Motor Trend Analysis - Effect of Transmission on Mileage

L N S S Ravi Teja
10th July 2020

Executive Summary

The goal of this project is to study the influence of variables from the mtcars dataset.Specifically we want to learn about how the mpg(Miles per gallon) is related to other variable predictors, in particular the transmission type(i.e., am(automatic or manual)). The main objective is not to build the perfect model but to explain the interaction between the predictors and response as good as possible. ### **1 : Synopsis**

Following the main aspects that we are going to explore :

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

2 : Data Processing

2.1 : Data Loading

Load the data from the mtcars.

```
library("data.table")
library("ggplot2")
library("tidyr")
library("knitr")
data(mtcars)

#converting to data.table
mtcars <- as.data.table(mtcars)

#Dimensions of data
dim(mtcars)
```

```
## [1] 32 11
```

2.2 : Find the column in data.table

```
names(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

Format of columns from ***?mtcars***

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

2.3 : Data Subsetting

As the main objective is related to transmission and mileage we subset the data to suit the requirements for Exploratory Data Analysis

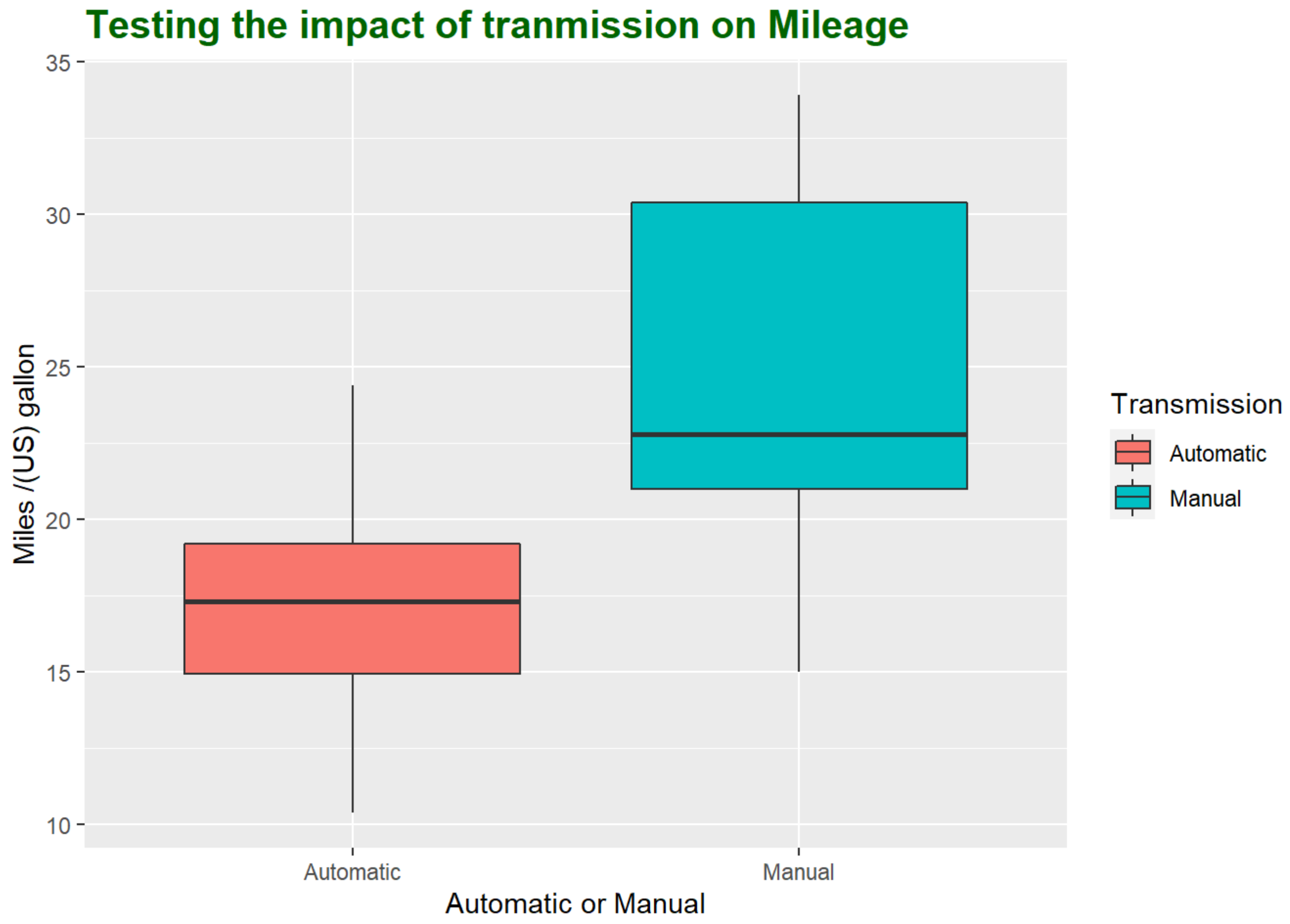
```
mtcarsDT <- mtcars[, c("mpg", "am")]
summary(mtcarsDT)
```

```
##       mpg              am        
##  Min.   :10.40   Min.   :0.0000  
##  1st Qu.:15.43   1st Qu.:0.0000  
##  Median :19.20   Median :0.0000  
##  Mean   :20.09   Mean   :0.4062  
##  3rd Qu.:22.80   3rd Qu.:1.0000  
##  Max.   :33.90   Max.   :1.0000  
```

3: Data Analysis

3.1 : Boxplot between Gas Mileage(mpg) and Transmission(am)

```
mtcarsDT$am <- factor(mtcarsDT$am)
g <- ggplot(mtcarsDT, aes(x = am, y = mpg, fill = am)) + geom_boxplot()
g <- g + scale_fill_discrete(name = "Transmission", labels = c("Automatic", "Manual"))
g <- g + scale_x_discrete(labels = c("Automatic", "Manual"))
g <- g + theme(plot.title = element_text(color = "DarkGreen", size = 15, face = "bold"))
g <- g + labs(x = "Automatic or Manual", y = "Miles /(US) gallon", title = "Testing the impac
t of tranmission on Mileage")
g
```



The above analysis definitely provides us the information that the transmissions have a large effect in mileages of cars. And Manual transmissions have better averages and provide high gas mileage than that of Automatic transmission.We can also get that the positions of means of the 2 sets are having a big gap between them.

- **for automatic transmissions** the median is at the middle of the data i.e. the data is simmetrical along the median.
- **for manual transmissions** the median is on the lower side of the data i.e. the data is largely dispersed along the higher values.

```
kable(mtcars[, .(mean = mean(mpg)), by = .(am)])
```

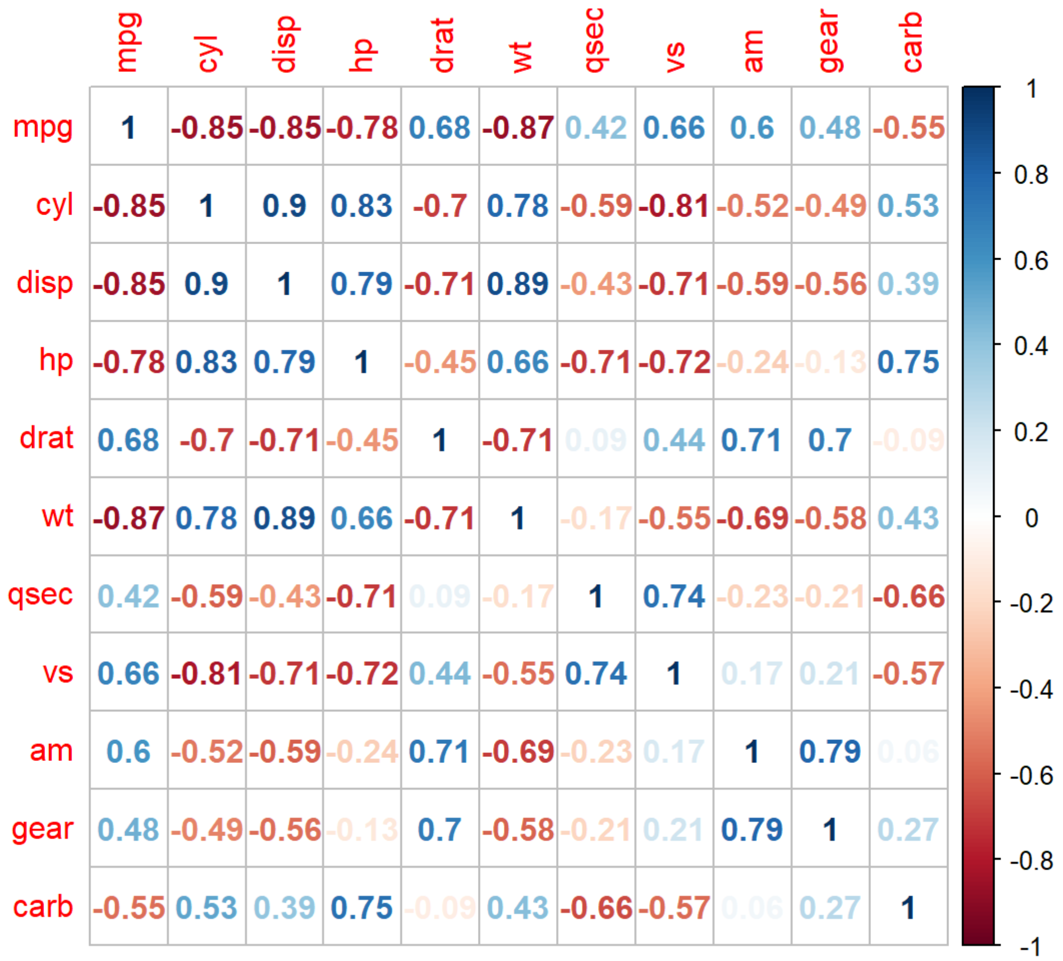
am	mean
1	24.39231
0	17.14737

4 : Modelling the Data

To model data we can follow 2 types of approaches
1 Forward Selection
2 Backward Elimination

Before Selecting any type of model we need to analyse the correlation between the predictors such that the variables that we select doesnot cause any biased predictions

```
corrplot::corrplot(cor(mtcars), method = "number")
```



4.1 : Basic Linear Model

Lets create a basic linear model mpg vs am

```
fit1 <- lm(mpg ~ am, mtcars)
summary(fit1)
```

```
##
## 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 ***
## am             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
```

We can see the adjusted R Square being approximately 33% which means that the transmission doesnot completely describe the mileage of motor cars.So there may be other variables that accompany the transmission system.

Here we are trying to implement ***Backward Elimination Model*** which basically states that we start with all the predictors first and then remove the useless predictor based on proofs. Since we already made a correlation matrix it makes the steps much easier.

4.2 : All Predictor Multi Variable Regression Model

Lets now include all predictors in the model that explains mpg.

```
fit2 <- lm(mpg ~ ., mtcars)
summary(fit2)
```

```
##
## 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
## am           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
```

As we can see that the Adjusted R square is much higher and thats the proof that other variables have a certain impact on mpg.

This is the point we have the need of multicollinearity among the variables so as to eliminate the variables.

4.3 : Eliminating disp, vs and hp as they have high collinearity and depends on other variables

```
fit3 <- lm(mpg ~ . - disp-vs-hp, mtcars)
summary(fit3)
```

```
## 
## Call:
## lm(formula = mpg ~ . - disp - vs - hp, data = mtcars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.0140 -1.0985 -0.2574  1.3572  4.3389 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  11.5557    17.8225   0.648   0.5229  
## cyl          -0.1519     0.8485  -0.179   0.8594  
## drat          0.9618     1.5564   0.618   0.5424  
## wt           -2.8571     1.0861  -2.631   0.0147 *
## qsec          0.7942     0.5940   1.337   0.1937  
## am            2.4682     1.9276   1.281   0.2126  
## gear          0.5701     1.4056   0.406   0.6887  
## carb         -0.7368     0.5636  -1.307   0.2034  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.541 on 24 degrees of freedom
## Multiple R-squared:  0.8624, Adjusted R-squared:  0.8223 
## F-statistic: 21.49 on 7 and 24 DF,  p-value: 6.892e-09
```

As we can see that the Adjusted R square is higher than that of previous model providing us the information that we are going in the right direction.(80.7% to 82.2%)

4.4 : Let’s test by eliminating drat

drat has positive correlaton between mpg and as well as “am” so we have to eliminate it.

```
fit4 <- lm(mpg ~ . - disp-vs-hp-drat, mtcars)
summary(fit4)
```

```
##
## Call:
## lm(formula = mpg ~ . - disp - vs - hp - drat, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2148 -1.1992 -0.2412  1.4018  4.4595
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.2624    15.9126   1.022   0.3166
## cyl          -0.3137     0.7971  -0.393   0.6973
## wt           -2.9548     1.0611  -2.785   0.0101 *
## qsec          0.7695     0.5853   1.315   0.2005
## am            2.6522     1.8807   1.410   0.1708
## gear          0.6415     1.3835   0.464   0.6469
## carb         -0.6764     0.5481  -1.234   0.2287
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.509 on 25 degrees of freedom
## Multiple R-squared:  0.8602, Adjusted R-squared:  0.8267
## F-statistic: 25.64 on 6 and 25 DF,  p-value: 1.542e-09
```

As we can see that the Adjusted R square is slightly higher than that of previous model providing us the information that we are going in the right direction.(82.2% to 82.7)

4.5 : Let’s test by eliminating carb

carb has negative correlation with mpg which is not a significant choice out of them but it has almost no correlation with “am” so lets test it.

```
fit5 <- lm(mpg ~ . - disp-vs-hp-drat-carb, mtcars)
summary(fit5)
```

```
##
## Call:
## lm(formula = mpg ~ . - disp - vs - hp - drat - carb, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7394 -1.3787 -0.8627  1.4275  4.6256
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.1752    15.9954   1.136 0.266206
## cyl          -0.4643     0.7956  -0.584 0.564566
## wt           -3.5917     0.9364  -3.836 0.000717 ***
## qsec          0.9349     0.5754   1.625 0.116261
## am            2.7317     1.8984   1.439 0.162097
## gear         -0.3941     1.1108  -0.355 0.725602
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.534 on 26 degrees of freedom
## Multiple R-squared:  0.8517, Adjusted R-squared:  0.8232
## F-statistic: 29.87 on 5 and 26 DF,  p-value: 5.429e-10
```

As we can see that the Adjusted R square is lesser than that of previous model providing us the information that we are not going in the right direction. Since excluding carb reduces the Adjusted R square lets check other variables.

4.6 : Let’s test by eliminating gear

```
fit6 <- lm(mpg ~ . - disp-vs-hp-drat-gear, mtcars)
summary(fit6)
```

```
##
## Call:
## lm(formula = mpg ~ . - disp - vs - hp - drat - gear, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2795 -1.2098 -0.3826  1.3961  4.4050
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20.0379    13.4638   1.488  0.14871
## cyl          -0.4612     0.7198  -0.641  0.52728
## wt           -3.0462     1.0268  -2.967  0.00638 **
## qsec          0.7272     0.5693   1.277  0.21276
## am            2.9417     1.7471   1.684  0.10419
## carb         -0.5222     0.4291  -1.217  0.23456
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.471 on 26 degrees of freedom
## Multiple R-squared:  0.859,  Adjusted R-squared:  0.8319
## F-statistic: 31.69 on 5 and 26 DF,  p-value: 2.847e-10
```

As we can see that the Adjusted R square is higher than that of previous model fit(fit4) providing us the information that we are going in the right direction.(82.7% to 83.2%) Here we tried to exclude gear because it hase very high p-value and not significant

4.7 : Let’s test by eliminating cyl

cyl has highly negative correlation with mpg and disturbs our analysis on automatic and manual transmissions.

```
fit7 <- lm(mpg ~ . - disp-vs-hp-drat-gear-cyl, mtcars)
summary(fit7)
```

```
##
## Call:
## lm(formula = mpg ~ . - disp - vs - hp - drat - gear - cyl, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1184 -1.5414 -0.1392  1.2917  4.3604
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.8972     7.4725   1.726 0.095784 .
## wt           -3.4343     0.8200  -4.188 0.000269 ***
## qsec          1.0191     0.3378   3.017 0.005507 **
## am            3.5114     1.4875   2.361 0.025721 *
## carb         -0.4886     0.4212  -1.160 0.256212
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.444 on 27 degrees of freedom
## Multiple R-squared:  0.8568, Adjusted R-squared:  0.8356
## F-statistic: 40.39 on 4 and 27 DF,  p-value: 5.064e-11
```

As we can see that the Adjusted R square is higher than that of previous model fit providing us the information that we are going in the right direction.(83.2% to 83.7%) Here we tried to exclude cyl because it hase very high p-value and not significant

4.8 : Let’s test by eliminating carb

carb has negetive correlation and as well as not having any significant p-value (>0.05)

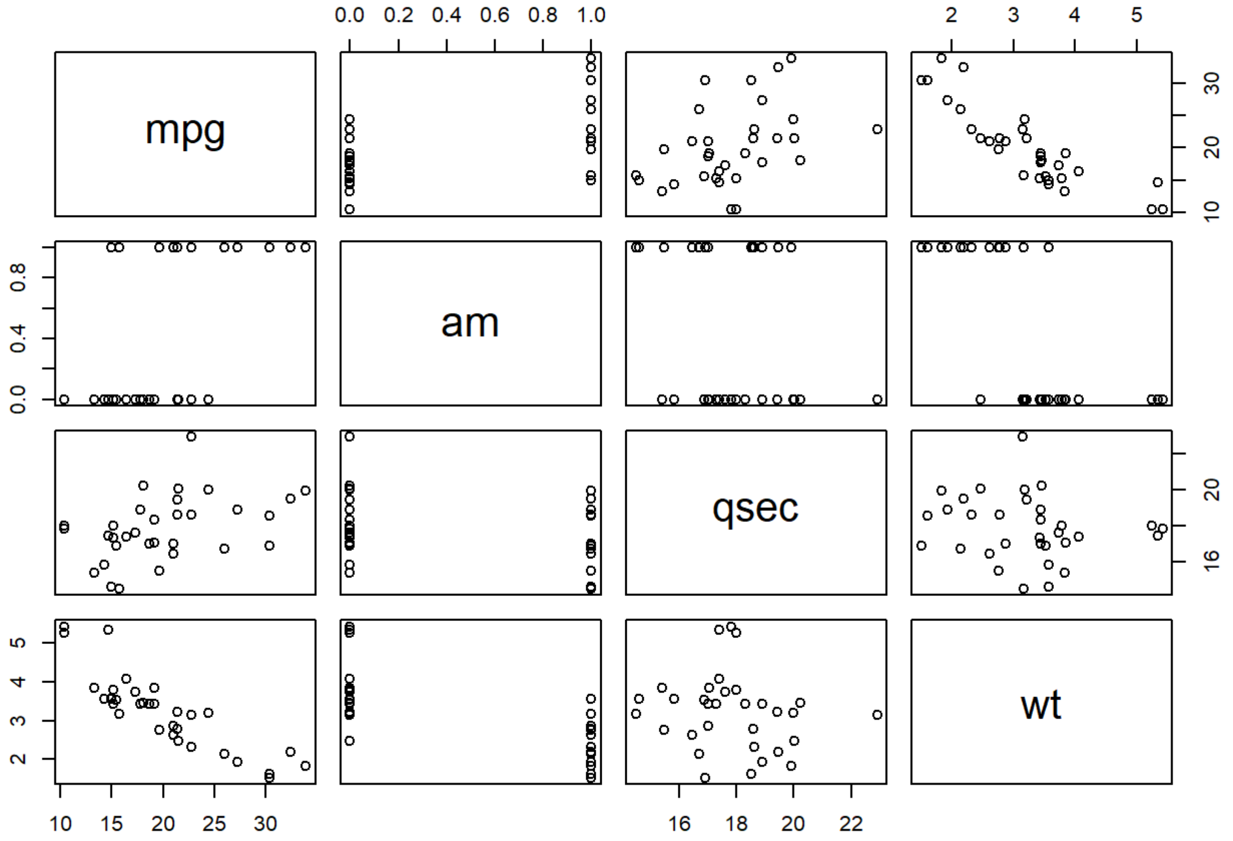
```
fit8 <- lm(mpg ~ . - disp-vs-hp-drat-gear-cyl-carb, mtcars)
summary(fit8)
```

```
##
## Call:
## lm(formula = mpg ~ . - disp - vs - hp - drat - gear - cyl - carb,
##     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 ***
## am            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
```

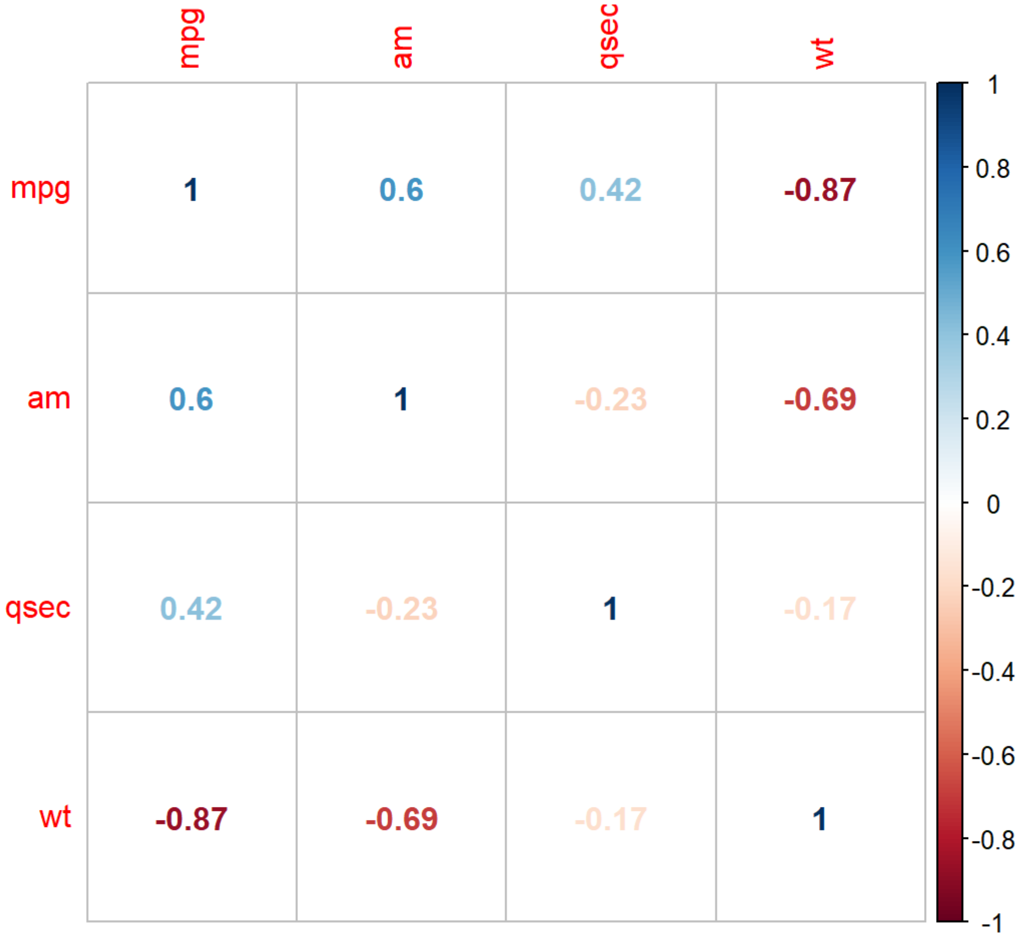
Though we find that the RSquare value is reduced we have significant P-values as well as F-Statistic that the above model is proves our statements.

The model is now settled at “wt”, “qsec”, “am”

```
mtcarsDT <- mtcars[, c("mpg", "am", "qsec", "wt")]
plot(mtcarsDT)
```



```
corrplot::corrplot(cor(mtcarsDT), method = "number")
```

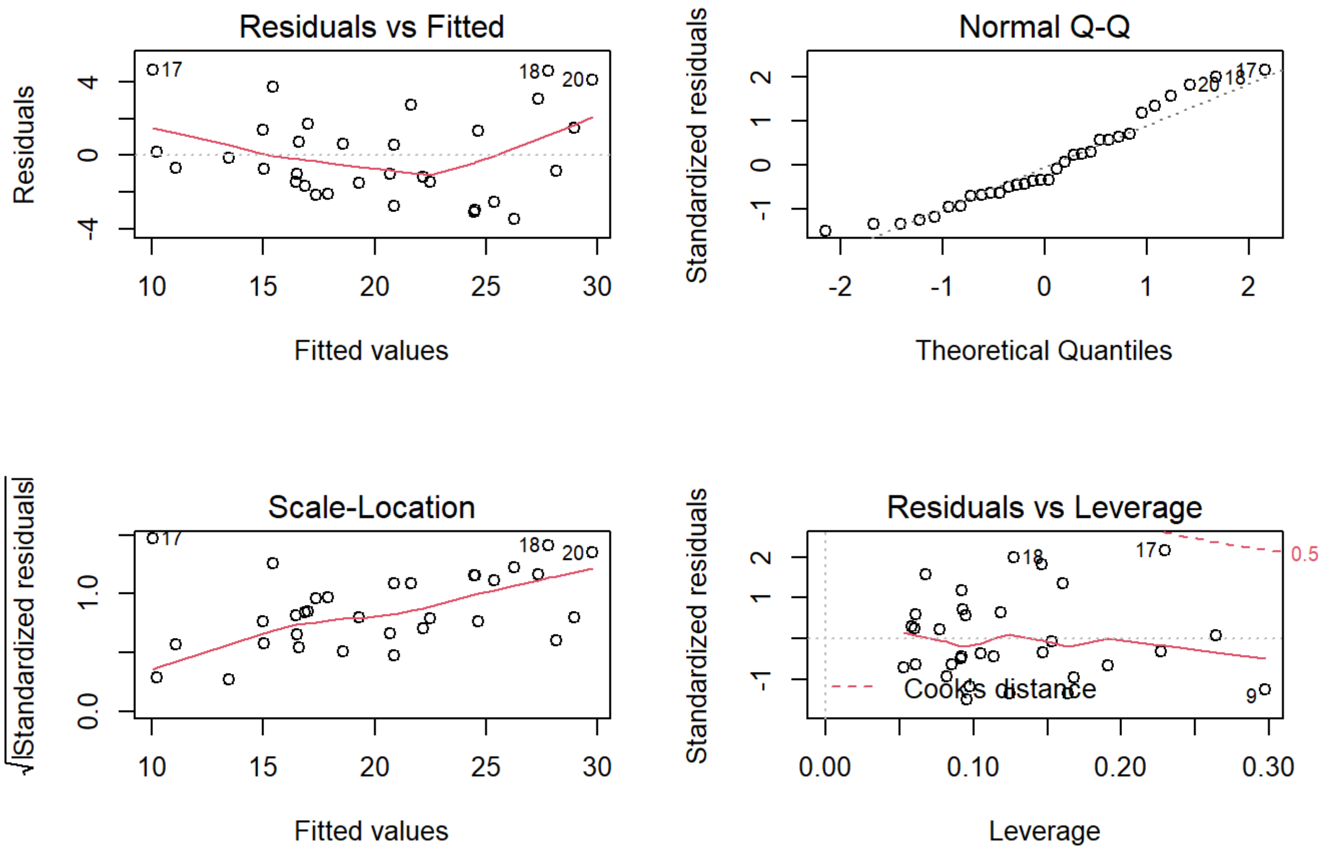


Here there is significant correlation between wt and am but that doesn't effect our analysis similarly to breath mint and smoke effect to lungs damage.

5 : Residual Analysis

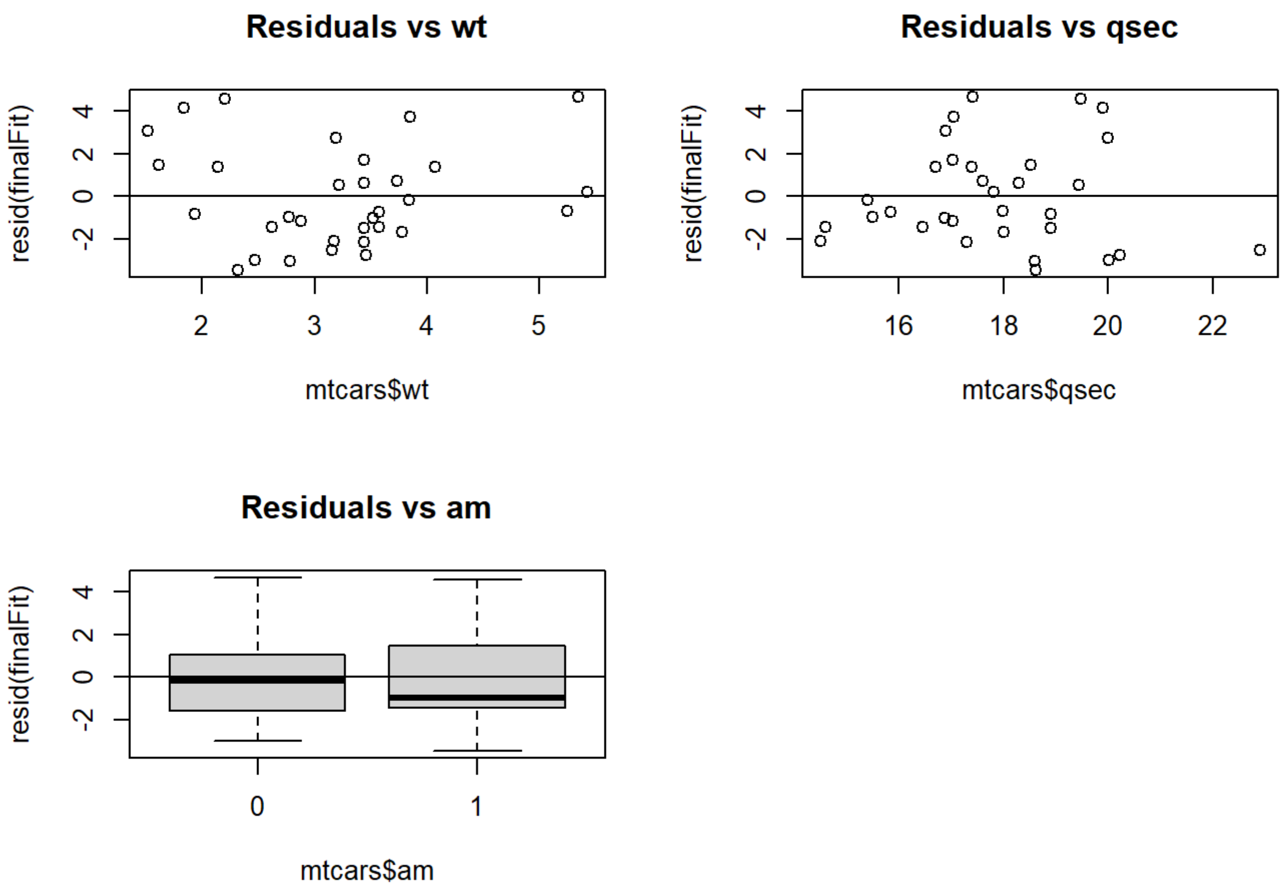
5.1 : Fitted vs Residual Values

```
finalFit <- lm(mpg ~ am + wt + qsec, mtcars)
par(mfrow=c(2,2))
plot(finalFit)
```



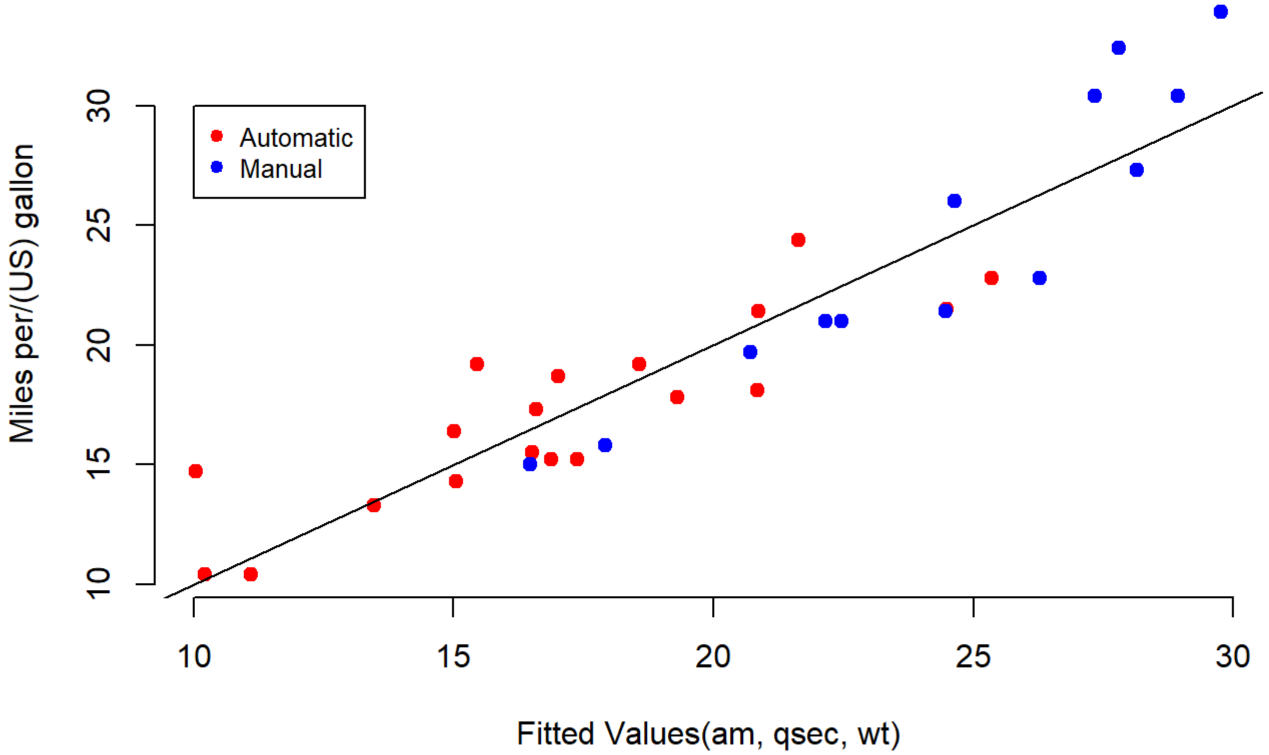
5.2 : Residuals vs Each predictor

```
par(mfrow=c(2,2))
plot(resid(finalFit) ~ mtcars$wt, main = "Residuals vs wt")
abline(h = mean(resid(finalFit)), lwd = 1)
plot(resid(finalFit) ~ mtcars$qsec, main = "Residuals vs qsec")
abline(h = mean(resid(finalFit)), lwd = 1)
boxplot(resid(finalFit) ~ mtcars$am, main = "Residuals vs am")
abline(h = mean(resid(finalFit)), lwd = 1)
```



6 : Conclusions

```
rbPal <- colorRampPalette(c('red','blue'))
mtcars$Col <- rbPal(10)[as.numeric(cut(mtcars$am,breaks = 10))]
plot(finalFit$fitted.values, mtcars$mpg,pch = 19, col  = mtcars$Col, xlab = "Fitted Values(a
m, qsec, wt)", ylab = "Miles per/(US) gallon", frame = FALSE, )
abline(lm(mtcars$mpg ~ finalFit$fitted.values))
legend(10, 30, legend=c("Automatic", "Manual"),
       col=c("red", "blue"), pch = 19, cex=0.8)
```



Based on the above graph we can conclude on various aspects of fit that **manual transmission** is better for High **mpg** values