# Regression Models Course Project

## Executive summary:

This report is try to make some exploring of the relationships between the a set of variables and miles per gallon (MPG), especially try to address wheter automatic or manual transmission better for MPG, and quantifying how different is the MPG between automatic and manual transmissions. The data used in this analysis is come from the Motor Trend,a magazine about the automobile industry. The analysis is with the linear fitting models.The final model show that manual transimission is better for MPG. It shows there is a 2.936 miles mpg increase with manual cars under the statistics significance.

## Data loading and exploratory analysis

The data loading is with the data() function. After that, we can do a inital check for mtcars:

```
head(mtcars,n=2)
```

```
##               mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4      21   6  160 110  3.9 2.620 16.46  0  1    4    4
## Mazda RX4 Wag  21   6  160 110  3.9 2.875 17.02  0  1    4    4
```

The graph plot can show there are some difference in mpg between the manual and automatic transmissions(see appendix).

## Initial Quantifying analysis

To quantify address the question, we can do a T testing for automatic and manual groups:

```
ya <- mtcars$mpg[mtcars$am == 0]
ym <- mtcars$mpg[mtcars$am == 1]
t<-t.test(ym, ya, paired=FALSE)
t$p.value; t$conf.int[1:2];
```

```
## [1] 0.001374
```

```
## [1]  3.21 11.28
```

The p-value = 0.001374. The difference of two groups' 95% confidence interval is [3.209684~11.280194]. These show that the difference of mpg between automatical and manual is confidential statistics significance.

## Approaches to regression Models

The linear regression models will be built with two approaches: correlation coefficients based and stepwise algorithm based.

### Correlation coefficients based approach

This approach starts from checking mpg's correlations coefficients to other vaiables in the mtcars:

```
sort(cor(mtcars)[1,])
```

```
##      wt     cyl    disp      hp    carb    qsec    gear      am      vs 
## -0.8677 -0.8522 -0.8476 -0.7762 -0.5509  0.4187  0.4803  0.5998  0.6640 
##    drat     mpg 
##  0.6812  1.0000
```

We can see that the absoulte value of correlation coeficients from high to low are: wt,cyl,disp and hp. As the first trial, we can fit linear model with these variables. But we also noticed that there is a correlation between cyl and disp which is 0.90203287 from the complted variables correlation coefficients matrix, so disp can be droped from model. Thus we can get the following possible models:

```
cormodel1 = lm(mpg ~ factor(am), data = mtcars)
cormodel2 = update(cormodel1, mpg ~ factor(am) + wt)
cormodel3 = update(cormodel2, mpg ~ factor(am) + wt + cyl)
cormodel4 = update(cormodel3, mpg ~ factor(am) + wt + cyl + hp)
```

The P-values of model base on anova analysis are:

```
anova(cormodel1, cormodel2, cormodel3, cormodel4)$Pr
```

```
## [1]        NA 5.390e-09 9.165e-04 7.855e-02
```

The results show the cormodel4's p-value bigger than the statistical signifiance level=0.0.5, so the best fitted model is cormodel3. The coefficients of it are:

```
## (Intercept) factor(am)1          wt         cyl 
##     39.4179      0.1765     -3.1251     -1.5102
```

The R-squared is 0.8303, after adjusting, R-Squareted becomes to 0.8122.This means this model can capture 83% of total variance and the adjusted captured rate is 81.2%.

### Approach with stepwise algorithm function step()

The second approach is with the stepwise algorithm function step(). The result is:

```
stepmodel = step(lm(data = mtcars, mpg ~ .),trace=0, steps=10000)
summary(stepmodel)$coefficients
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    9.618     6.9596   1.382 1.779e-01
## wt            -3.917     0.7112  -5.507 6.953e-06
## qsec           1.226     0.2887   4.247 2.162e-04
## am             2.936     1.4109   2.081 4.672e-02
```

We can see that step function selects the wt, qsec and am as the model variables. The r-squared is 0.8497. The adjusted r-squared is 0.8336.This is better than the first approach, which shows there are some strong correlationship between first approach's model predictors. So this model was selected as the final predictiong model.

## Conclussion

With above approach, the final selected fitting model can have 85% total variance capturing rate, the adjusted rate is 83.4%. The final model selected the wt, qsec and am as the model variables to predict mpg. The coefficients are: -3.917, 1.226 and 2.936. From these coefficents, we ccould see when the weight increased 1000 lbs, the mpg decreased, -3.917 miles, when the qsec increase 1 second, the mpg increasing 1.226 miles. More important, the fitted model shows that manual cars has a 2.936 miles increase of the mpg under statistic significance.

## Appendix

In this appendix, we will show the plots of mpg distribution and the residual and diagnostic plots for the studied regression models.

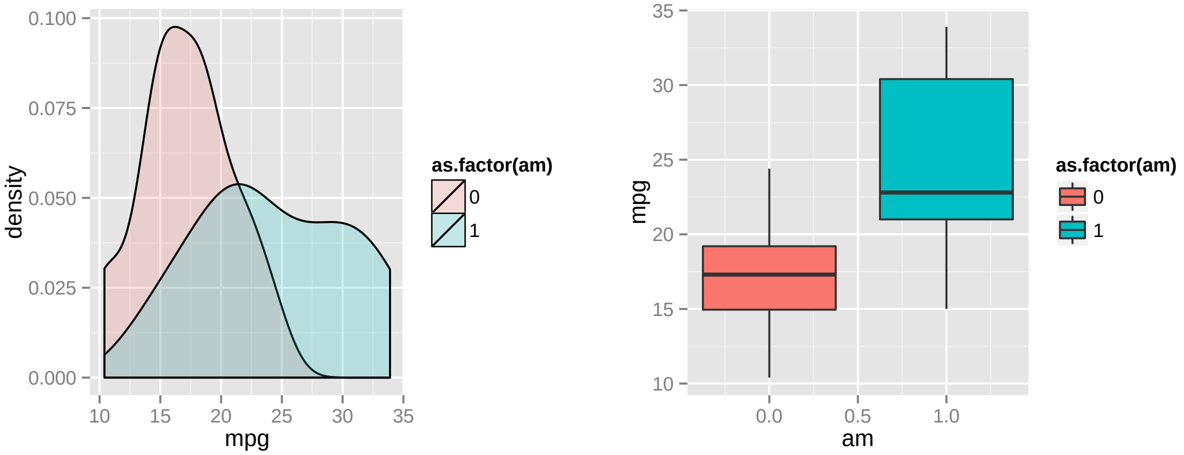
- The graph to show mpg difference between automatic and manual transmission:

```
library(ggplot2)
library(gridExtra)
```

```
## Loading required package: grid
```

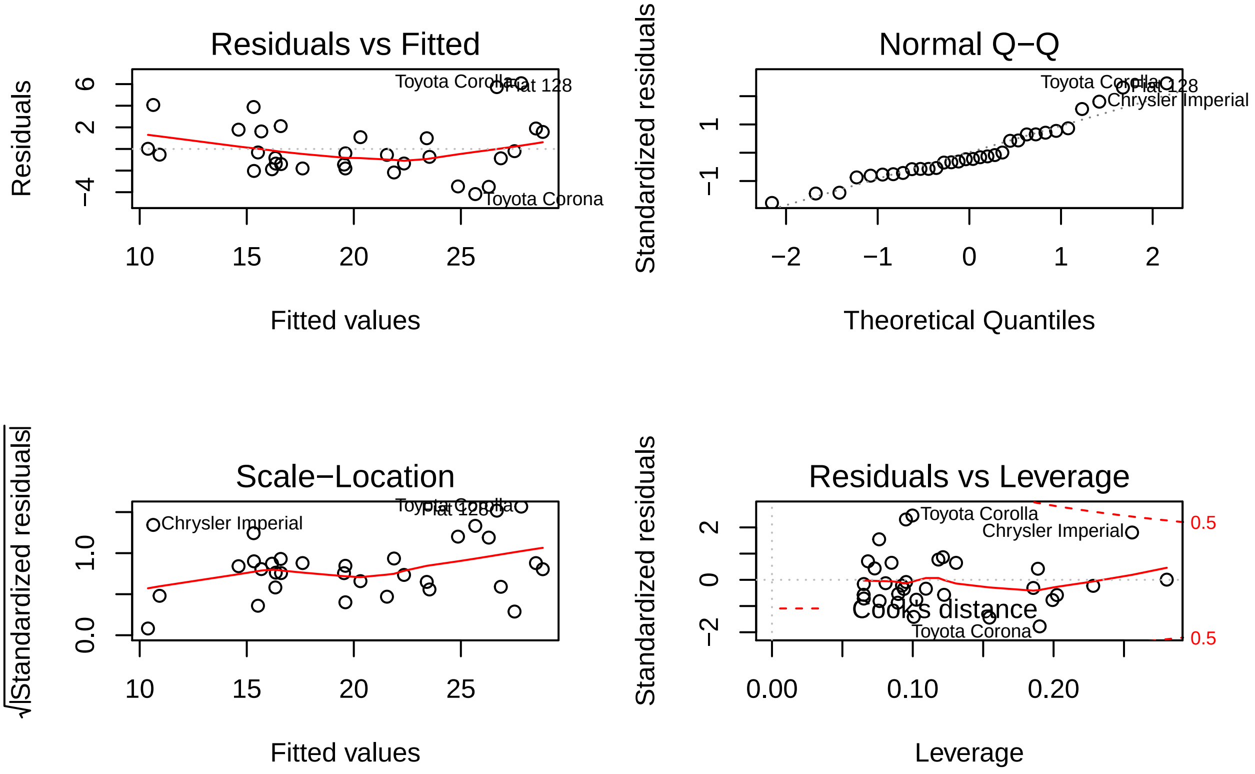
```
# density distribiution
g1<-ggplot(mtcars, aes(mpg, fill = as.factor(am))) + geom_density(alpha = 0.2)
 # box plotting
g2<-ggplot(mtcars, aes(x=am, y=mpg, fill=as.factor(am))) + geom_boxplot()

grid.arrange(g1,g2, ncol=2)
```



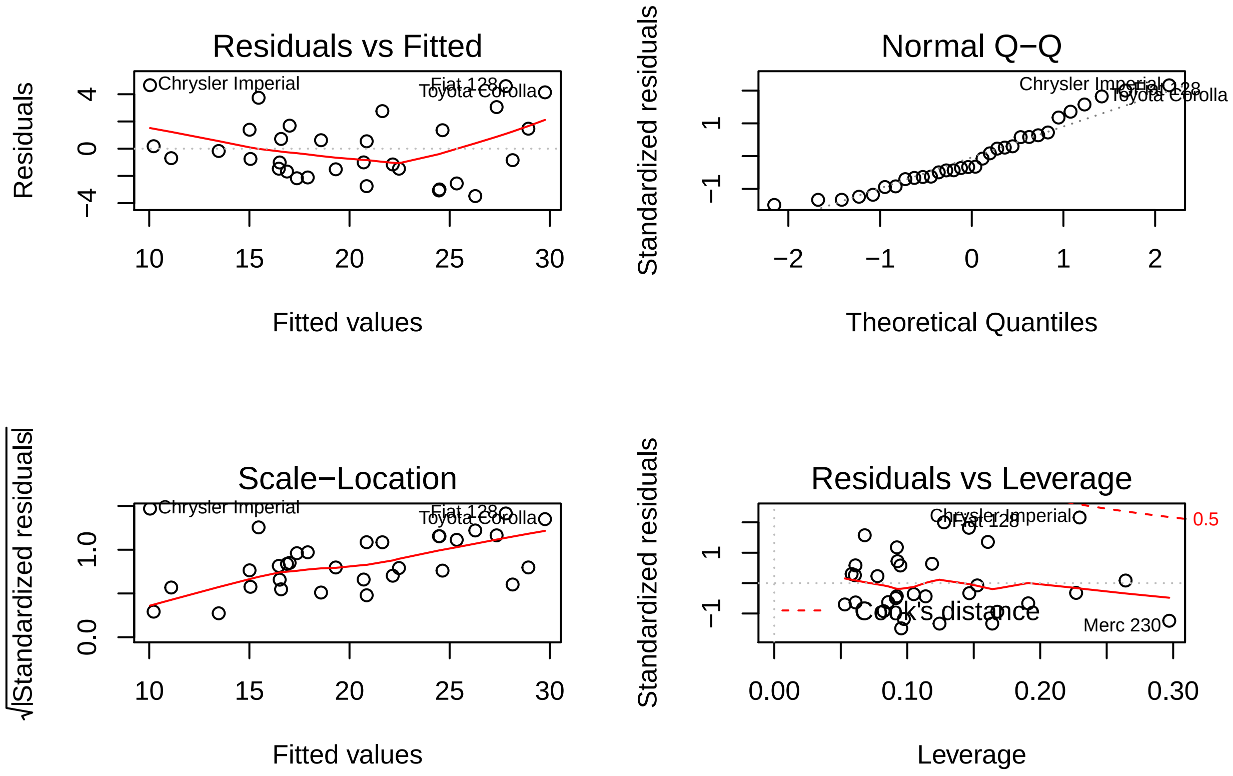
- The correlation coefficients based model:

```
par(mfrow= c(2,2))
plot(cormodel3)
```



- The final Stepwise algorithm based model:

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


Residuals vs Fitted
Residuals
4
0
−4
10
15
20
25
30
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
1
−1
−2
−1
0
1
2
Theoretical Quantiles
Chrysler Imperial
Fiat 128
Toyota Corolla
Scale−Location
√|Standardized residuals|
1.0
0.0
10
15
20
25
30
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Residuals vs Leverage
Standardized residuals
1
−1
0.00
0.10
0.20
0.30
Leverage
Chrysler Imperial
Fiat 128
Merc 230
Cook's distance
0.5