# Does Transmission Type affect Mileage?

*Wang Xin*

*2016-4-25*

## Executive Summary

In responding of the insterests of *Motor Trend*, a magazine about the automobile industry, this report explores the relationship between a set of variables and mile per gallon(MPG), trying to answer two questions:

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

**Notice**: This work is part of Coursera regression model courese's assignment. The results could be wrong or outdated.

## Load and Prepare Data

```
data("mtcars")
dt <- mtcars
dt$am<-factor(dt$am,levels=c(0,1),labels=c('Automatic','Manual'))
```

## Exploratory Data Analyse #1

By simply boxplot mpg by different transmission type (See Appendix #1), it shows that manual got higher mpg than automatic.

## Building Model #1

```
fit1 <- lm(mpg ~ am, data = dt)
summary(fit1)
```

It shows that an automatic car with 17.147 mpg, gets 7.245 mpg more if it is manual transimission. The Adjusted R-squared vaue is 0.3385, which meacs the model explain only 34% of the MPG varibles. So, more variables need to be indroduced into the model.

## ANOVA test

Implement ANOVA test to determine those most relevant variables.

```
dt_anova <- aov(formula = mpg ~ ., data = dt)
summary(dt_anova)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## cyl          1  817.7   817.7 116.425 5.03e-10 ***
## disp         1   37.6    37.6   5.353  0.03091 *  
## hp           1    9.4     9.4   1.334  0.26103    
## drat         1   16.5    16.5   2.345  0.14064    
## wt           1   77.5    77.5  11.031  0.00324 ** 
## qsec         1    3.9     3.9   0.562  0.46166    
## vs           1    0.1     0.1   0.018  0.89317    
## am           1   14.5    14.5   2.061  0.16586    
## gear         1    1.0     1.0   0.138  0.71365    
## carb         1    0.4     0.4   0.058  0.81218    
## Residuals   21  147.5     7.0                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results suggest that disp(displacement), wt(Weight) and cyl(Number of cylinders) are those top 3 significant variables.

## Exploratory Data Analyse #2

Visiualize more variables (See appendix #2) and it shows that disp,wt,cyl all have stronger relationship with mpg.

## Building Model #2

Build new model with the other 3 varibales:

```r
fit2 <- lm(mpg ~ cyl + disp + wt + am, data = dt)
summary(fit2)
```

```
## 
## Call:
## lm(formula = mpg ~ cyl + disp + wt + am, data = dt)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -4.318 -1.362 -0.479  1.354  6.059 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 40.898313   3.601540  11.356 8.68e-12 ***
## cyl         -1.784173   0.618192  -2.886  0.00758 ** 
## disp         0.007404   0.012081   0.613  0.54509    
## wt          -3.583425   1.186504  -3.020  0.00547 ** 
## amManual     0.129066   1.321512   0.098  0.92292    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.642 on 27 degrees of freedom
## Multiple R-squared:  0.8327, Adjusted R-squared:  0.8079 
## F-statistic: 33.59 on 4 and 27 DF,  p-value: 4.038e-10
```

It shows that disp is not significant for the model. So I drop this varible.

## Building Model #3

```r
fit3 <- lm(mpg ~ cyl + wt + am, data = dt)
summary(fit3)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + wt + am, data = dt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1735 -1.5340 -0.5386  1.5864  6.0812
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.4179     2.6415  14.923 7.42e-15 ***
## cyl          -1.5102     0.4223  -3.576  0.00129 **
## wt           -3.1251     0.9109  -3.431  0.00189 **
## amManual      0.1765     1.3045   0.135  0.89334
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.612 on 28 degrees of freedom
## Multiple R-squared:  0.8303, Adjusted R-squared:  0.8122
## F-statistic: 45.68 on 3 and 28 DF,  p-value: 6.51e-11
```

This model has Residual standard error as 2.612 on 28 degree of freedom, and Adjusted R-squared value os 0.8122, which means it can explain 81% of the MPG variables. All of the coeddicients are significant at 0.05 significant level. This is a good one.

## Residual plot and diagnostics

According to the residual plots (See Appendix #3):

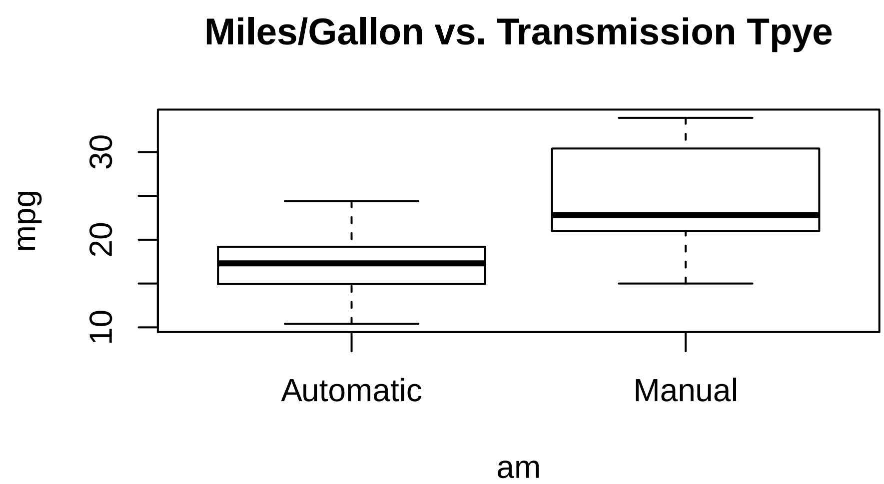
1. The residuals vs fitted plot shows no consistent pattern
2. The points lie closely to the line, so the residuals are normally distributed
3. The scale-location plot confirms the constant variance assumption
4. No outliers are present.

## Conclusion

In this dataset, although manual vehicles achieve more mpg that automatics, thransmission type is more a good factor to predict mpg. The number of cylinders and the gross weight are much more directly linked to fuel usage.

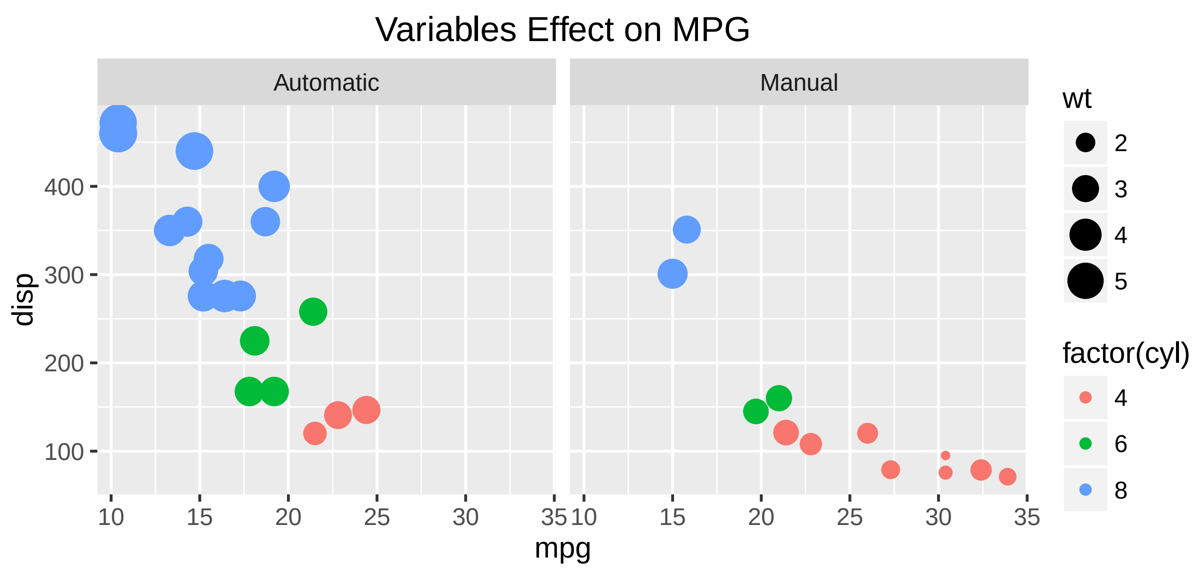
## Appendix #1 - botplot

```
plot(mpg ~ am, data = dt, main = "Miles/Gallon vs. Transmission Tpye")
```




## Appendix #2 - plot more variables

```
library(ggplot2)
qplot(x = mpg, y = disp, data = dt,
      size = wt, color = factor(cyl), facets=.~am,
      main='Variables Effect on MPG')
```




## Appendix #3 - residual plots

```r
par(mfrow = c(2,2))
plot(fit3)
```