Some Analysis on MTCARS dataSet

Executive summary

In this document, we are going to study and to analyze the possible relation between the transmission type of a car and its miles per gallon. In our study, we will perform the following steps :
1. loading data and getting a preview
2. exploring data
3. fitting a regressive model
4. checking if our model is robust when adding new variables.
5. trying to conclude based on the results of our tests.

Preview

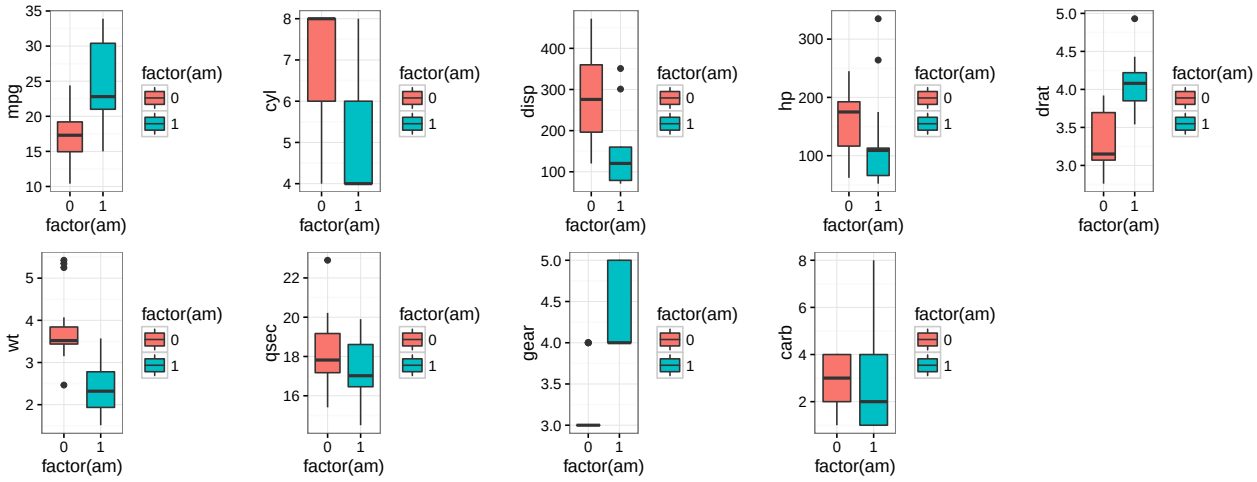
The data set we are going to study is "mtcars".

```
data("mtcars")
head(mtcars,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
```

Exploratory Analysis

let's divide and compare data into two groups of cars by the type of transmission: "auto" and "manuel" transmission.



We can see that mpg are differents in the subgroups of transmission mode. However,we can see also that generally the two groups don't have the same characteristics : This means that we have to be careful before stating that only type transmission affects the mpg -> there might other factors that cause the mpg changes.

Inference

Let's check if subgrouping by type of transmission creates samples with different means :

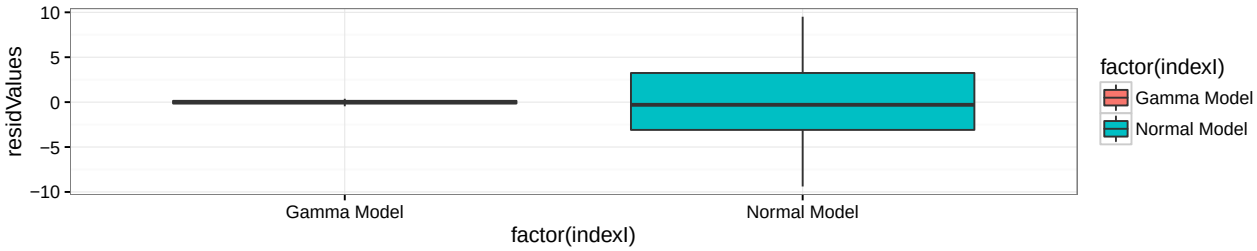
```
t.test(mtcars[mtcars$am==1,]$mpg,mtcars[mtcars$am==0,]$mpg,paired = FALSE,var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars[mtcars$am == 1, ]$mpg and mtcars[mtcars$am == 0, ]$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:
##   3.209684 11.280194
## sample estimates:
## mean of x mean of y
##  24.39231  17.14737
```

with a p-value of 0.001374 , we are pretty sure that the two subgroups have different means and we can infer it.

Fitting a model : mpg Vs am

We know that mpg is positive and isn't always an integer. Naturally, we can't think of using the Gamma model. Lets compare the residuals of a normal(classical)linear model with a generalised linear model using a gamma family.



Gamma model seems to fit greatly the data.

What if .. ? other variables might explain the difference

Let's take a look at the coefficients of our gamma model :

```
## factor(am)0 factor(am)1
##  0.05831799  0.04099653
```

Let's see what happens if we add the variable weight :

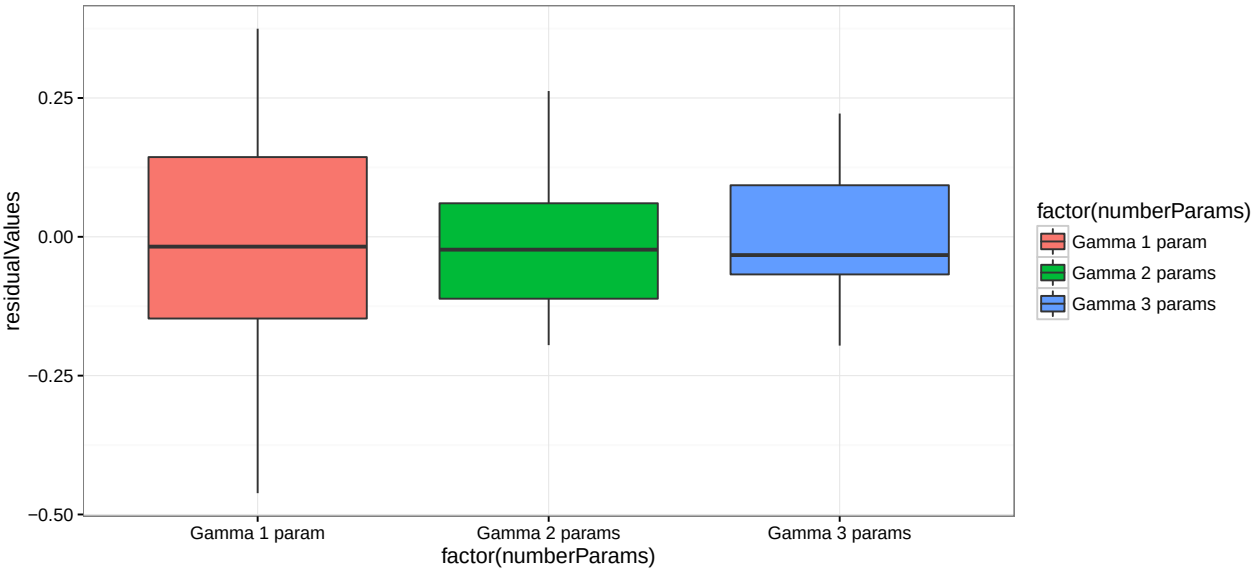
```
## factor(am)0 factor(am)1          wt
## 0.001362639 0.005260111 0.015650230
```

now let's add weight and cylinders too :

```
##  factor(am)0  factor(am)1           wt          cyl
## 0.0002845236 0.0027332157 0.0102939618 0.0030814591
```

Yes , we can see that including “cyl” and “wt”" has changed radically the coefficients of the factors of transmission mod.

Let’s compare the residuals of the three models




```
## [1] 1.7601419 0.5102102 0.3881136
```

as we can see, introducing the new variables “cyl” and “wt” has reduced the residuals and improved the quality of our model.

let’s compare the three models using anova :

```
## Analysis of Deviance Table
##
## Model 1: mpg ~ factor(am) - 1
## Model 2: mpg ~ factor(am) + wt - 1
## Model 3: mpg ~ factor(am) + wt + cyl - 1
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        30    1.76014
## 2        29    0.51021  1   1.2499 < 2.2e-16 ***
## 3        28    0.38811  1   0.1221  0.003208 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova confirms that the additional variables bring information.

Conclusion

The analysis above show that we have no evidence that the transmission mode is responsable of mpg value. To make a more valuable analysis , It might be better to have data that compares cars that differ only in mode of transmission and are similar in other characterisics(than mpg and am).

Appendix

1) To produce the exploratory figures , i used the following code:

```
part1<-subset(mtcars,am==0)
part2<-subset(mtcars,am==1)
# please uncomment the four following lines
#library(ggplot2)
#library(gridExtra)
#library(grid)
#library(lattice)

x1<-ggplot(aes(y = mpg, x = factor(am),fill=factor(am)), data = mtcars) + geom_boxplot()+ theme_bw()
x2<-ggplot(aes(y = cyl, x = factor(am),fill=factor(am)), data = mtcars) + geom_boxplot()+ theme_bw()
x3<-ggplot(aes(y = disp, x = factor(am),fill=factor(am)), data = mtcars) + geom_boxplot()+ theme_bw()
x4<-ggplot(aes(y = hp, x = factor(am),fill=factor(am)), data = mtcars) + geom_boxplot()+ theme_bw()
x5<-ggplot(aes(y = drat, x = factor(am),fill=factor(am)), data = mtcars) + geom_boxplot()+ theme_bw()
x6<-ggplot(aes(y = wt, x = factor(am),fill=factor(am)), data = mtcars) + geom_boxplot()+ theme_bw()
x7<-ggplot(aes(y = qsec, x = factor(am),fill=factor(am)), data = mtcars) + geom_boxplot()+ theme_bw()
x8<-ggplot(aes(y = gear, x = factor(am),fill=factor(am)), data = mtcars) + geom_boxplot()+ theme_bw()
x9<-ggplot(aes(y = carb, x = factor(am),fill=factor(am)), data = mtcars) + geom_boxplot()+ theme_bw()
 # please , uncomment the following line :
#grid.arrange(x1, x2, x3, x4,x5,x6,x7,x8,x9, ncol=5)
```

2)To fit regressive models , i used the following code :

```
mdlN<-lm(mpg~factor(am)-1,data=mtcars)
mdlG<-glm(mpg~factor(am)-1,data=mtcars,family=Gamma)

residG<-resid(mdlG)
residN<-resid(mdlN)
indexI<-c(rep("Normal Model",length(resid((mdlG)))),rep("Gamma Model",length(resid((mdlG)))))

residValues<-c(residN,residG)

df<-data.frame(residValues,indexI)
# please , uncomment the following lines
#ggplot(aes(y = residValues, x = factor(indexI),fill=factor(indexI)), data = df) + geom_boxplot()+ them
```

3) To add new variables into model and use anova , i used the following code :

```
mdlG2<-glm(mpg~factor(am)+wt -1,data=mtcars,family=Gamma)
mdlG3<-glm(mpg~factor(am)+wt + cyl -1,data=mtcars,family=Gamma)
# please ,  uncomment the following lines
#mdlG$coefficients
#mdlG2$coefficients
#mdlG3$coefficients
#anova(mdlG,mdlG2,mdlG3,test="Chisq")
```