Regression Model Course Project

Sachin Sharma

10/7/2021

Brief Summary

In this report, we will examine the **mtcars** data set and explore how miles per gallon (MPG) is affected by different variables. As per the requiremnt of the project, we will answer the following two questions:

1. Is an automatic or manual transmission better for MPG, and
2. Quantify the MPG difference between automatic and manual transmissions.

Exploratory Data Analysis

Importing Libraries

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(naniar)
library(dplyr)
library(datasets)
library(tinytex)
library(DT)
```

Reading Data

```
data("mtcars")


head(mtcars)
```

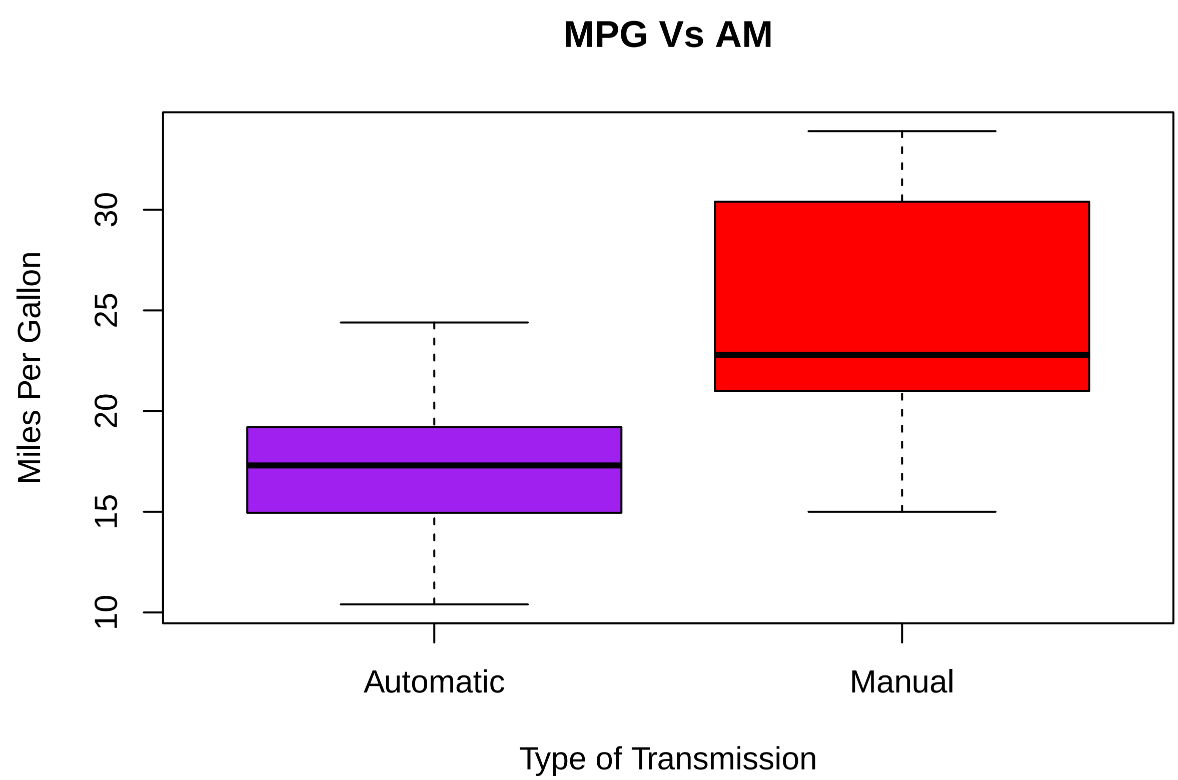
```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
```

```
## 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
```

Transform certain variables into factors

```
mtcars$cyl  <- factor(mtcars$cyl)
mtcars$am   <- factor(mtcars$am,labels=c("Automatic","Manual"))
mtcars$vs   <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
```

```
boxplot(mpg ~ am, data = mtcars, col = (c("purple","red")), ylab = "Miles Per Gallon", xlab = "Type of '
```



Regression Analysis

With the help of plot, We've visually seen that automatic is better for MPG, but we will now quantify his difference

```
aggregate(mpg~am, data = mtcars, mean)
```

```
##          am      mpg
## 1 Automatic 17.14737
## 2    Manual 24.39231
```

Difference of MPG between Automatic and Manual

```
24.39231 - 17.14737
```

```
## [1] 7.24494
```

Therefore, we can see that the Manual cars have an MPG of 7.245 (approx.) more than automatic cars

We can now use a t-test here

What is t-test ?

The t-test assesses whether the means of two groups are statistically different from each other.This analysis is appropriate whenever you want to compare the means of two groups

```
automatic_car <- mtcars[mtcars$am == "Automatic",]
manual_car <- mtcars[mtcars$am == "Manual",]
t.test(automatic_car$mpg, manual_car$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  automatic_car$mpg and manual_car$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
```

We can see that the p-value is 0.001374, thus we can state this is a significant difference. Now to quantify this, we can use the following code :

```
model_1  <- lm(mpg ~ am, data = mtcars)
summary(model_1)
```

```
##
## 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 ***
## amManual       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
```

The above data shows us that the average MPG for automatic is 17.1 MPG, while manual is 7.2 MPG higher.

The R^2 value is 0.36 which states that, this model only explains us 36% of the variance. As a result, here we require to build a multivariate linear regression.

Lets see with the help of corrplot , to check the correlation among the variables with mpg.

Before plotting the corrplot, we will check the structure of the data ;

```
df_1 <- subset(mtcars, select = c(mpg,cyl,disp,hp,drat,wt,qsec,vs))

head(df_1)
```

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

```
str(df_1)
```

```
## 'data.frame':    32 obs. of  8 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ 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  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
```

Here we can see that, cyl and vs columns are in factor, we will now convert this into numeric to plot corrplot and check the correlation.

```
df_1$cyl <- as.character(df_1$cyl)

df_1$cyl <- as.numeric(df_1$cyl)

df_1$vs <- as.character(df_1$vs)

df_1$vs <- as.numeric(df_1$vs)
```

```r
# Now we can check the structure of the data again
str(df_1)
```

```
## 'data.frame':	32 obs. of  8 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 ...
```

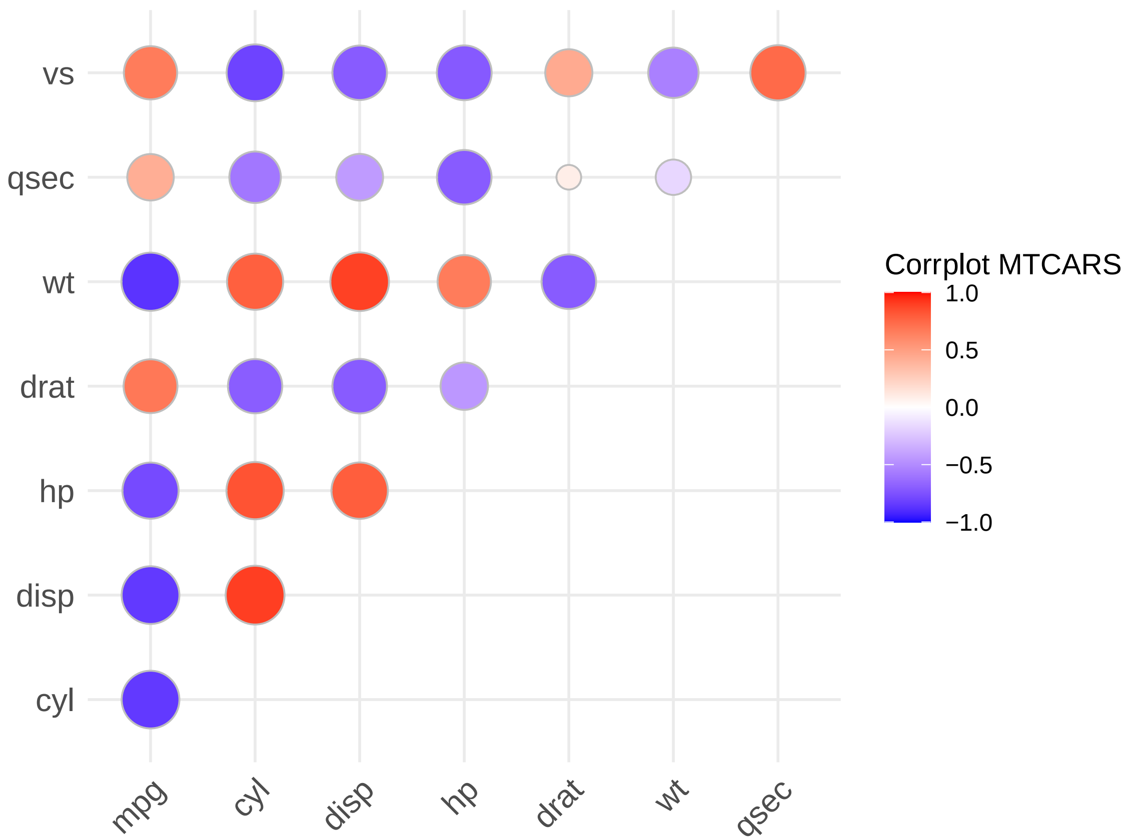
Now we can see that all the columns are in numeric, now we can plot wit the help of ggcorrplot and corrplot to check the correlation :

```r
library(ggcorrplot)

r <- cor(df_1)

ggcorrplot(r,method = "circle", type = c("upper"), legend.title = "Corrplot MTCARS")
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```

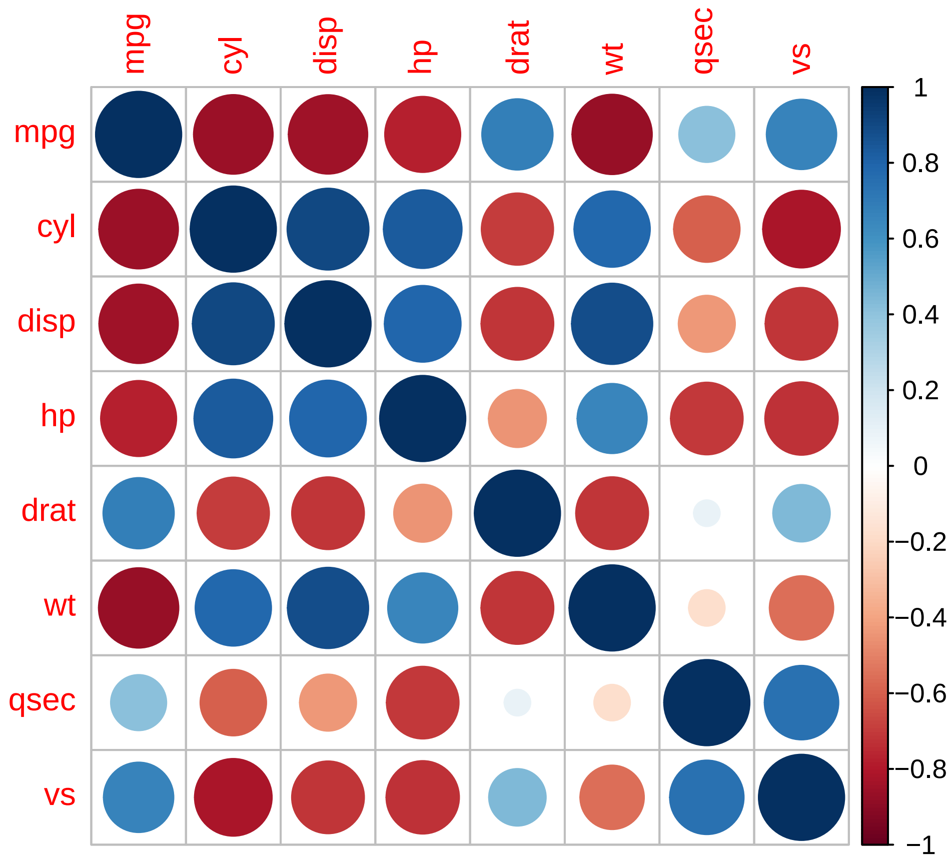



```
library(corrplot)
```

```
## corrplot 0.90 loaded
```

```
r <- cor(df_1)


corrplot(r, method = "circle")
```



With the help of above two plots, we can easily say that cyl,disp, hp and wt have strong correlation with mpg

We build a new model using these variables and compare them to the initial model with the anova function.

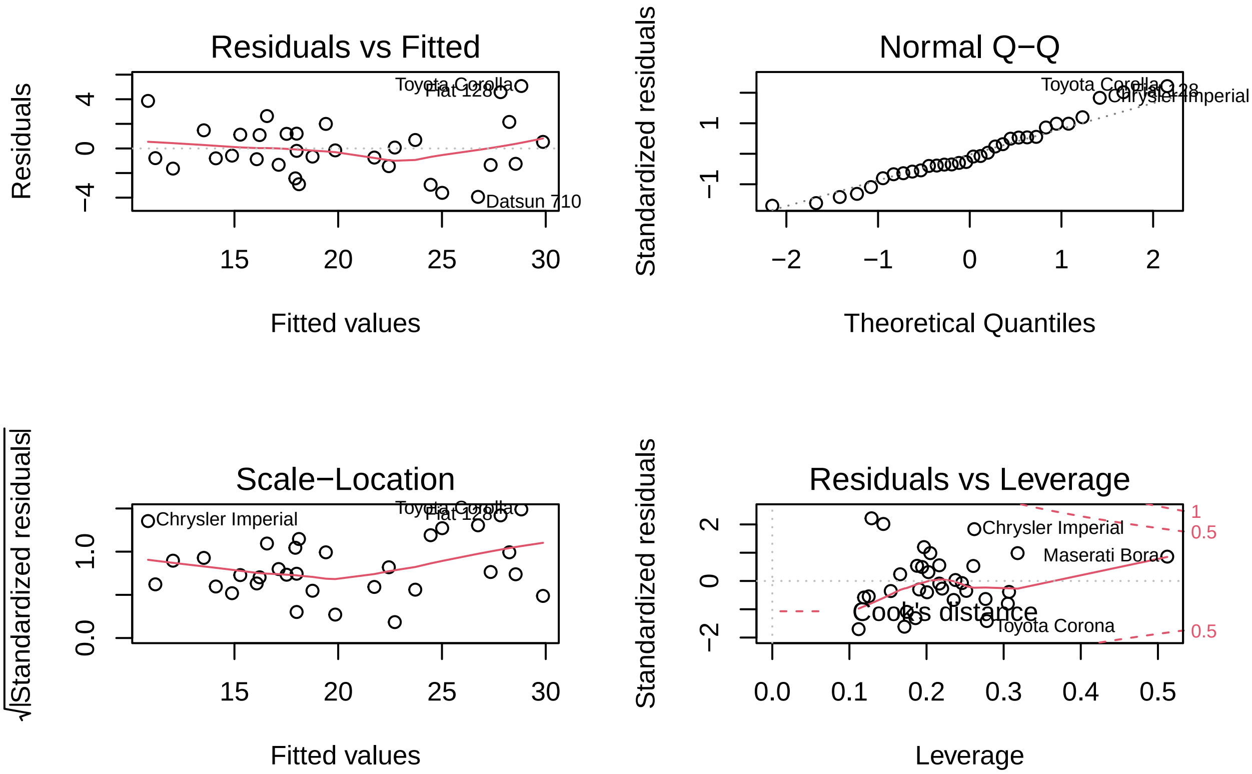
```
model_2  <- lm(mpg~am + cyl + disp + hp + wt, data = mtcars)
anova(model_1, model_2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + disp + hp + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     25 150.41  5    570.49 18.965 8.637e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

\textcolor{blue}{\textbf{ Here we can see that the result of p-value is 8.637e-08, and hence we can say that our model_2 is significantly better than our model_1 which is a simple model.}}

We can plot the graph to check the residuals for non - normality and see whether they are normally distributed or not.

```r
par(mfrow = c(2,2))
plot(model_2)
```



Now we will check the summary of our model_2

```r
summary(model_2)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9374 -1.3347 -0.3903  1.1910  5.0757
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 33.864276   2.695416  12.564 2.67e-12 ***
## amManual     1.806099   1.421079   1.271   0.2155
## cyl6        -3.136067   1.469090  -2.135   0.0428 *
## cyl8        -2.717781   2.898149  -0.938   0.3573
## disp         0.004088   0.012767   0.320   0.7515
## hp          -0.032480   0.013983  -2.323   0.0286 *
## wt          -2.738695   1.175978  -2.329   0.0282 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.453 on 25 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8344
## F-statistic: 27.03 on 6 and 25 DF,  p-value: 8.861e-10
```

With the help of the above summary, we can say that model explain that there is a variance of 86.64% and as a result variables like cyl, disp, hp, wt did affect the correlation between mpg and am.

Hence, we can say the difference between automatic and manual transmissions is 1.81 MPG