Regression Models Project

Eric VACHON

June 2015

Summary

In this project, we work for a magazin about automobile industry : Motor Trend. They ask us two questions : - Question 1 : Is an automatic or manual transmission better for MPG ? - Question 2 : Quantify the MPG difference between automatic and manual transmissions ?

And to answer this two questions, we are going to use the R dataset "mtcars", data which was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

Dataset : mtcars

In first look at the dataset : data frame with 32 observations on 11 variables : **mpg** (Miles/(US) gallon), **cyl** (Number of cylinders), **disp** (Displacement), **hp** (Gross horsepower), **drat** (Rear axle ratio), **wt** (Weigh), **qsec** (1/4 mile time), **vs** (V/S), **am** (Transmission (0 = automatic, 1 = manual)), **gear** (Number of forward gears) and **carb** (Number of carburetors).

```
##                mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4     21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710    22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
```

Question 1 : Is an automatic or manual transmission better for MPG ?

First we must analyse the data for automatic and manual transmmission (cf plot 1 in Appendix) and run a t.test between this 2 subsets :

```
##
##  Welch Two Sample t-test
##
## data:  mtcars[mtcars$am == 0, "mpg"] and mtcars[mtcars$am == 1, "mpg"]
## 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 of x mean of y
##  17.14737  24.39231
```

The p value is less than 0.05, 0 is not include in the 95 percent confidence interval, the mean of automatic transmission (17.15) is less than the manual (24.39), and (plot 1) we can see that automatic transmission consume globally less than the manual.

So we can conclude that : **automatic transmission is better for MPG**.

Question 2 : Quantify the MPG difference between automatic and manual transmissions ?

In this section we are going to use in first a simple model of regression : **Is type of transmission can explain the mpg ?**

```
simpleLM<-lm(mpg~am, data = mtcars)
summary(simpleLM)$r.square
```

```
## [1] 0.3597989
```

```
summary(simpleLM)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am           7.244939   1.764422  4.106127 2.850207e-04
```

```
confint(simpleLM)
```

```
##                 2.5 %   97.5 %
## (Intercept) 14.85062 19.44411
## am           3.64151 10.84837
```

So R-squared is low (0.3597989), p value is less than 0.05, 0 is not include in the confidence interval : So 35.98% of the variation of mpg can be explain by the type of transmission … **we can find a better model**

So we are going to find a another variables to quantify better the MPG difference between automatic and manual transmissions.

First look at the correlation between all the 11 variables (cf plot 3 in the appendix), and we must find the better combinaison of variables to fin the best model. To do this we are going to use the R-step function :

```
multipleLM <- step(lm(data = mtcars, mpg ~ .),trace=0)
summary(multipleLM)$r.square
```

```
## [1] 0.8496636
```

```
summary(multipleLM)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## am           2.935837  1.4109045  2.080819 4.671551e-02
```

```
confint(multipleLM)
```

```
##                   2.5 %    97.5 %
## (Intercept) -4.63829946 23.873860
## wt          -5.37333423 -2.459673
## qsec         0.63457320  1.817199
## am           0.04573031  5.825944
```

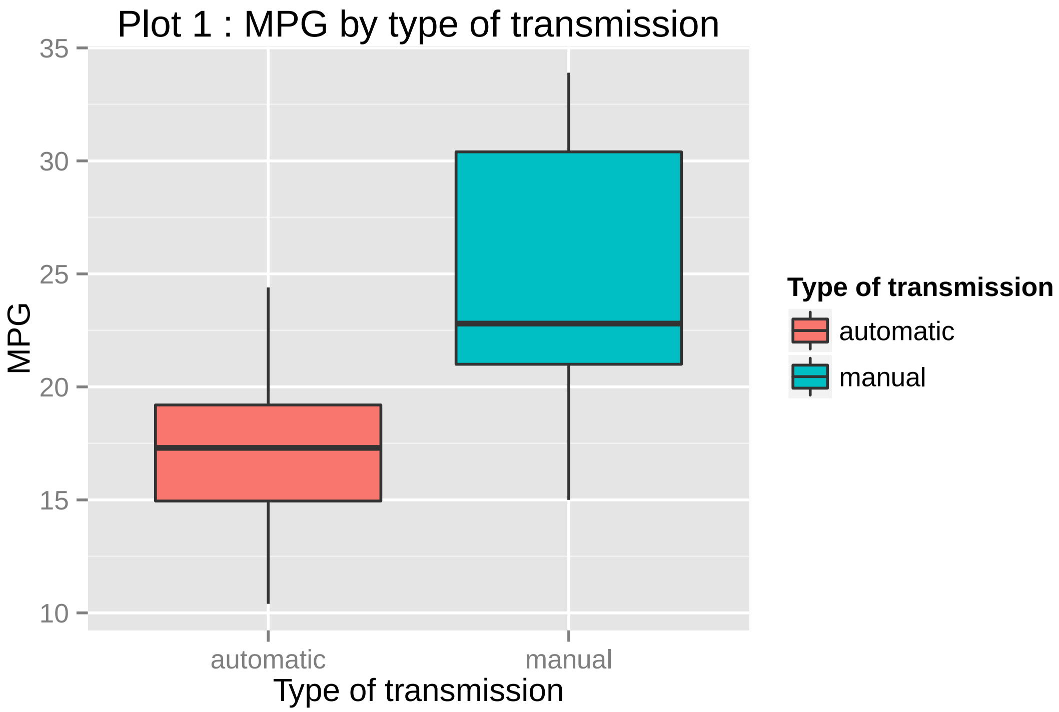
Now we have a model that 84.97% of the variation of mpg can be explain by : the weight + 1/4 mile time + type of transmission.
Now look at the normality of residuals (plot 4) : the residuals seem to be normal (an anova help too).
In fact this model explain that :
1. We estimate an expected 1.22 miles/gallon increase in fuel consumption for every 1 seconde increase in speed to do 1/4 miles in holding the remaining variables constan t(wt and am) : plot 3.
2. We estimate an expected 3.91 miles/gallon decrease in fuel consumption for every lb/1000 increase in weight in holding the remaining variables constan t(qsec and am) : plot 4.
3. We estimlate in mean an expected increase by 2.9 mpg the fuel consumption if we use a manual transmission instead of an automatic transmission (plot 3 and 4).

Appendix

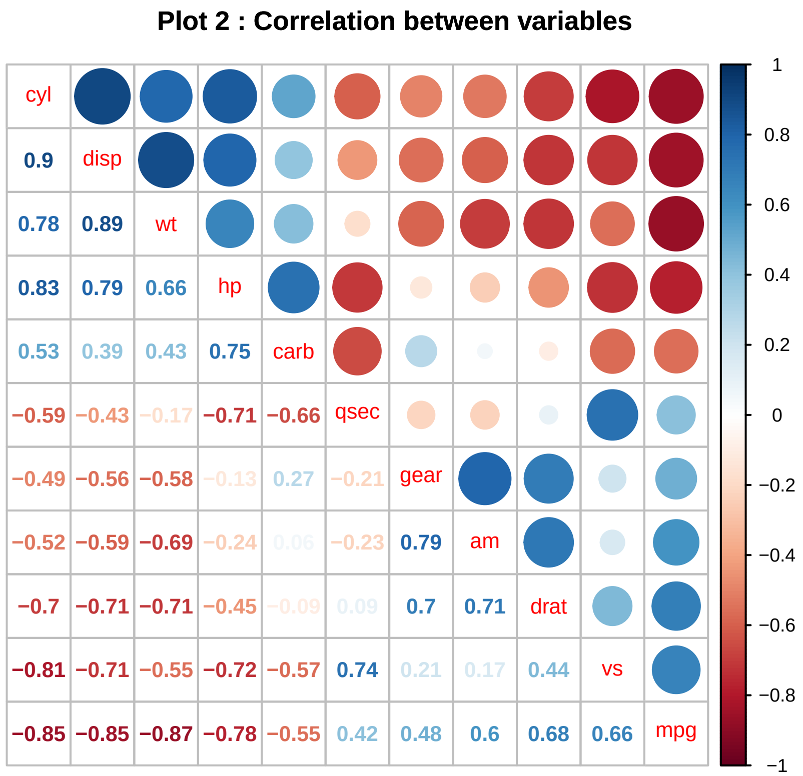
Libraries

```
library(ggplot2);library(corrplot);library(plyr)
mtcars$TypeTransmission <- factor(mtcars$am)
mtcars$TypeTransmission <- revalue(mtcars$TypeTransmission, c("0"="automatic", "1"="manual"))
```

```
ggplot(mtcars, aes(x=TypeTransmission, y=mpg, fill=TypeTransmission)) + geom_boxplot() +
       labs(x="Type of transmission",y = "MPG",title = "Plot 1 : MPG by type of transmission")+
       guides(fill=guide_legend(title="Type of transmission"))
```

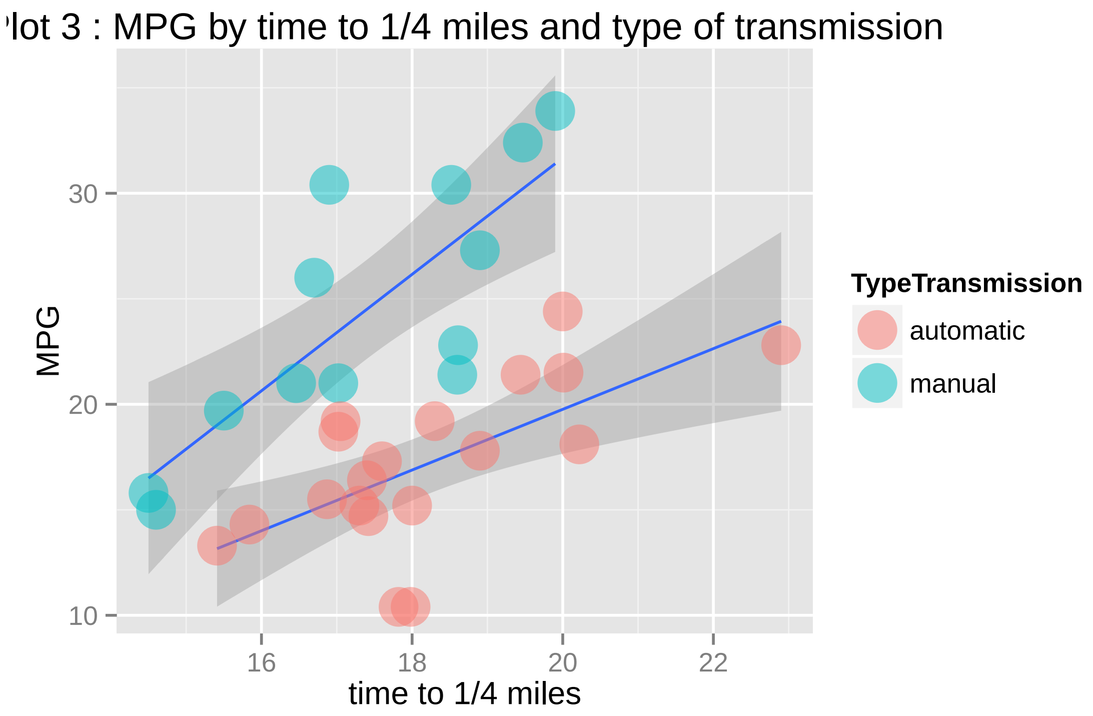


```
par(cex = 0.7)
corrplot.mixed(cor(mtcars[,c(1:11)]),order = "FPC",mar=c(0,0,1,0),
               title="Plot 2 : Correlation between variables")
```



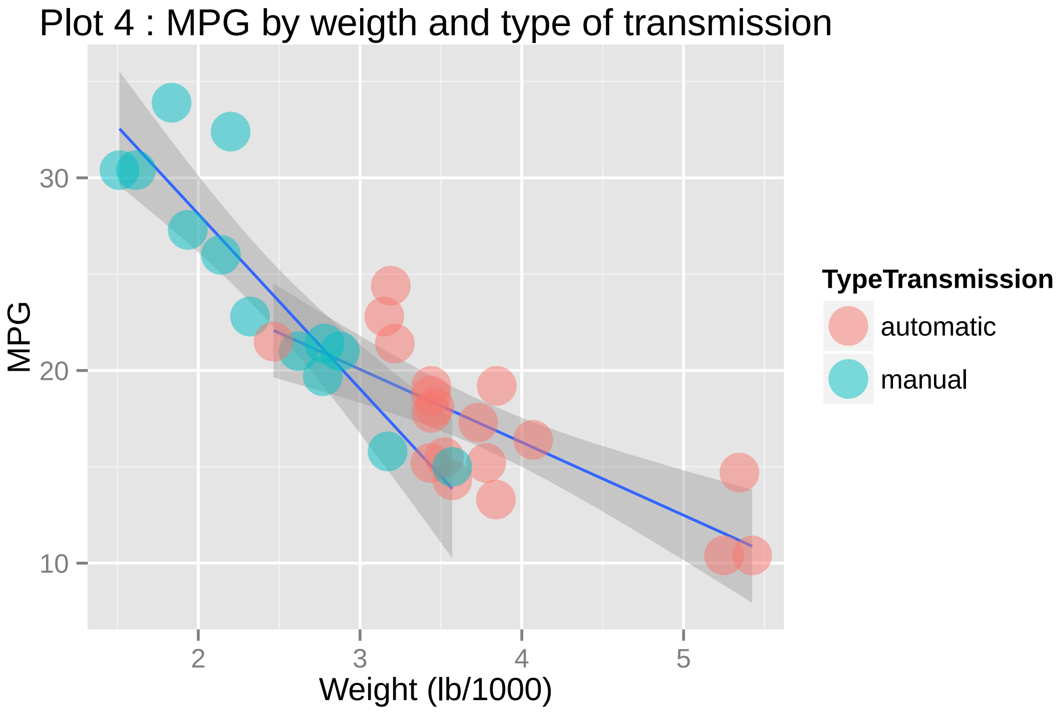
Plot 2 : Correlation between variables

```
ggplot(mtcars, aes(x = qsec, y = mpg, group =
  TypeTransmission))+geom_smooth(method = "lm")+
       geom_point(size = 7, aes(colour = TypeTransmission), alpha=0.5) +
       labs(x = "time to 1/4 miles",y = "MPG", title = "Plot 3 : MPG by time to 1/4 miles and type of transmi
```

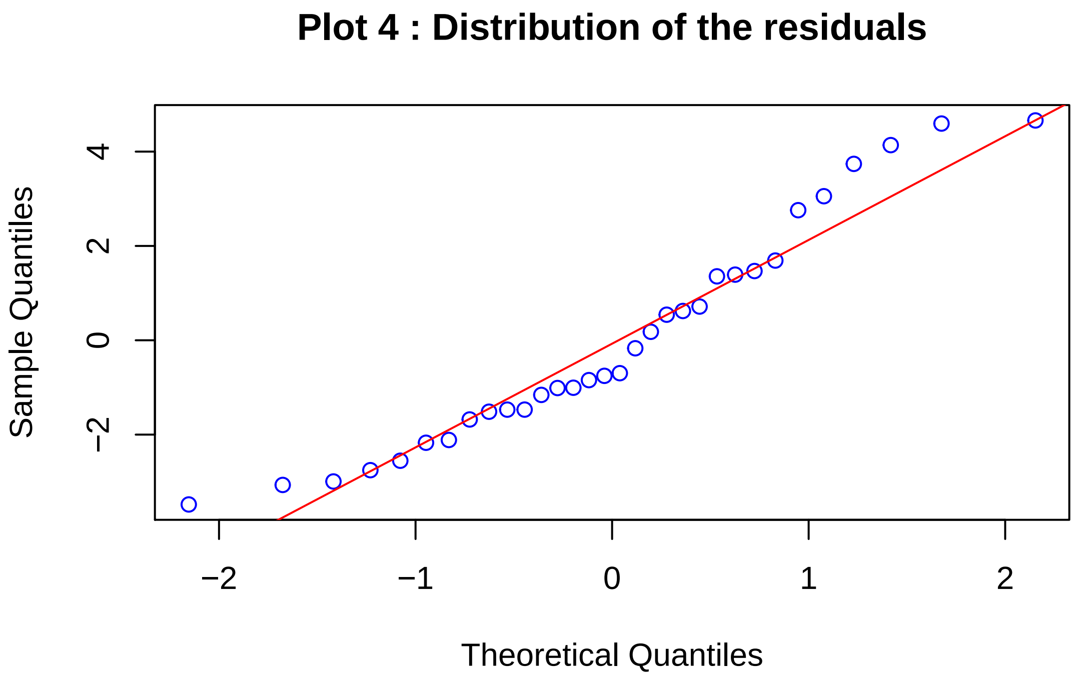


'lot 3 : MPG by time to 1/4 miles and type of transmission

```
ggplot(mtcars, aes(x = wt, y = mpg, group=TypeTransmission))+geom_smooth(method = "lm")+
       geom_point(size = 7, aes(colour = TypeTransmission), alpha=0.5) +
       labs(x = "Weight (lb/1000)",y = "MPG", title = "Plot 4 : MPG by weigth and type of transmission")
```



```
qqnorm(resid(multipleLM), col = 'blue',main = 'Plot 4 : Distribution of the residuals')
qqline(resid(multipleLM), col = 'red')
```



End of the document