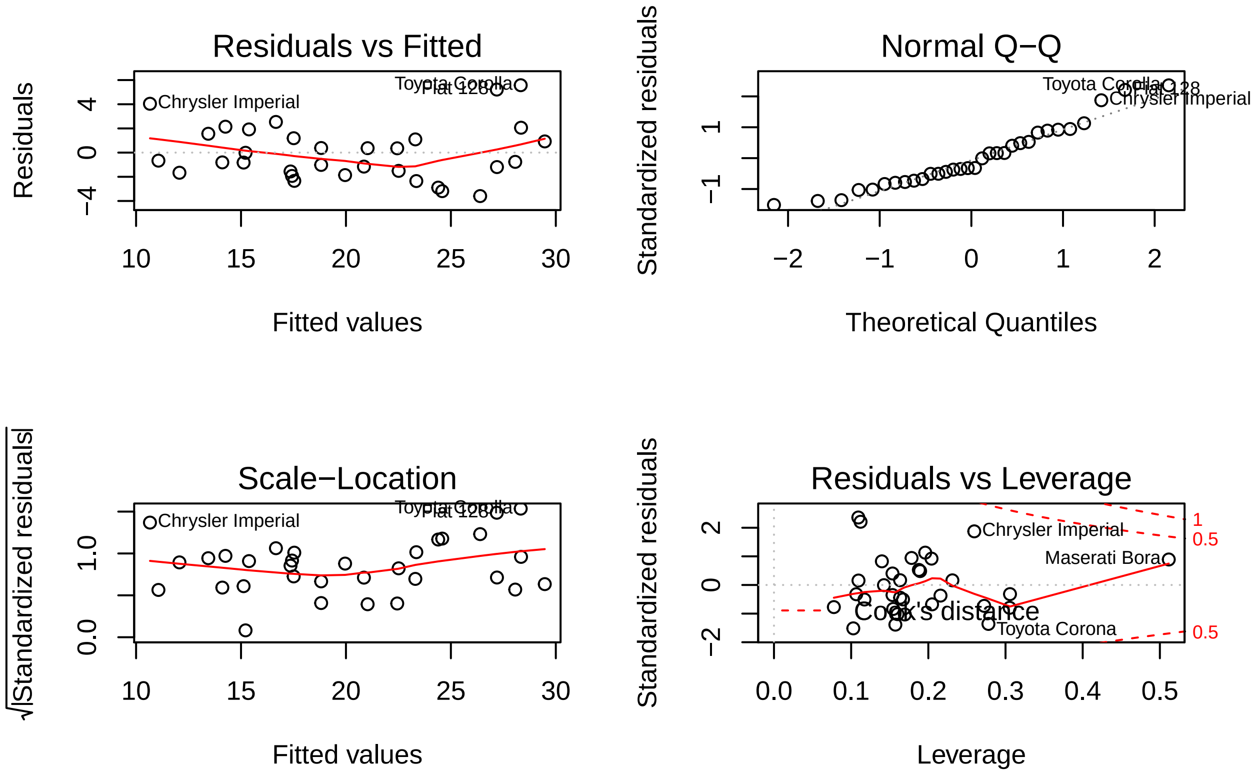
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + hp + wt + disp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 163.12  4    557.78 22.226 4.507e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### Plots of the regression model residuals

```r
## regression residual plots considering all attributes in the datset
par(mfrow = c(2,2))
plot(R_all)
```



```r
## regression residual plots considering selected attributes in the datset
par(mfrow = c(2,2))
plot(RS)
```



In the linear regression between the mileage and transmission type, manual transmission had 7.245 mpg better mileage than automatic. However, viewing the R-squared value, we can infer mileage as per transmission types affects only around 36% of mpg performance metric.

Hence to take into account of the effect of other atributes over mpg, multivariate regression is performed.

First, all the variables are taken into account for regression analysis.An increase in 2.5 mpg can be observed while factoring in all the varibles. The R-squared value is around 87% indicating a greater impact.

To have a more concrete analysis, I have done a regression on mpg to am, with number of cylinders, horsepower,weight and displacement of the vehicle as additional variables to the regression model. An increase in 1.5 mpg is observed for manual transmission when compared with automatic transmission. The impact (R-squared value) is also around 85 %.