Regression Analysis of mpg between automatic and manual vehicles

Executive Summary

The is popular belief that manual transfmission cars perform better in term of mpg than automatic.The purpose of this study is to either accept that notion or disprove it by answering the questions below.

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

Tho one cannot answer the first question without diving into the details of analysing the model, we can safely say that Manual, all things equal, tend to have better mpg than Automatic.

load the dataset and library

```
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.2.5

data(mtcars)
names(mtcars)

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

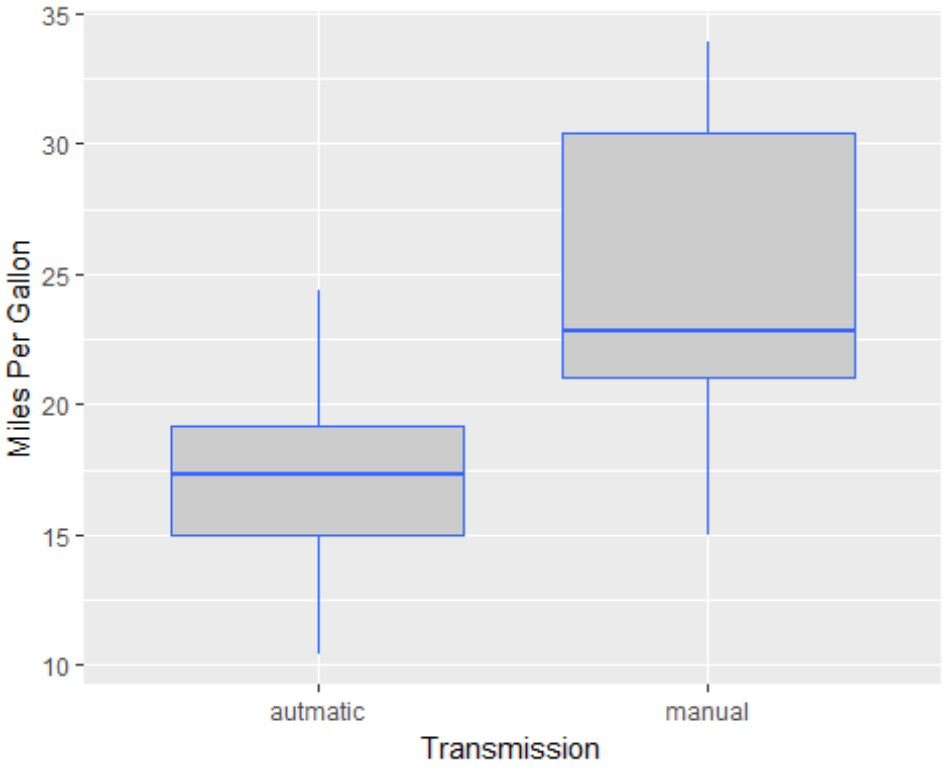
Here the variable of interest is the 'am' which means automatic/manual. Transmission (0 = automatic, 1 = manual)

```
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("autmatic", "manual")
```

Explore the dataset

Plotting a boxplot are able to see the difference in MPG between Manual and Automatic transmission

```
p <- ggplot(mtcars, aes(factor(mtcars$am), mpg))
p = p + geom_boxplot(fill = "grey80", colour = "#3366FF")
p = p + xlab("Transmission")
p = p + ylab("Miles Per Gallon")
p
```



As you see from the plot above, the average manual mpg is higher than the average automatic mpg. That shows Manual is mostly better mpg than Automatic.

Now the issue is that we are not accounting for other factors such as weight, number of cylinders etc... These factors are important in determining whether Transmission plays an overall factor in MPG and not just on a case by case bases.

Correlation amongst variables

```
##         wt        cyl       disp         hp       carb       qsec
## -0.8676594 -0.8521620 -0.8475514 -0.7761684 -0.5509251  0.4186840
##       gear         am         vs       drat        mpg
##  0.4802848  0.5998324  0.6640389  0.6811719  1.0000000
```

We see 'wt', 'cycl', 'disp', 'hp' and 'carb' are negatively correlated to 'mpg'. On the other hand 'am' is positively correlated with 'mpg' but in a moderate level. Maybe other factors need to be included so that 'am' strongly influence 'mpg'.

Models

Lineal Model

Here we will first test with a Linear Regression to gage the relationship between these varaibles

```
##
## 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
```

The results above show that with all things equal, the mpt for a manual car is 7.245 mpg better than automatic.Looking at the R^2 ~ 36%, we are not very confident on the robustness of this model. Only 36% of the variance is explained by the model.

Multivariate Regression

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## vs           0.31776281  2.10450861  0.1509915 0.88142347
## am           2.52022689  2.05665055  1.2254035 0.23398971
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
```

Looking at the above results we see that the P-value for the variables are fairly large. Now we will eleminate some of the variables one by one and fit the model to check the p-value.

This will help us find out which variable are significan in term of mpg.

OR we can fit a model with the variables wt(weight) and hp(horsepower) and see what sort of results we get. Here this stems from the thought made about, that transmission and mpg are also related to the weight and horsepower of the vehicle. Plus the above correlation results somewhat confirm our hypothesis.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + hp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     28 180.29  2    540.61 41.979 3.745e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Call:
## lm(formula = mpg ~ am + wt + hp, data = mtcars)
##
## Residuals:
```

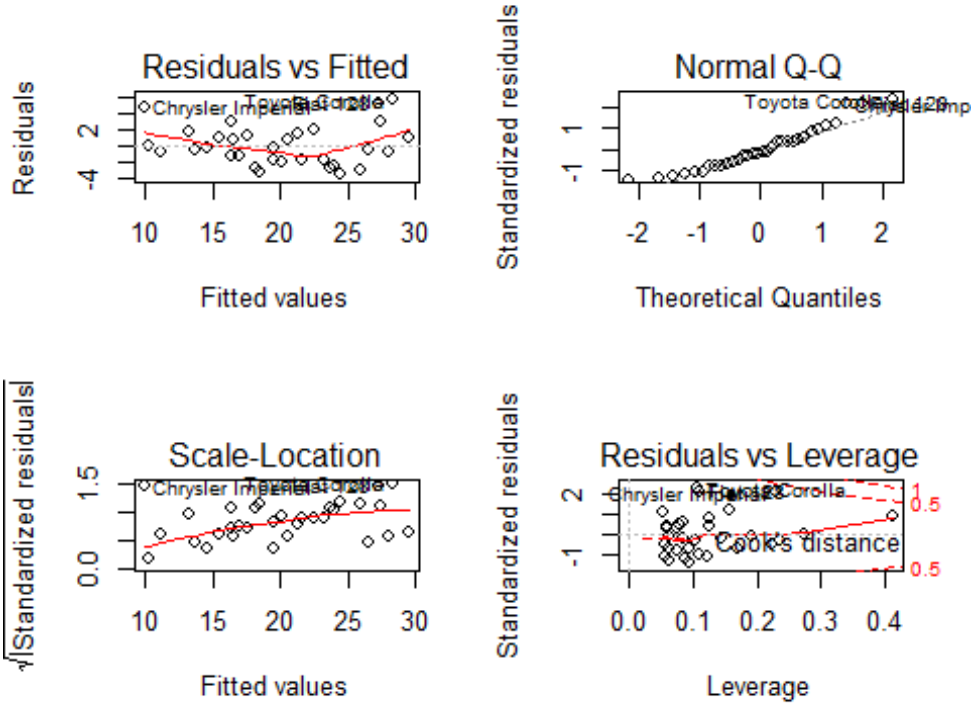
```
##     Min      1Q  Median      3Q     Max
## -3.4221 -1.7924 -0.3788  1.2249  5.5317
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.002875   2.642659  12.867 2.82e-13 ***
## am           2.083710   1.376420   1.514 0.141268
## wt          -2.878575   0.904971  -3.181 0.003574 **
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```

We can see that our R^2 increased quite a bit to about 84%(~83.99%). It means that model is robust as most of the variance is explained by model.

Also from the coefficients, we can say that on average, manual car have 2.084 more MPG than automatic. Actually this is a widely accepted notion in the car industry that manual, in general, have better mileage than automatic..All things equal.

Appendix

Residuals Plot



Full Anova analysis

Results mpg ~ am

```
anova(fit1, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + disp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 300.28  1    420.62 40.621 5.748e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Results mpg ~ am + disp

```
anova(fit1, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 271.36  1    449.53 48.041 1.285e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Results mpg ~ am + wt

```
anova(fit1, fit4)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 46.115 1.867e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```