Lecture 4 : Methods for Regression

Anova 1 factor

K. Meziani

Dauphine | PSL
UNIVERSITÉ PARIS

Example under R

Atherosclerosis is the leading cause of death for men after age 35 and for women after age 45 in most developed countries. It is a thickening and a loss of elasticity of the internal walls of the arteries, one of the consequences of which is myocardial infarct. The arterial wall consists of three layers respectively from the arterial lumen: the intima, the media and the adventitia. The thickness of the intima-media is a recognized marker of atherosclerosis. It was measured ultrasonically on a sample of 110 subjects in 1999 at the Bordeaux University Hospital. Information on the main risk factors was also collected, including on smoking and alcohol consumption among patients:

- `Smoking status` is measured in 3 modalities: `0="do not smoke"`, `"1=quit smoking"`, `2="smoke"`.
- `Consumption of alcohol` is measured in 3 modalities: `0="do not drink"`, `1="drink occasionally"`, `2="drink regularly"`.

We want to conduct an analysis of the influence of these factors on the thickness of the intima-media.

Packages

```
library(carData)
library(car)
library(knitr)
## library multcomp necessite
library (survival)
library(MASS)
library(TH.data)
library(mvtnorm)
library(multcomp)
```

Section 1

I. Upload the dataset

Upload the dataset

Consider in this section an Anova single factor model. Consider the example of the influence of the `Consumption of alcohol` on the thickness of the intima-media.

We recall that the Consumption of alcohol :`alcool` has $J = 3$ modalities

"0" ="do not drink"

"1" ="drink occasionally"

"2" ="drink regularly"

Read the data and select a subsample

First load and read the dataset.

```
Marqueur = read.table("Intima_Media.txt", header=T,
                      sep=" ", dec=",")
names(Marqueur)
```

```
## [1] "SEXE"   "AGE"    "taille" "poids"  "tabac"  "paqan"  "SPORT"  "mesure"
## [9] "alcool"
```

```
marqueur=Marqueur[,c(8,9)]
names(marqueur)
```

```
## [1] "mesure" "alcool"
```

Rename some modalties

For sake of simplicity in the interpretation, we change the name of the modalities of the variable `alcool`

```
marqueur$alcool=replace(marqueur$alcool,marqueur$alcool==0,"NotDrink")
marqueur$alcool=replace(marqueur$alcool,marqueur$alcool==1,"DrinkOcc")
marqueur$alcool=replace(marqueur$alcool,marqueur$alcool==2,"DrinkReg")
```

Check the nature of the features

Check that the variables have been correctly defined.

```
str(marqueur)
```

```
## 'data.frame':    110 obs. of  2 variables:
##  $ mesure: num  0.52 0.42 0.65 0.48 0.45 0.49 0.42 0.45 0.65 0.52 ...
##  $ alcool: chr  "DrinkOcc" "DrinkOcc" "NotDrink" "DrinkOcc" ...
```

Correction of the nature of the feature alcool

The variable alcool has not been correctly defined. Then, we have to declare it as a factor as follows.

```
marqueur$alcool=as.factor(marqueur$alcool)
str(marqueur$alcool)
```

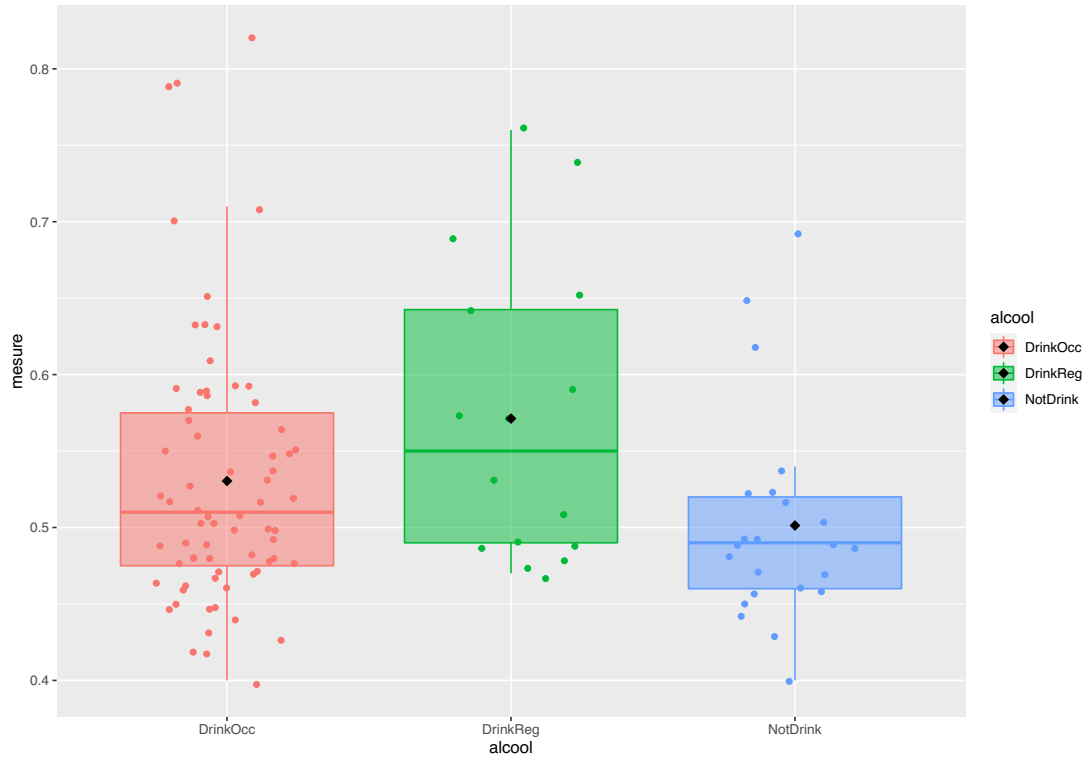
```
##  Factor w/ 3 levels "DrinkOcc","DrinkReg",..: 1 1 3 1 1 1 1 1 2 1 ...
```

Section 2

II. Descriptive analysis

Boxplots of the mesure per modality

```
library(cowplot); library(ggplot2)
ggplot(marqueur, aes(y=mesure, x=alcool,colour=alcool ,fill=alcool))+
geom_boxplot(alpha=0.5, outlier.alpha=0)+geom_jitter(width=0.25)+
stat_summary(fun=mean, colour="black", geom="point",shape=18, size=3)
```



Comments on the used functions

- First underline that the black diamonds represent the empirical mean.
- In the function `geom_boxplot`, the argument `outlier.alpha=0` allows to not represent twice an outlier point (once with the function `geom_boxplot`, once with the function `geom_jitter`).
- The function `geom_jitter function` is used to represent points without overlapping (`width = 0.25` allows to manage the spacing of the points.)

Some resumes of the dataset

Display the number of modalities J of the factor . Display the n_j , $j = 1, \cdots, J$ the number of observations of the modality j. Note that, in this dataset, the plan is unbalanced. Here, $\overline{n}_1 = 71$, $\overline{n}_2 = 16$ and $\overline{n}_3 = 23$

```
J =length(levels(marqueur$alcool))
n_j =table(marqueur$alcool);
knitr::kable(n_j,col.names =c("Modalities","Counts"))
```

Modalities	Counts
DrinkOcc	71
DrinkReg	16
NotDrink	23

How is the plan ?

Display empirical mean (EM) per cell

Here, $\overline{Y}_{.1} = 0.5304225$, $\overline{Y}_{.2} = 0.57125$ and $\overline{Y}_{.3} = 0.5013043$.

```
moy_j=tapply(marqueur$mesure,list(Alcool=marqueur$alcool),
             mean,na.rm=TRUE)
knitr::kable(moy_j,col.names =c("Empirical mean"))
```

	Empirical mean
DrinkOcc	0.5304225
DrinkReg	0.5712500
NotDrink	0.5013043

Display other EM

The **EM of the EM by cell**: $\overline{\overline{Y}}_{..} = 0.5343256$

```
mean(moy_j)
```

```
## [1] 0.5343256
```

Display the **EM of the variable `mesure`**. Here, $\overline{Y}_{..} = 0.5302727$.

```
mean(marqueur$mesure)
```

```
## [1] 0.5302727
```

Why these two means are not equal?

Section 3

III. How declare constraints?

How declare constraints?

Consider the anova single factor model under one constraint

$$Y = \mu \mathbb{1}_n + A\alpha + \varepsilon.$$

It can be done with the function `lm()` (or with the function `aov()`, we get the same result).

Constraint $\alpha_1 = 0$

This is the constraint by default in **R**. (called also *"Contrast treatement hypotheses"*).

```
mod1=lm(mesure~alcool, data=marqueur);mod1
```

```
##
## Call:
## lm(formula = mesure ~ alcool, data = marqueur)
##
## Coefficients:
##    (Intercept)  alcoolDrinkReg  alcoolNotDrink
##        0.53042         0.04083        -0.02912
```

Here,

$$\widehat{\alpha_1} = 0, \quad \widehat{\mu} = \overline{Y}_{.1} = 0.53042, \quad \widehat{\alpha_2} = \overline{Y}_{.2} - \overline{Y}_{.1} = 0.04083 \text{ and } \widehat{\alpha_3} = \overline{Y}_{.3} - \overline{Y}_{.1} = -0.02912$$

Constraint $\alpha_2 = 0$

```
marqueur$alcool = relevel(marqueur$alcool, ref="DrinkReg")
lm(mesure~alcool, data=marqueur)
```

```
##
## Call:
## lm(formula = mesure ~ alcool, data = marqueur)
##
## Coefficients:
##    (Intercept)  alcoolDrinkOcc  alcoolNotDrink
##        0.57125        -0.04083        -0.06995
```

Here,

$$\widehat{\alpha_2} = 0, \quad \widehat{\mu} = \overline{Y}_{.2} = 0.57125, \quad \widehat{\alpha_1} = \overline{Y}_{.1} - \overline{Y}_{.2} = -0.04083 \text{ and } \widehat{\alpha_3} = \overline{Y}_{.3} - \overline{Y}_{.2} = -0.06995$$

Constraint $\alpha_3 = 0$

```
marqueur$alcool = relevel(marqueur$alcool, ref="NotDrink")
lm(mesure~alcool, data=marqueur)
```

```
##
## Call:
## lm(formula = mesure ~ alcool, data = marqueur)
##
## Coefficients:
##    (Intercept)  alcoolDrinkReg  alcoolDrinkOcc
##        0.50130         0.06995         0.02912
```

Here,

$$\widehat{\alpha_3} = 0, \quad \widehat{\mu} = \overline{Y}_{.3} = 0.50130, \quad \widehat{\alpha_1} = \overline{Y}_{.1} - \overline{Y}_{.3} = 0.02912 \text{ and } \widehat{\alpha_2} = \overline{Y}_{.2} - \overline{Y}_{.3} = 0.06995$$

Constraint $\mu = 0$

As the calculation of R^2 and R_a^2 are done by considering an intercept, the output of these coefficient for this constraint are false.

```
lm(mesure~-1+alcool,data=marqueur)
```

```
##
## Call:
## lm(formula = mesure ~ -1 + alcool, data = marqueur)
##
## Coefficients:
## alcoolNotDrink  alcoolDrinkReg  alcoolDrinkOcc
##         0.5013          0.5713          0.5304
```

Here,

$$\widehat{\mu} = 0, \quad \widehat{\alpha}_1 = \overline{Y}_{.1} = 0.5304, \quad \widehat{\alpha}_2 = \overline{Y}_{.2} = 0.5713 \text{ and } \quad \widehat{\alpha}_3 = \overline{Y}_{.3} = 0.5013$$

Constraint $\sum_{j=1}^{J} n_j \alpha_j = 0$

Note that one coefficient $\widehat{\alpha_j}$ has to be calculated by hand (it depends of the way you defined your matrix of constraint (this contraint is called *"orthogonality constraint"*).

*Here, **R** does rename the modalities.*

```
contrasts(marqueur$alcool)=cbind(c(1,0,-n_j[3]/n_j[1]),
                                 c(0,1,-n_j[2]/n_j[1]))
contrasts(marqueur$alcool)
```

```
##                 [,1]       [,2]
## NotDrink   1.0000000  0.0000000
## DrinkReg   0.0000000  1.0000000
## DrinkOcc -0.3239437 -0.2253521
```

```
#lm(mesure~alcool,data=marqueur)
```

Constraint $\sum_{j=1}^{J} n_j \alpha_j = 0$

```
lm(mesure~alcool,data=marqueur)
```

```
##
## Call:
## lm(formula = mesure ~ alcool, data = marqueur)
##
## Coefficients:
## (Intercept)      alcool1      alcool2
##     0.53027     -0.02897      0.04098
```

Here, $\widehat{\mu} = \overline{Y}_{..} = 0.53027$, $\widehat{\alpha}_3 = \overline{Y}_{.3} - \overline{Y}_{..} = -0.02897$, $\widehat{\alpha}_2 = \overline{Y}_{.2} - \overline{Y}_{..} = 0.04098$

Moreover

$$\widehat{\alpha}_1 = -(n_2/n_1) \times \widehat{\alpha}_2 - (n_3/n_1) \times \widehat{\alpha}_3 = -(0.2253521) \times \widehat{\alpha}_2 - (0.3239437) \times \widehat{\alpha}_3$$

Constraint $\sum_{j=1}^{J} \alpha_j = 0$

*Here, **R** does rename the modalities.*

```
contrasts=list(alcool="contr.sum")
lm(mesure~alcool,contrasts=list(alcool="contr.sum"),data=marqueur)
```

```
##
## Call:
## lm(formula = mesure ~ alcool, data = marqueur, contrasts = list(alcool = "con
##
## Coefficients:
## (Intercept)      alcool1      alcool2
##     0.53433     -0.03302      0.03692
```

$$\widehat{\mu} = \overline{\overline{Y}}_{..} = 0.53433, \quad \widehat{\alpha}_3 = \overline{Y}_{.3} - \overline{\overline{Y}}_{..} = -0.03302, \quad \widehat{\alpha}_2 = \overline{Y}_{.2} - \overline{\overline{Y}}_{..} = 0.03692 \text{ and}$$
$$\widehat{\alpha}_1 = -(\widehat{\alpha}_2 + \widehat{\alpha}_3)$$

Section 4

IV. Study with the constraint $\alpha_1 = 0$

How to display the used constraint

We can display the contraint used by default as follows.

```
getOption( "contrasts")
```

```
##         unordered           ordered
## "contr.treatment"      "contr.poly"
```

Anova 1 factor model with $\alpha_1 = 0$

$$Y = \mu \mathbb{1}_n + A\alpha + \varepsilon.$$

```
summary(mod1)
```

```
##
## Call:
## lm(formula = mesure ~ alcool, data = marqueur)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.13042 -0.05814 -0.02042  0.03642  0.28958
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.53042    0.01008  52.607   <2e-16 ***
## alcoolDrinkReg  0.04083    0.02351   1.736   0.0854 .
## alcoolNotDrink -0.02912    0.02038  -1.429   0.1561
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08496 on 107 degrees of freedom
## Multiple R-squared:  0.05641,    Adjusted R-squared:  0.03877
## F-statistic: 3.198 on 2 and 107 DF,  p-value: 0.04477
```

Comments

- Note that here "`alcool0`" correspond to α_1, so with our constraint "`alcool0`" does not appear as $\alpha_1 = 0$.
- Here, in each line, it is tested if the difference between the EM of the cell $j \neq 1$ and the reference cell $j = 1$ is significant

$$H_0 : \alpha_j = 0 \quad vs \quad H_1 : \alpha_j \neq 0$$

We conclude with the *p-value*.

Comment

☛ In the setting of an anova single factor, the output of `anova(mod1)` displays the global Fisher test.

$$H_0 : Y = \mu \mathbb{1}_n + \varepsilon \quad vs \quad H_1 : Y = \mu \mathbb{1}_n + A\alpha + \varepsilon = X\beta + \varepsilon$$

```
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: mesure
##            Df  Sum Sq  Mean Sq F value  Pr(>F)
## alcool      2 0.04617 0.023084  3.1982 0.04477 *
## Residuals 107 0.77232 0.007218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comments

☛ The global fisher test answers to this question : does the factor `alcool` has an influence on the response variable `mesure`?

☛ Compare to a risk of $\alpha = 5\%$, the *p-value* is smallest, then we reject H_0 at the level α. Thus, the factor is relevant/influent. In other words, this result indicates that the measurements of the intima with the different alcohol status are globally different.

Comments

☛ The command anova applies to the simplest intercept model (mod0) compare to the full one (mod1) gives the *RSS*, the *TSS* and the *MSS*. Here,

$$RSS = 0.78108, \quad TSS = 0.81849, \quad MSS = 0.037415$$

```
mod0 = lm(mesure~1,data=marqueur)
anova(mod0,mod1)
```

```
## Analysis of Variance Table
##
## Model 1: mesure ~ 1
## Model 2: mesure ~ alcool
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1    109 0.81849
## 2    107 0.77232  2  0.046169 3.1982 0.04477 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tukey test

The output (`summary(mod1)`) displays tests which compare the difference between the EM of the cell $j \neq 1$ and the reference cell $j = 1$

$$H_0 : \alpha_j = 0 \quad vs \quad H_1 : \alpha_j \neq 0$$

A natural question is how to test the difference between the EM of the 2 different cells ? To compare all the EM two by two, we can use the Tukey test and compare the p-value to 5%. If at least one *p-value* is larger than 5%, it means that at least one cell (one modality of the factor) influes on the response variable. This is the case here.

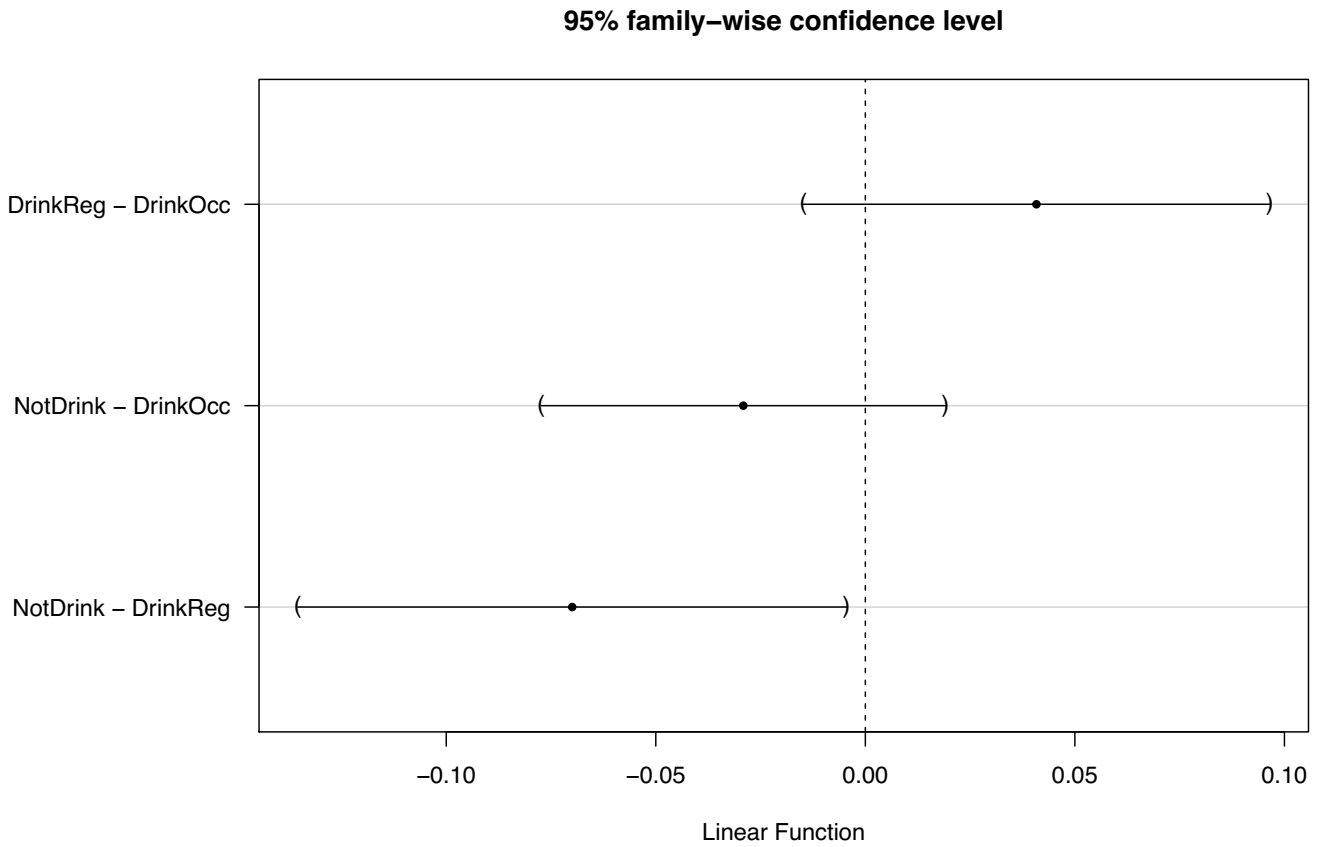
Tukey test

```
library(multcomp)
mc_tukey =glht(mod1, linfct=mcp(alcool="Tukey"))
summary(mc_tukey)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = mesure ~ alcool, data = marqueur)
##
## Linear Hypotheses:
##                            Estimate Std. Error t value Pr(>|t|)
## DrinkReg - DrinkOcc == 0  0.04083    0.02351   1.736   0.1925
## NotDrink - DrinkOcc == 0 -0.02912    0.02038  -1.429   0.3248
## NotDrink - DrinkReg == 0 -0.06995    0.02766  -2.529   0.0332 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Tukey test : graphical comparaisons

```
par(mar=c(9,10,3,3)) #les marges
plot(mc_tukey)
```



Tukey test : “letter” comparaisons

The `multcomp` package also contains the function `cld` that allows, as part of the Tukey test, to indicate by letters the significance of the comparisons. When two modalities share the same letter, it means that their differences are not significantly different. On the other hand, when two modalities do not share letters in common, then it means that their EM are significantly different.

```
cld(mc_tukey)
```

```
## DrinkOcc DrinkReg NotDrink
##     "ab"      "b"      "a"
```

Section 5

V. Residuals analysis

Residuals analysis

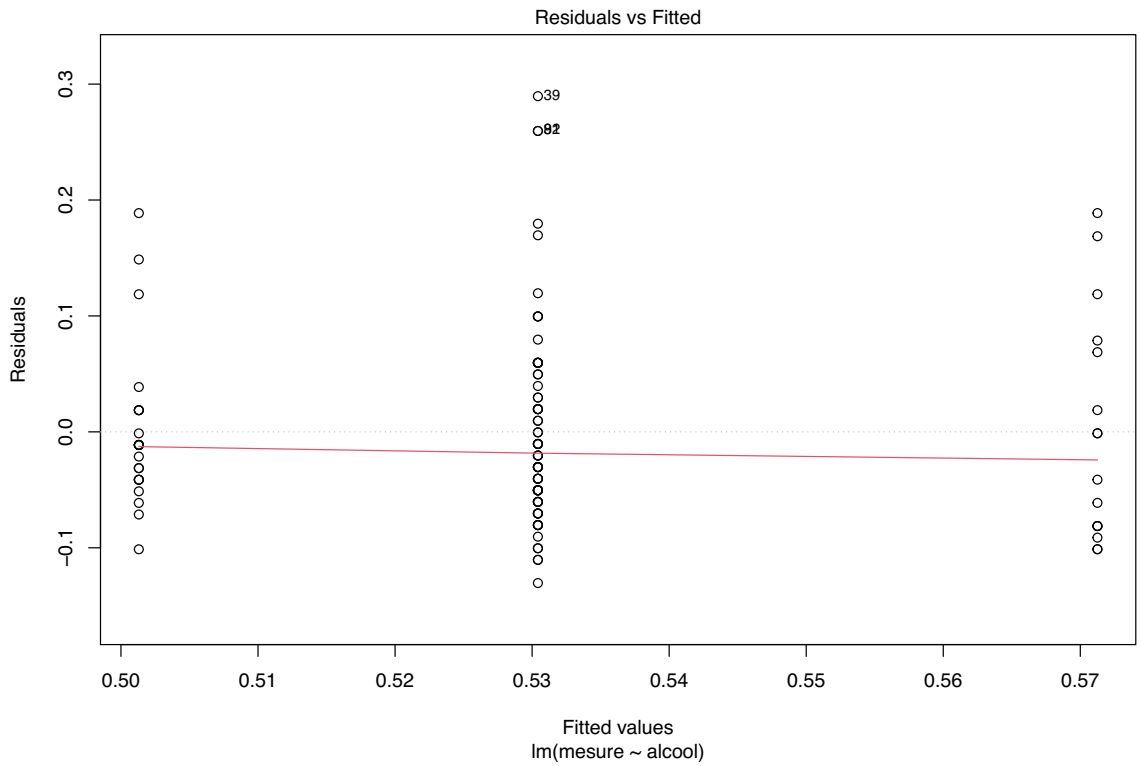
The anova single factor, is a linear model

$$Y = \mu \mathbb{1}_n + A\alpha + \varepsilon = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0_n, \sigma^2 \mathbb{I}_n)$$

So, we have to validate the postulats as usual. We study the estimated residuals.

Postulat [P1] : $\mathbb{E}[\varepsilon] = 0_n$ validated

```
plot(mod1,1)
```



Postulat [P2] : residuals have homoscedastic variance

We can use the Bartlett test, (H_0: Homoscedasticity and $H_1 = \overline{H_0}$). In our setting, the p-value is larger than 5%, we can't reject H_0

```
#plot(mod1,3)
bartlett.test(residuals(mod1)~marqueur$alcool)$p.value
```

```
## [1] 0.307024
```

Postulat [P3] : residuals are uncorrelated

The Durbin Watson test (H_0: uncorrelated). It is therefore concluded that there is no autocorrelation as the test *p-value* is here greater than 5%.

```
set.seed(111)
durbinWatsonTest(mod1)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     0.002975737      1.991699    0.91
##  Alternative hypothesis: rho != 0
```

Postulat [P4] : residuals are gaussian

The *p-value* of the Shapiro test (H_0: uncorrelated) is very small, so we reject the postulat on the normality of the residues.

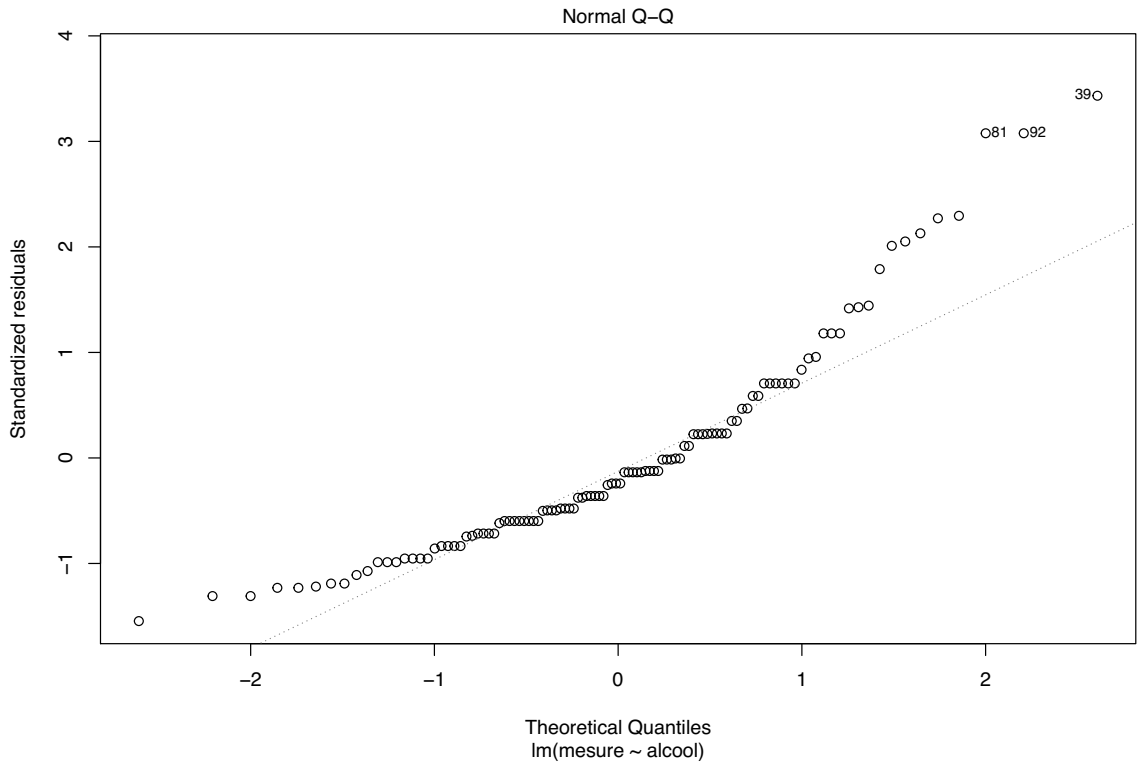
```
shapiro.test(mod1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod1$residuals
## W = 0.89873, p-value = 4.472e-07
```

Postulat [P4] : residuals are gaussian

Graphically, the result of the Shapiro test is confirmed

```
plot(mod1,2)
```



Leverage points: Cook's plot

There is no leverage points to study.

```
plot(mod1,4)
```