# Motor Trend Analysis_Course(RM) Project

Pengfei LI

7/9/2020

## Executive Summary

### Introduction

This report aims to provide an analysis report for the magazine *Motor Trend*. The whole analysis is to explore the relationship between a set of variables and miles per gallon (MPG). Here we shall focus on two main issues,

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

To answer these issues, we start from the dataset of a collection of cars.

## ** Exploratory Analysis**

### Data Description

Here I analyze the dataset (mtcars) through several steps in the following section. This dataset could be loaded via R command *data(mtcars)*, in which the data was extracted from the 1974 Motor Trend US magazine. This dataset colltects fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

The dataset presents the mentioned data in the following columns.

- mpg Miles/(US) gallon
- cyl Number of cylinders
- disp Displacement (cu.in.)
- hp Gross horsepower
- drat Rear axle ratio
- wt Weight (1000 lbs)
- qsec 1/4 mile time
- vs Engine (0 = V-shaped, 1 = straight)
- am Transmission (0 = automatic, 1 = manual)
- gear Number of forward gears
- carb Number of carburetors

## Data Pre-processing

Here the procedure of data pre-processing is provided.

```
# Load the package
library(datasets)
library(ggplot2)

# Load the data
data(mtcars)

# Checking the dataset
dim(mtcars)
```

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

```
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 ...
```

```
names(cars)
```

```
## [1] "speed" "dist"
```

```
# Clean the data
mtcars$cyl  <- factor(mtcars$cyl)
mtcars$vs   <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am   <- factor(mtcars$am,labels=c("Auto","Manual"))

# View the data
summary(mtcars)
```

```
##       mpg        cyl         disp             hp             drat
##  Min.   :10.40   4:11   Min.   : 71.1   Min.   : 52.0   Min.   :2.760
##  1st Qu.:15.43   6: 7   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080
##  Median :19.20   8:14   Median :196.3   Median :123.0   Median :3.695
##  Mean   :20.09          Mean   :230.7   Mean   :146.7   Mean   :3.597
##  3rd Qu.:22.80          3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920
```

```
##  Max.   :33.90          Max.   :472.0   Max.   :335.0   Max.   :4.930
##        wt             qsec       vs          am     gear   carb
##  Min.   :1.513   Min.   :14.50   0:18   Auto  :19   3:15   1: 7
##  1st Qu.:2.581   1st Qu.:16.89   1:14   Manual:13   4:12   2:10
##  Median :3.325   Median :17.71                      5: 5   3: 3
##  Mean   :3.217   Mean   :17.85                             4:10
##  3rd Qu.:3.610   3rd Qu.:18.90                             6: 1
##  Max.   :5.424   Max.   :22.90                             8: 1
```

## Analysis

### Regression Analysis

Based on the data pre-processing section, the dataset splits into several variables. To explore the relationship between other variables and mpg, I set the regression model with all variables inside as the predictor, and then gradually select the most significant predictors for the final model. With the AIC algorithm, I use the forward selection and backward elimination method to run the model selection below.

```
start_model <- lm(mpg~.,data = mtcars)
final_model <- step(start_model,direction = "both")
```

```
summary(final_model)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

From the outcome of this process, the adjusted R squared value is 0.8401, showing that the above model could explain 84% variablity.

Then, I turn to anova analysis to check the regression model (**am** as the predictor) and the final model.

```
rg1 <- lm(mpg~am,data = mtcars)
anova(rg1,final_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result of anova process shows that **am** is the key predictor contributing to the accuracy of the model.

In the end, I go for the T-test to check whether two different transmission types are significantly different.

```
t.test(mpg~am, data = mtcars)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## 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 in group Auto mean in group Manual
##             17.14737             24.39231
```

Here the p-value is 0.001374, which means there is a significant difference between automatic cars and manual cars.

To quantify this difference, I run the following analysis.

```
summary(rg1)
```

```
##
## 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 outcome shows that the average MPG for automatic cars is 17.147 while manual is 7.2. However, the low $R^2$ value shows that the model only could explain 36% of variance. This infers that the analysis need the supplementary multivariate linear regression. The new model is as follows.

```
rg2 <- lm(mpg~am+cyl+disp+hp+wt, data = mtcars)
anova(rg1,rg2)
```

```
## 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
```

The result presents a low p-value, meaning that the supplementary model is better. Thus, I double-check the residuals, which is in the normal distribution and all are homoskedastic.

```
summary(rg2)
```

```
##
## 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
```

This result shows that 86.64% of the variance could be explained. It also shows that the difference between automatic cars and manual cars is 1.806 mpg.

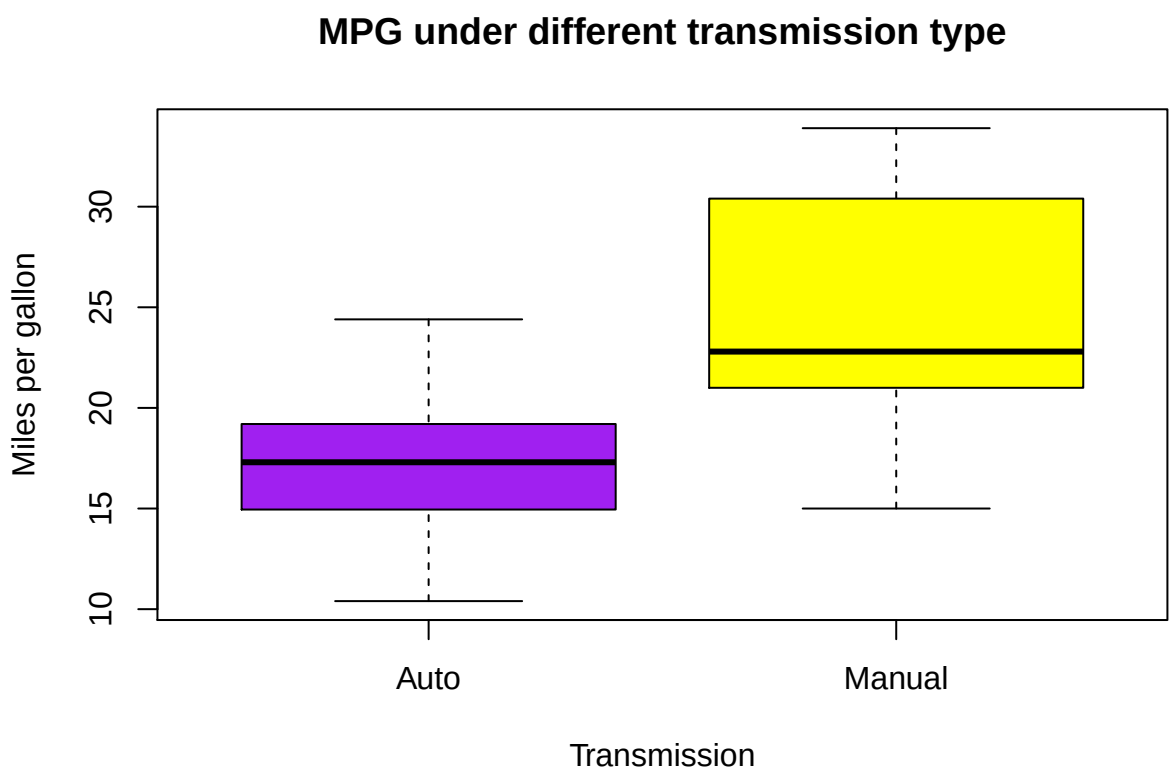
## Conclusion

Based on the analysis in upper section, the manual transmission is better for MPG. Moreover, it is **1.806 larger** than automatic transmission on MPG.

# APPENDIX

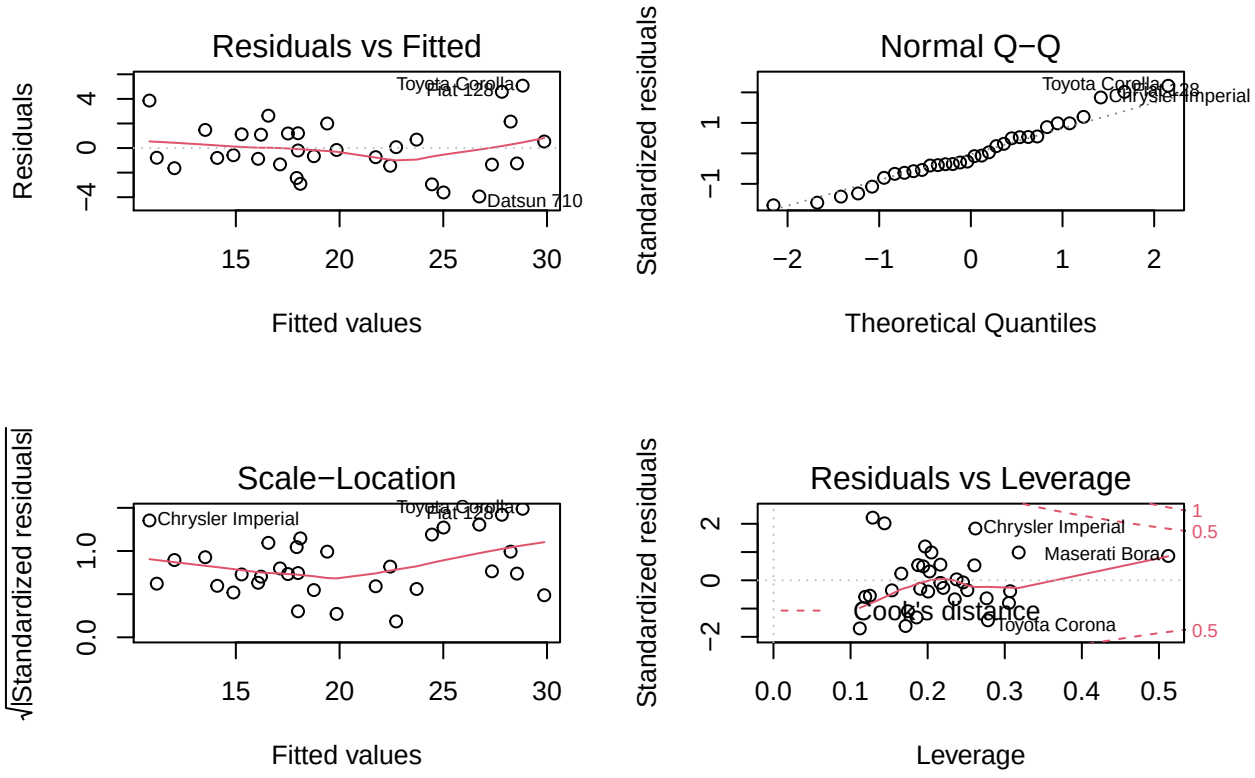
*Plot 1*

```
boxplot(mpg ~ am, data = mtcars, col = (c("purple","yellow")), ylab = "Miles per gallon", xlab = "Transm
```




*Plot 2*

```
par(mfrow = c(2,2))
plot(rg2)
```




*Plot 3*

```
pairs(mpg ~ ., data = mtcars,main= "Pairs Plot")
```