MAP553 Regression

Chapter 6: ANOVA

Contents

6.1. Introduction

Until now, we have only studied the case of quantitative variables, but some variables are often qualitative variables. Variance analysis (ANOVA) is a method that makes it possible to study the modification of the average of the phenomenon studied according to the influence of one or more factors of qualitative experiments. A **factor** is a qualitative variable with a limited number of modalities. Let's illustrate the problem of variance analysis on the following example:

Example 1 *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.

6.2. Anova single factor

6.2.1. Definition

We want to explain a variable Y according to one factor. Consider a factor with J modalities, such that $J \in \mathbb{N}^*$. We denote by

- Y_{ij} an observation i that admits j as modality for the factor;
- n_j the number of observations Y_{ij} associated with the modality j of the factor such as :

$$\sum_{j=1}^{J} n_j = n.$$

When there is only one factor, we are talking about Anova single factor model. Formally, we model the quantitative variable Y according to a qualitative explanatory variable that has J possible modalities.

Definition 1 *the plan is said to be*

- *complete if* $\forall j,\ n_j \geq 1$,
- *imcomplete if* $\exists j,\ n_j = 0$,

- *balanced if* $\forall\, j,\ n_j = I$.

Regular model:

Modality 1	$\cdots$	Modality j	$\cdots$	Modality J
$Y_{i1} = \mu_1 + \varepsilon_{i1}$	$\cdots$	$Y_{ij} = \mu_j + \varepsilon_{ij}$	$\cdots$	$Y_{iJ} = \mu_J + \varepsilon_{iJ}$

$$Y_{ij} = \mu_j + \varepsilon_{ij}, \quad i \in \{1, \ldots, n_j\},\ j \in \{1, \ldots, J\} \tag{1}$$

where ε_{ij} is the random error variable. We suppose in this chapter

$$\varepsilon_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2).$$

Of course, this asumption can be verified as we saw in previous Chapters.

Singular model:

Consider the following decompostion of μ_j :

$$\forall\ j \in \{1, \ldots, J\}, \quad \mu_j = \mu + \alpha_j$$

where, the coefficient α_j represents **the main effect of the factor** j.

Modality 1	$\cdots$	Modality j	$\cdots$	Modality J
$Y_{i1} = \mu + \alpha_1 + \varepsilon_{i1}$	$\cdots$	$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$	$\cdots$	$Y_{iJ} = \mu + \alpha_J + \varepsilon_{iJ}$

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}, \quad i \in \{1, \ldots, n_j\},\ j \in \{1, \ldots, J\} \tag{2}$$

where

$$\varepsilon_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2).$$

Matrix form of the singular model:

We use the lexicographic scheduling, to define

$$Y = (Y_{11}, \ldots, Y_{n_1 1}, Y_{12}, \ldots, Y_{n_2 2}, Y_{1J}, \ldots, Y_{n_J J})^T,$$

$$\varepsilon = (\varepsilon_{11}, \ldots, \varepsilon_{n_1 1}, \varepsilon_{12}, \ldots, \varepsilon_{n_2 2}, \varepsilon_{1J}, \ldots, \varepsilon_{n_J J})^T$$

and the design matrix X is defined as follows

$$X = [\mathbb{1}_n \mid A] \qquad \text{where} \qquad A = \begin{pmatrix} \mathbb{1}_{n_1} & \cdots & 0_{n_1} \\ \vdots & \ddots & \vdots \\ 0_{n_J} & \cdots & \mathbb{1}_{n_J} \end{pmatrix}$$

with $\mathbb{1}_{n_j}$ is the one vector of size n_j. By setting $\beta = (\mu, \alpha^T)$ and $\alpha = (\alpha_1, \ldots, \alpha_J)^T$, we can recover our well known matrix form of our model

$$Y = X\beta + \varepsilon = \mu\mathbb{1}_n + A_c\alpha + \varepsilon. \tag{3}$$

We still assume

$$\varepsilon \sim \mathcal{N}(0_n, \sigma^2\mathbb{I}_n).$$

6.2.2. Estimation of the model

Recall that we want to minimize the following quadratic distance $\|Y - X\beta\|^2$, which is miminal for the orthogonal projection of Y into the space generated by the columns of X, say $[X]$. Moreover, it is important to underline that the projection denoted by

$$P_X Y = X\widehat{\beta}$$

is unique. If X is full rank, then the vecteur $\widehat{\beta}$ is also unique, otherwise there is a problem of identifiability, there is an infinity of solutions for $\widehat{\beta}$. In the Anova single factor model, the rank of $n \times (J + 1)$ matrix X is not full as it equals to J. We have to add

$$(J + 1) - \text{Rank}(X) = (J + 1) - J = 1$$

constraint to exhib one of the possible solutions which will depends on the setted constraint.

The classic used constraints:

1. $\mu = 0$.
2. $\alpha_k = 0$ (choice of the cell k as the reference cell).
3. $\sum_{j=1}^{J} \alpha_j = 0$.
4. $\sum_{j=1}^{J} n_j\alpha_j = 0$. (orthogonality constraint)

Comments:

- For the constraint 1., under **R**, we just add -1 in the function `lm()`. This constraint is called the ***Contrast treatement***.
- The constraint 2. for $k = 1$ is the constraint by default. For $k > 1$, it can be done with the `relevel()` under **R**.
- The constraint 3. is called the ***Contrast sum***. Under **R**, we declare it at `contr.sum`.
- The constraint 4. is not coded in **R**, so we have to code it by ourselves.

Some notations:

Empirical mean of	Definition
the observations Y_{ij} having the modality j	$\overline{Y}_{\cdot j} = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij}$
all the observations Y_{ij}	$\overline{Y}_{\cdot\cdot} = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j} Y_{ij} = \frac{1}{n}\sum_{j=1}^{J} n_j \overline{Y}_{\cdot j}$
all the empirical mean $\overline{Y}_{\cdot j}$	$\overline{\overline{Y}}_{\cdot\cdot} = \frac{1}{J}\sum_{j=1}^{J} \overline{Y}_{\cdot j}$

Comments:

- Note that when the plan is balanced, all n_j are equal and $\overline{\overline{Y}}_{\cdot\cdot} = \overline{Y}_{\cdot\cdot}$.
- **Under the constraint $\mu = 0$, the OLSE $\widehat{\alpha}_j$ correspond to $\overline{Y}_{\cdot j}$ the average in the cell j.**

Proposition 1

Consider the singular model defined in (2).

Constraints/Estimators	$\widehat{\mu}$ and $\widehat{\alpha}_j$	
$\mu = 0$	$\rightarrow \widehat{\mu} = 0$	$\widehat{\alpha}_j = \overline{Y}_{\cdot j}, \forall j \in \{1, \dots, J\}$
$\alpha_k = 0$	$\rightarrow \widehat{\mu} = \overline{Y}_{\cdot k}$	$\widehat{\alpha}_j = \overline{Y}_{\cdot j} - \overline{Y}_{\cdot k}, \forall j \in \{1, \dots, J\}$ and $j \neq k$
$\sum_{j=1}^{J} n_j \alpha_j = 0$	$\rightarrow \widehat{\mu} = \overline{Y}_{\cdot\cdot}$	$\widehat{\alpha}_j = \overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot}, \forall j \in \{2, \dots, J\}$ $\widehat{\alpha}_1 = -\left(\sum_{j=2}^{J} n_j \widehat{\alpha}_j\right)/n_1$
$\sum_{j=1}^{J} \alpha_j = 0$	$\rightarrow \widehat{\mu} = \overline{\overline{Y}}_{\cdot\cdot}$	$\widehat{\alpha}_j = \overline{Y}_{\cdot j} - \overline{\overline{Y}}_{\cdot\cdot}, \forall j \in \{2, \dots, J\}$ $\widehat{\alpha}_1 = -\sum_{j=2}^{J} \widehat{\alpha}_j$

Proof : According to the regular model

$$\widehat{\mu}_j = \overline{Y}_{\cdot j}, \quad \forall j \in \{1, \ldots, J\}$$

As the projection is unique, the singular model gives the same estimation

$$\widehat{y}_{ijk} = \widehat{\mu}_j = \widehat{\mu} + \widehat{\alpha}_j.$$

Then by identification we have for all $j \in \{1, \ldots, J\}$

$$\widehat{\mu}_j = \widehat{\mu} + \widehat{\alpha}_j \Leftrightarrow \widehat{\alpha}_j = \widehat{\mu}_j - \widehat{\mu} = \overline{Y}_{\cdot j} - \widehat{\mu}.$$

By using the different constaints in $\boxed{\widehat{\mu} = \overline{Y}_{\cdot j} - \widehat{\mu}}$, we get the results.

- $\boxed{\mu = 0 \Rightarrow \widehat{\mu} = 0}$ and for all $j \in \{1, \ldots, J\}$

  $$\widehat{\alpha}_j = \overline{Y}_{\cdot j} - 0 = \overline{Y}_{\cdot j}$$

- $\boxed{\alpha_k = 0 \Rightarrow \widehat{\alpha}_k = 0}$

  $$\widehat{\alpha}_k = 0 = \overline{Y}_{\cdot k} - \widehat{\mu} \Rightarrow \widehat{\mu} = \overline{Y}_{\cdot k}$$

  And for all $j \neq k$

  $$\widehat{\alpha}_j = \overline{Y}_{\cdot j} - \widehat{\mu} = \overline{Y}_{\cdot j} - \overline{Y}_{\cdot k}$$

- $\boxed{\sum_{j=1}^{J} \alpha_j = 0 \Rightarrow \sum_{j=1}^{J} \widehat{\alpha}_j = 0}$ For all $j = 1, \cdots, J$

  $$\widehat{\alpha}_j = \overline{Y}_{\cdot j} - \widehat{\mu} \Rightarrow \sum_{j=1}^{J} \widehat{\alpha}_j = \sum_{j=1}^{J} \overline{Y}_{\cdot j} - \sum_{j=1}^{J} \widehat{\mu} \Rightarrow 0 = \sum_{j=1}^{J} \overline{Y}_{\cdot j} - J\widehat{\mu} \Rightarrow \widehat{\mu} = \overline{\overline{Y}}_{\cdot\cdot}$$

  Then, for all $j = 1, \cdots, (J-1)$

  $$\widehat{\alpha}_j = \overline{Y}_{\cdot j} - \widehat{\mu} = \overline{Y}_{\cdot j} - \overline{\overline{Y}}_{\cdot\cdot}$$

  And to be sure that the constraint is satisfied we calculate $\widehat{\alpha}_J$ as follows :

  $$\widehat{\alpha}_J = -\sum_{j=1}^{J-1} \widehat{\alpha}_j$$

- $\boxed{\sum_{j=1}^{J} n_j \alpha_j = 0.}$ (orthogonality constraint) : Let in exercice. □

Important comments:

☛ Under the constraint $\alpha_k = 0$. The cell k is the reference cell. Therefore, the coefficient $\widehat{\mu}$ is equal to the empirical average in the cell k of reference. The others coefficients $\widehat{\alpha}_j$ traduce the diffential effect between the average of the cell j and the average of the reference cell k.

☛ Under the constraint $\sum_{j=1}^{J} n_j\alpha_j = 0$. The estimator of the fix effect $\widehat{\mu}$ is the general empirical average $\overline{Y}_{..}$. The others coefficients $\widehat{\alpha}_j$ traduce the diffential effect between the average of the cell j and the general empirical average (the reference cell).

☛ Under the constraint $\sum_{j=1}^{J} \alpha_j = 0$. The estimator of the fix effect $\widehat{\mu}$ is $\overline{\overline{Y}}_{..}$, the mean (average) of the empirical means (of each cell). The others coefficients $\widehat{\alpha}_j$ traduce the diffential effect between the average of the cell j and the average of the empirical averages (the reference cell).

Proposition 2

• *The given estimators in proposition 1 are unbiased under the posutlats **[P1]–[P3]**.*

• *With any constraint, an unbiased estimator of σ^2 is*

$$\widehat{\sigma}^2 = \frac{\|Y - P_X\|^2}{n-J} = \frac{\|Y - P_{[\mathbb{1}_n \mid A]}\|^2}{n-J} = \frac{1}{n-J}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_{\cdot j})^2.$$

• *Under the gaussian assumption **[P4]***

$$\frac{(n-J)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-J).$$

Proof : Immediate according to previous Chapters. □

6.2.2. Tests

To test the impact/influence of the factor on the response variable Y, we can use a global Fisher test.

Theorem 1

➦ *Consider the model (2) with $\varepsilon \sim \mathcal{N}(0_n, \sigma^2\mathbb{I}_n)$. We want to test*

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

➦ *Let us define the following test statistic*

$$F = \frac{\|P_{1_n}Y - P_X Y\|^2/(J-1)}{\widehat{\sigma}^2},$$

➦ *Moreover, under H_0,*

$$F \sim \mathcal{F}_{(J-1,n-J)}.$$

For $\alpha \in]0,1[$, we denote by $q_{J-1,n-J,1-\alpha}$ the quantile of order $1-\alpha$ of the Fisher law at $(J-1, n-J)$ degrees of freedom. Then the Fisher global test of size α forH_0 vs H_1 is

$$\{F > q_{J-1,n-J,1-\alpha}\}.$$

Sketch of proof:

• First note that

$$\text{Rank}(X) - \text{Rank}(\mathbb{1}_n) = \text{Rank}([\mathbb{1}_n \mid A]) - \text{Rank}(1_n) = J - 1.$$

• By proposition 2

$$\frac{(n-J)\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-J).$$

• We conclude by the theorem 3 ("donuts" theorem) chapter 2. □

Comment:

☛ Note that F is such that

$$F = \frac{\sum_{j=1}^{J} n_j(\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot})^2}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_{\cdot j})^2} \times \frac{(n-J)}{(J-1)} = \frac{(RSS_{H_0} - RSS)/(J-1)}{RSS/(n-J)}$$

where

$$\begin{aligned} RSS &= \|Y - P_X Y\|^2 = \|Y - P_{[\mathbb{1}_n \mid A]}Y\|^2 = \|Y - P_A Y\|^2 = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_{\cdot j})^2 \\ RSS_{H_0} &= \|Y - P_{1_n} Y\|^2 = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_{\cdot\cdot})^2 = TSS. \end{aligned}$$

Exercise :

Prove that

1. Under the constraint $\mu = 0$, it stands for all $j \in \{1, \ldots, J\}$,

$$\mathbb{V}\text{ar}(\widehat{\alpha_j}) = \frac{\sigma^2}{n_j}.$$

2. Under the constraint $\sum_{j=1}^{J} n_j \alpha_j = 0$, it stands for all $j \in \{1, \ldots, J\}$,

$$\mathbb{V}\text{ar}(\widehat{\mu}) = \frac{\sigma^2}{n}, \quad \mathbb{V}\text{ar}(\widehat{\alpha_j}) = \left(\frac{1}{n_j} - \frac{1}{n}\right)\sigma^2.$$

6.3. R example Anova 1 factor model

6.3.1. Descriptive analysis

Consider in this section an Anova single factor model. Consider the example introduce section 6.1. about the influence of the `Consumption of alcohol` on the thickness of the intima-media. We recall that the smoking Consumption of alcohol :`alcool` has 3 modalities

- "0"="do not drink"
- "1"="drink occasionally"
- "2"="drink regularly"

Read the dataset

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"
```

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 that the variables have been correctly defined.

```
str(marqueur$mesure)
```

```
##  num [1:110] 0.52 0.42 0.65 0.48 0.45 0.49 0.42 0.45 0.65 0.52 ...
```

```
str(marqueur$alcool)
```

```
##  chr [1:110] "DrinkOcc" "DrinkOcc" "NotDrink" "DrinkOcc" "DrinkOcc" ...
```

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 ...
```

Plot of the dataset

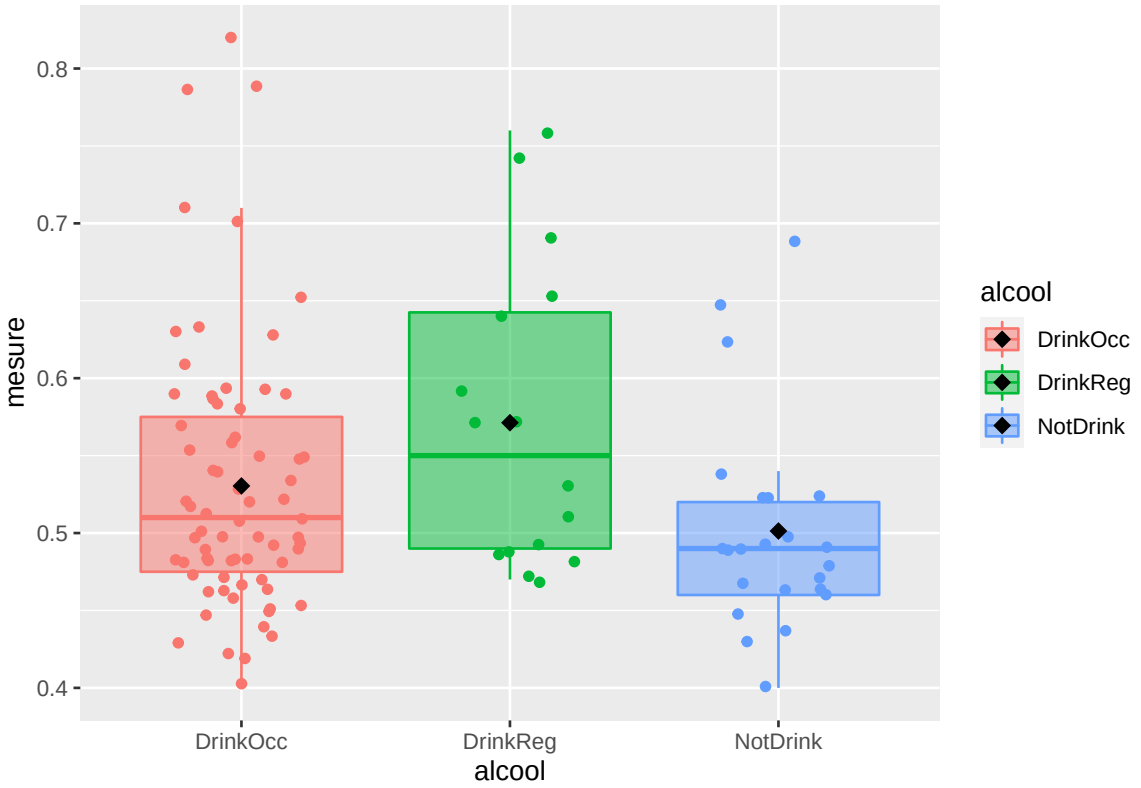
We can plot our dataset with the function `ggplot` of the package `ggplot2`. We will comment the outplut in lecture class.

Comments on the used functions:

- ☛ First underline that the black diamonds represent the averages.
- ☛ 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.)

```
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.y=mean, colour="black", geom="point",shape=18, size=3)
```

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```



Some resumes of the dataset

Display the number of modalities J of the factor

```
J =length(levels(marqueur$alcool))
print(paste("J=",J))
```

```
## [1] "J= 3"
```

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 $\overline{n}_3 = 23$

```
n_j =table(marqueur$alcool);n_j
```

```
##
## DrinkOcc DrinkReg NotDrink
##       71       16       23
```

Note that an easy way to display the average by cell is the following. Here, $\overline{Y}_{.1} = 0.5304225$, $\overline{Y}_{.2} = 0.57125$ and $\overline{Y}_{.3} = 0.5013043$.

```
tapply(marqueur$mesure,list(Alcool=marqueur$alcool),mean,na.rm=TRUE)
```

```
## Alcool
##  DrinkOcc  DrinkReg  NotDrink
## 0.5304225 0.5712500 0.5013043
```

Then , to display the average of the average by cell. Here, $\overline{\overline{Y}}_{..} = 0.5343256$

```
mean(tapply(marqueur$mesure,list(Alcool=marqueur$alcool),mean,na.rm=TRUE))
```

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

To display the average of the variable `mesure`. Here, $\overline{Y}_{..} = 0.5302727$.

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

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

6.3.2. 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). We will comment the output in lecture class.

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,

$$\widehat{\alpha}_1 = 0, \quad \widehat{\mu} = \overline{Y}_{.1} = 0.53042, \quad \widehat{\alpha}_2 = \overline{Y}_{.2} - \overline{Y}_{.1} = 0.04083 \text{ and } \quad \widehat{\alpha}_3 = \overline{Y}_{.3} - \overline{Y}_{.1} = -0.02912$$

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,

$$\widehat{\alpha}_2 = 0, \quad \widehat{\mu} = \overline{Y}_{.2} = 0.57125, \quad \widehat{\alpha}_1 = \overline{Y}_{.1} - \overline{Y}_{.2} = -0.04083 \text{ and } \quad \widehat{\alpha}_3 = \overline{Y}_{.3} - \overline{Y}_{.2} = -0.06995$$

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,

$$\widehat{\alpha}_3 = 0, \quad \widehat{\mu} = \overline{Y}_{.3} = 0.50130, \quad \widehat{\alpha}_1 = \overline{Y}_{.1} - \overline{Y}_{.3} = 0.02912 \text{ and } \quad \widehat{\alpha}_2 = \overline{Y}_{.2} - \overline{Y}_{.3} = 0.06995$$

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.*

```r
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
```

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

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

Here,

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

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.*

```r
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 = "contr.sum
##
## Coefficients:
## (Intercept)      alcool1      alcool2
##     0.53433     -0.03302      0.03692
```

Here,

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

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

We can display the contraint used by default as follows.

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

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

Consider the anova single factor model under the constraint $\alpha_1 = 0$

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

```
library(carData)
library(car)
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 average 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*.

- Note that the last line of the above output gives the global fisher test which tests

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

This is the test described in theorem 1. We can have the same test in the case of an anova single factor with the functions `anova` and `Anova`. Note that theses 2 last functions will not give the same result for other models (see below).

```
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:

- In the setting of an anova single factor, the output of `anova(mod1)` displays the global fisher test.
- 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.
- The command `anova` applies to the simplest intercept model (`mod0`) compare to the full one (`mod1`) gives the RSS, the TSS and the MSS (see bellow).

```
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
```

Comments:

- Here,
$$RSS = 0.78108, \quad TSS = 0.81849, \quad MSS = 0.037415$$
- We can check that
$$TSS = RSS + MSS$$

The output (`summary(mod1)`) displays tests which compare the difference between the average 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 average of the 2 different cells ? To compare all the averages 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.

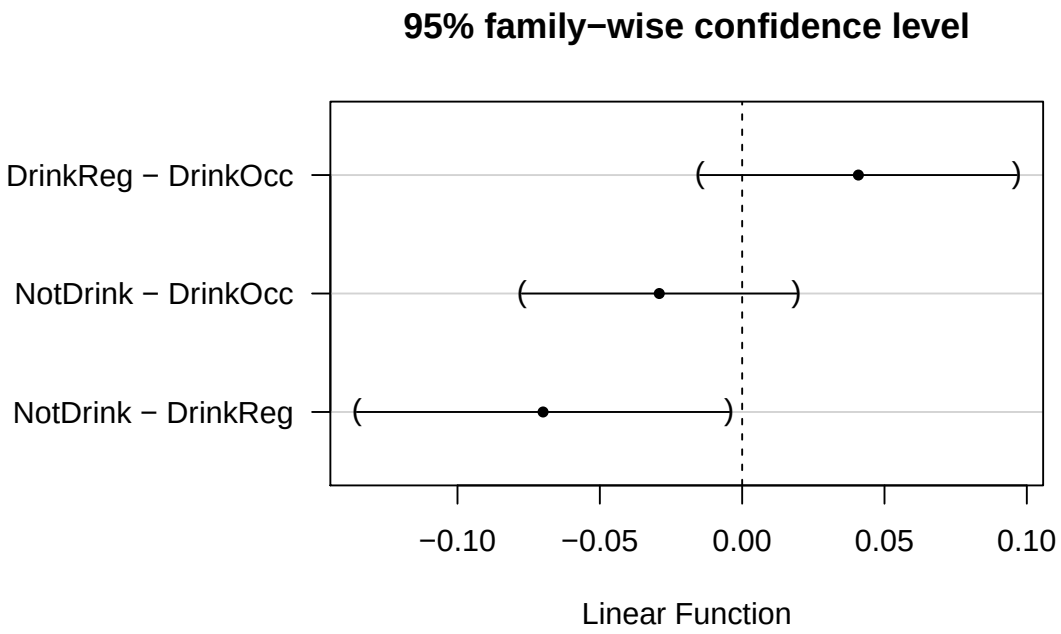
```
library(multcomp)
mc_tukey =glht(mod1, linfct=mcp(alcool="Tukey"))
summary(mc_tukey)
```

```
##
##   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)
```

We may want to graphically view the comparisons:

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



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 averages are significantly different.

```
tuk.cld <- cld(mc_tukey)
tuk.cld
```

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

6.3.4. 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 [P3] : residuals are uncorrelated

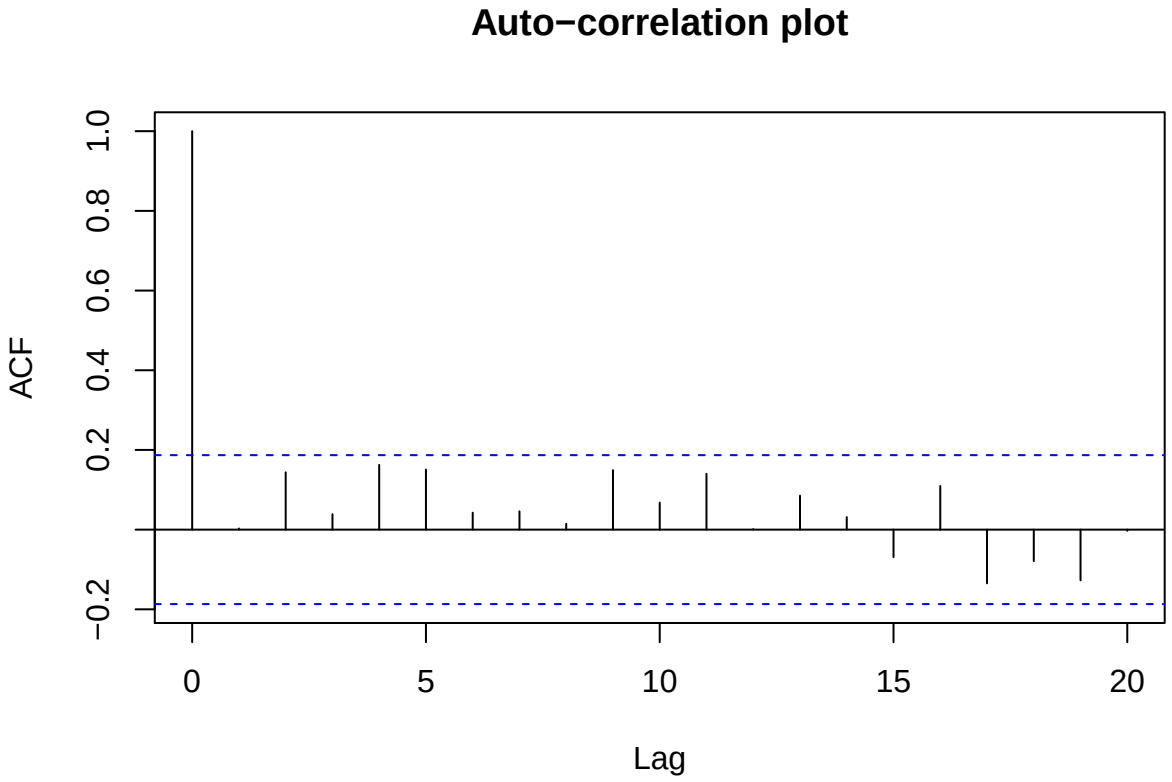
The Durbin Watson test tests the auto correlation. It is therefore concluded that there is no autocorrelation as the test *p-value* is heregreater than 5%.

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

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

Graphically, we come to same conclusion, the residuals are uncorrelated.

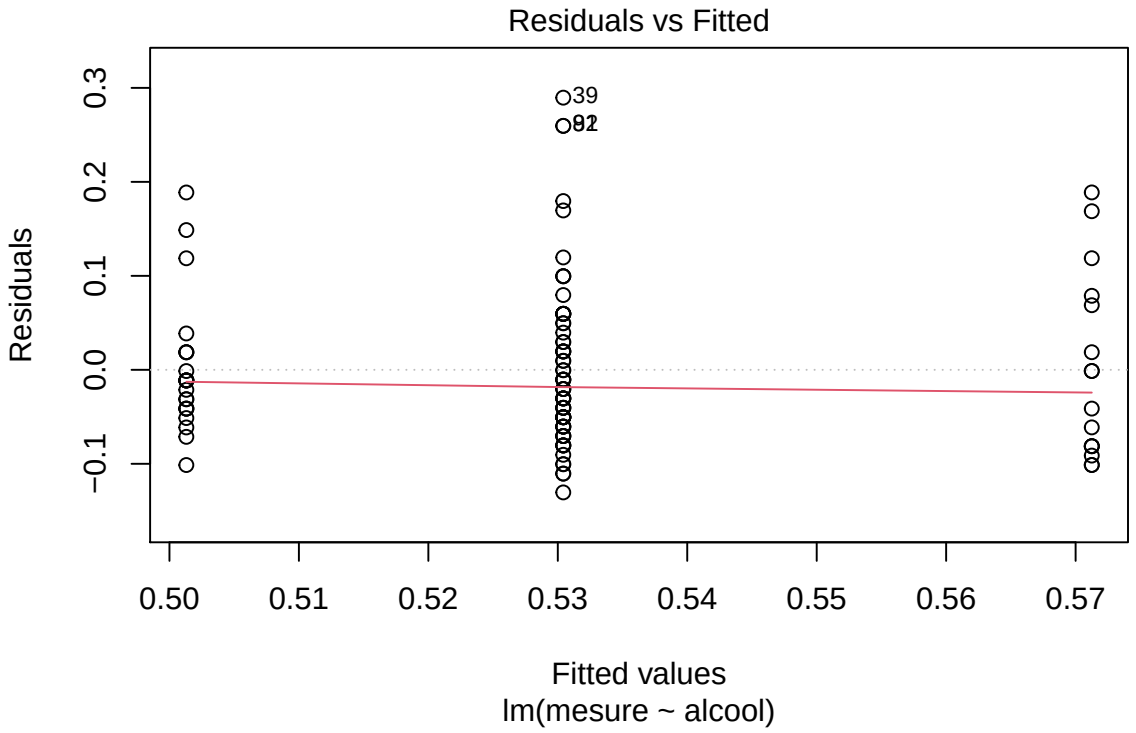
```
acf(residuals(mod1),main="Auto-correlation plot")
```



Postulat [P1] : residuals are centered

Here, the value of the residues does not seem to depend on the treatment since they are all globally centered on 0. So we validate the assumption $\mathbb{E}[\varepsilon] = 0_n$.

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



Postulat [P4] : residuals are uncorrelated

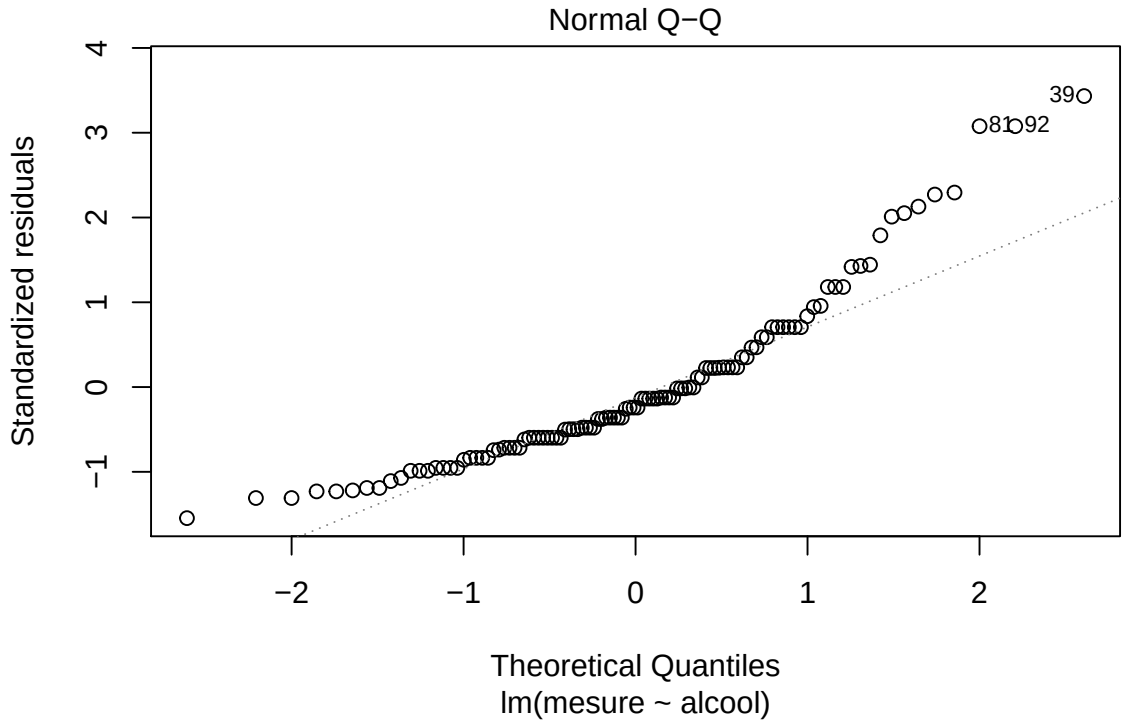
The *p-value* of the Shapiro test is very small, so we reject the postulat on the normality of the residues.

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

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

Graphically, the result of the Shapiro test is confirmed.

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



Postulat [P2] : residuals have homoscedastic variance

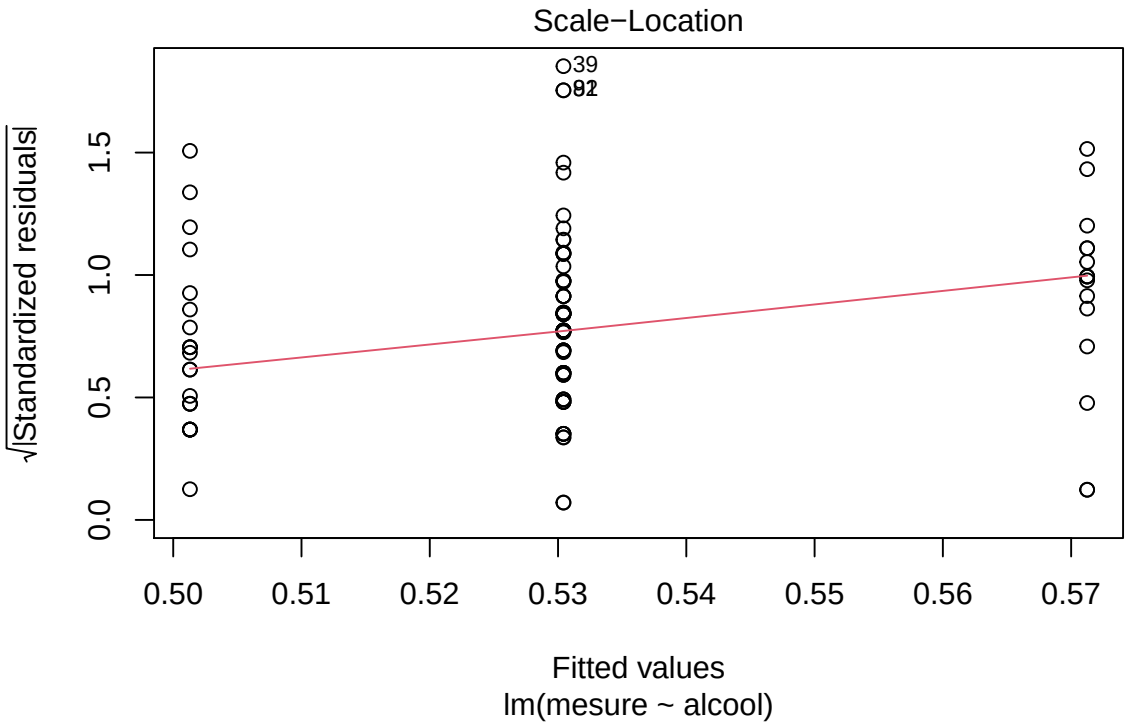
We can use the Bartlett test, (H_0: the variances of the different groups are globally identical and $H_1 = \overline{H_0}$). In our setting, the p-value is larger than 5%, we can't reject H_0

```
bartlett.test(residuals(mod1)~marqueur$alcool)$p.value
```

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

Graphicaly, we see here that the dispersions of the residues (their vertical spacings) relative to each treatment modality are globally identical, the assumption of homogeneity of the residues is accepted.

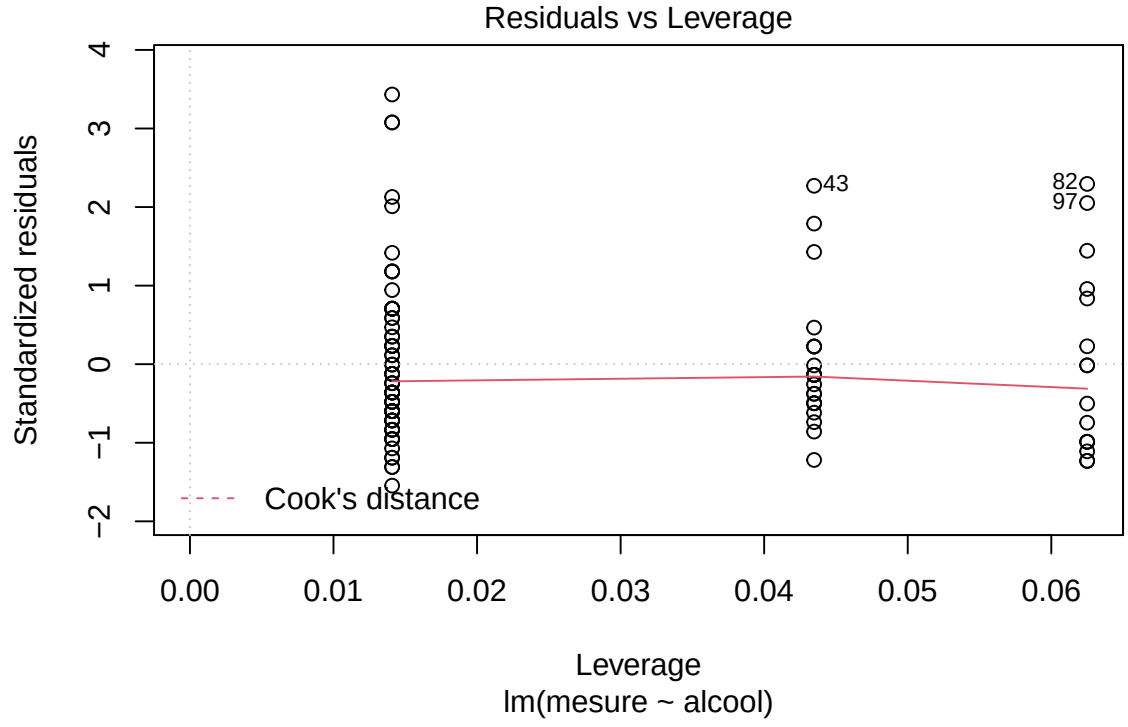
```
plot(mod1,3)
```



Leverage point

There is no leverage points to study.

```
plot(mod1,5)
```



6.4. Anova 2 factors

6.4.1. Definition

The case of two factors is now considered. This section is an extension of the previous one while introducing the possible interactions between factors. The following results can easily be generalized to the ANOVA 3 factors, 4 factors, ... We want to explain a variable Y according to two factors. Consider two factors with respectively J and K modalities, such that $J \in \mathbb{N}^*$ and $K \in \mathbb{N}^*$. We denote by

- Y_{ijk} an observation i that admits j as modality for the first factor and k as modality for the second factor;
- n_{jk} the number of observations associated with the modality j of the first factor and k of the second factor, such as :

$$\sum_{j=1}^{J}\sum_{k=1}^{K} n_{jk} = n, \qquad \sum_{j=1}^{J} n_{jk} = n_{\cdot k} \quad \text{and} \quad \sum_{k=1}^{K} n_{jk} = n_{j\cdot}.$$

Definition 2 *the plan is said to be*

- *complete if* $\forall (j,k)$, $n_{jk} \geq 1$,
- *imcomplete if* $\exists (j,k)$, $n_{jk} = 0$,
- *balanced if* $\forall (j,k)$, $n_{jk} = I$.

Regular model:

Factor I | Factor II	Modality 1	$\cdots$	Modality k	$\cdots$	Modality K
Modality 1	$Y_{i11} = \mu_{11} + \varepsilon_{i11}$	$\cdots$	$Y_{i1k} = \mu_{1k} + \varepsilon_{i1k}$	$\cdots$	$Y_{i1K} = \mu_{1K} + \varepsilon_{i1K}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Modality j	$Y_{ij1} = \mu_{j1} + \varepsilon_{ij1}$	$\cdots$	$Y_{ijk} = \mu_{jk} + \varepsilon_{ijk}$	$\cdots$	$Y_{ijK} = \mu_{jK} + \varepsilon_{ijK}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Modality J	$Y_{iJ1} = \mu_{J1} + \varepsilon_{iJ1}$	$\cdots$	$Y_{iJk} = \mu_{Jk} + \varepsilon_{iJk}$	$\cdots$	$Y_{iJK} = \mu_{JK} + \varepsilon_{iJK}$

For $J \in \mathbb{N}^*$ and $K \in \mathbb{N}^*$, the **Anova 2 factors model** is written as :

$$Y_{ijk} = \mu_{jk} + \varepsilon_{ijk}, \quad i \in \{1, \ldots, n_{jk}\},\ j \in \{1, \ldots, J\},\ k \in \{1, \ldots, K\} \tag{4}$$

where ε_{ijk} is the random error variable and we still assume

$$\varepsilon_{ijk} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2).$$

Of course, this asumption can be verified as we saw in previous Chapters.

Singular model:

In this paragraph, to better analyze the influence of the two factors, we will consider the following decomposition of μ_{jk} :

$$\forall\ j \in \{1, \dots, J\}, \quad \forall\ k \in \{1, \dots, K\}, \quad \mu_{jk} = \mu + \alpha_j + \beta_k + \gamma_{jk}.$$

The coefficients α_j represent **the main effect of the factor** j, the coefficients β_k represent **the main effect of the factor** k and the coefficients γ_{jk} represent **the interaction between the factors** j **and** k.

Factor I | Factor II	Modality 1	$\cdots$	Modality K
Modality 1	$Y_{i11} = \mu + \alpha_1 + \beta_1 + \gamma_{11} + \varepsilon_{i11}$	$\cdots$	$Y_{i1K} = \mu + \alpha_1 + \beta_K + \gamma_{1K} + \varepsilon_{i1K}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Modality J	$Y_{iJ1} = \mu + \alpha_J + \beta_1 + \gamma_{J1k} + \varepsilon_{iJ1}$	$\cdots$	$Y_{iJK} = \mu + \alpha_J + \beta_K + \gamma_{JK} + \varepsilon_{iJK}$

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk}, \quad \forall\ j \in \{1, \dots, J\}, \quad \forall\ k \in \{1, \dots, K\}. \tag{5}$$

where

$$\varepsilon_{ijk} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2).$$

Matrix form of the singular model:

Lexicographic scheduling reading cells from the table above from left to right line by line

$$Y = (Y_{111}, \cdots, Y_{n_{11}11}, \dots, Y_{11K}, \cdots, Y_{n_{1K}1K}, \dots, Y_{1J1}, \cdots, Y_{n_{J1}J1}, \dots, Y_{1JK}, \cdots, Y_{n_{JK}JK})^T$$

$$\varepsilon = (\varepsilon_{111}, \cdots, \varepsilon_{n_{11}11}, \dots, \varepsilon_{11K}, \cdots, \varepsilon_{n_{1K}1K}, \dots, \varepsilon_{1J1}, \cdots, \varepsilon_{n_{J1}J1}, \dots, \varepsilon_{1JK}, \cdots, \varepsilon_{n_{JK}JK})^T.$$

Moreover, let us define $\alpha = (\alpha_1, \cdots, \alpha_J)^T$, $\beta = (\beta_1, \cdots, \beta_K)^T$ and $\gamma = (\gamma_{11}, \cdots, \gamma_{1K}, \cdots, \gamma_{JK})^T$. Then, we can rewrite the model in its following matrix form

$$Y = X\theta + \varepsilon = \mathbb{1}_n \mu + A\alpha + B\beta + C\gamma, \quad \varepsilon \sim \mathcal{N}(0_n, \sigma^2 \mathbb{I}_n) \tag{6}$$

where $\theta = (\mu, \alpha^T, \beta^T, \gamma^T)^T \in \mathbb{R}^{1+J+K+JK}$ and the design matrix X is defined as follows

$$X = [\mathbb{1}_n \mid A \mid B \mid C],$$

where A is a matrix of size $n \times J$, B is a matrix of size $n \times K$ and C is a matrix of size $n \times JK$ such

that

$$A = \begin{pmatrix} \mathbb{1}_{n_{1.}} & 0 & & \\ 0 & \mathbb{1}_{n_{2.}} & & \vdots \\ \vdots & & \ddots & 0 \\ & & & \mathbb{1}_{n_{J.}} \end{pmatrix}, \quad B = \begin{pmatrix} \mathbb{1}_{n_{11}} & 0 & & \\ 0 & \mathbb{1}_{n_{12}} & & \vdots \\ \vdots & & \ddots & 0 \\ & & & \mathbb{1}_{n_{1K}} \\ \mathbb{1}_{n_{21}} & 0 & & \\ 0 & \mathbb{1}_{n_{22}} & & \vdots \\ \vdots & & \ddots & 0 \\ & & & \mathbb{1}_{n_{2K}} \\ \vdots & & & \vdots \\ \mathbb{1}_{n_{J1}} & 0 & & \\ 0 & \mathbb{1}_{n_{J2}} & & \vdots \\ \vdots & & \ddots & 0 \\ & & & \mathbb{1}_{n_{JK}} \end{pmatrix} \quad \text{and} \quad C = \begin{pmatrix} \mathbb{1}_{n_{11}} & 0 & & \\ 0 & \mathbb{1}_{n_{12}} & & \vdots \\ \vdots & & \ddots & 0 \\ & & & \mathbb{1}_{n_{JK}} \end{pmatrix}.$$

6.4.2. Estimation of the model

Again the model is not identifiable because we have $1 + J + K + JK$ parameters to estimate. and $rang(X) = JK$. Therefore, constraints are necessary.

The classic used constraints:

1. **Constraint of type *cell analysis*:** $\forall j = 1, \cdots, J$ and $forall k = 1, \cdots, K$

$$\mu = \alpha_j = \beta_k = 0.$$

2. **Constraint of type *reference cell*:** $\forall j = 1, \cdots, J$ and $forall k = 1, \cdots, K$

$$\alpha_1 = \beta_1 = \gamma_{j1} = \gamma_{1k} = 0.$$

3. **Constraint of type *sum*:** $\forall j' = 1, \cdots, J$ and $\forall k' = 1, \cdots, K$

$$\sum_{j=1}^{J} \alpha_j = \sum_{k=1}^{K} \beta_k = \sum_{k=1}^{K} \gamma_{j'k} = \sum_{j=1}^{J} \gamma_{jk'} = 0.$$

Comment:

☛ Constraint of type *sum* allows to have only JK free parameters and thus guarantee the identifiability of the model. Indeed, the $2 + J + K$ imposed linear relations are not independent and define only $1 + J + K$ constraints, so that the space of the acceptable parameters is of dimension:

$$(1 + J + K + JK) - (1 + J + K) = JK.$$

Some notations:

If the plan is balanced, then the n_{jk} do not depend of j and k; and the n_{jk} are all equal to $I \in \mathbb{N}^*$. Therefore, it comes $n = IJK$. Let us define the following empirical mean/average in the general setting and in the case of a balanced plan.

Empirical mean of	General setting	Balanced plan $n_{jk} = I$
all the observations Y_{ijk}	$\overline{Y}_{\cdots} = \frac{1}{n}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}} Y_{ijk}$	$\overline{Y}_{\cdots} = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{ijk}$
the observations Y_{ijk} having the modalities (j,k)	$\overline{Y}_{\cdot jk} = \frac{1}{n_{jk}}\sum_{i=1}^{n_{jk}} Y_{ijk}$	$\overline{Y}_{\cdot jk} = \frac{1}{I}\sum_{i=1}^{I} Y_{ijk}$
the observations Y_{ijk} having the modalities j	$\overline{Y}_{\cdot j\cdot} = \frac{1}{n_{j\cdot}}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}} Y_{ijk}$ $= \frac{1}{n_{j\cdot}}\sum_{k=1}^{K} n_{jk}\overline{Y}_{\cdot jk}$	$\overline{Y}_{\cdot j\cdot} = \frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K} Y_{ijk}$ $= \frac{1}{K}\sum_{k=1}^{K}\overline{Y}_{\cdot jk}$
the observations Y_{ijk} having the modalities k	$\overline{Y}_{\cdot\cdot k} = \frac{1}{n_{\cdot k}}\sum_{j=1}^{J}\sum_{i=1}^{n_{jk}} Y_{ijk}$ $= \frac{1}{n_{\cdot k}}\sum_{j=1}^{J} n_{jk}\overline{Y}_{\cdot jk}$	$\overline{Y}_{\cdot\cdot k} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ijk}$ $= \frac{1}{J}\sum_{j=1}^{J}\overline{Y}_{\cdot jk}$
the empirical means $\overline{Y}_{\cdot jk}$ having the modalities j	$\overline{\overline{Y}}_{\cdot j\cdot} = \frac{1}{K}\sum_{k=1}^{K}\overline{Y}_{\cdot jk}$ $= \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_{jk}}\sum_{i=1}^{n_{jk}} Y_{ijk}$	$\overline{Y}_{\cdot j\cdot} = \frac{1}{K}\sum_{k=1}^{K}\overline{Y}_{\cdot jk}$ $= \frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K} Y_{ijk}$
the empirical means $\overline{Y}_{\cdot jk}$ having the modalities k	$\overline{\overline{Y}}_{\cdot\cdot k} = \frac{1}{J}\sum_{j=1}^{J}\overline{Y}_{\cdot jk}$ $= \frac{1}{J}\sum_{j=1}^{J}\frac{1}{n_{jk}}\sum_{i=1}^{n_{jk}} Y_{ijk}$	$\overline{Y}_{\cdot\cdot k} = \frac{1}{J}\sum_{k=j}^{J}\overline{Y}_{\cdot jk}$ $= \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ijk}$
the empirical means $\overline{Y}_{\cdot jk}$	$\overline{\overline{Y}}_{\cdots} = \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K}\overline{Y}_{\cdot jk}$ $= \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{1}{n_{jk}}\sum_{i=1}^{n_{jk}} Y_{ijk}$	$\overline{Y}_{\cdots} = \frac{1}{JK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\overline{Y}_{\cdot jk}$ $= \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{ijk}$

From now and for sake of simplicity, we suppose that the plan is balanced. Therefore, it comes $n_{jk} = I$ and $n = IJK$. We set

$$\boxed{n = IJK}$$

Proposition 3

Consider the singular modeldefined in (5) and ***the setting of balanced plan*** *($n = IJK$)*

Constraints/Estimators	$\widehat{\mu}$	$\widehat{\alpha}_j$ and $\widehat{\beta}_k$	$\widehat{\gamma}_{jk}$
$\mu = 0$ $\alpha_j = \beta_k = 0, \forall (j,k)$	$\Rightarrow \widehat{\mu} = 0$	$\Rightarrow \widehat{\alpha}_j = 0, \forall j$ $\Rightarrow \widehat{\beta}_k = 0, \forall k$	$\Rightarrow \widehat{\gamma}_{jk} = \overline{Y}_{\cdot jk}, \forall (j,k)$
$\alpha_1 = \beta_1 = 0$ $\gamma_{j1} = 0, \forall j$ $\gamma_{1k} = 0, \forall k$	$\Rightarrow \widehat{\mu} = \overline{Y}_{\cdot 11}$	$\Rightarrow \widehat{\alpha}_1 = \widehat{\beta}_1 = 0$ $\Rightarrow \widehat{\alpha}_j = \overline{Y}_{\cdot j1} - \overline{Y}_{\cdot 11}, \forall j \neq 1$ $\Rightarrow \widehat{\beta}_k = \overline{Y}_{\cdot 1k} - \overline{Y}_{\cdot 11}, \forall k \neq 1$	$\Rightarrow \widehat{\gamma}_{j1} = 0, \forall j$ $\Rightarrow \widehat{\gamma}_{1k} = 0, \forall k$ $\Rightarrow \forall j \neq 1$ and $\forall k \neq 1$, $\widehat{\gamma}_{jk} = \overline{Y}_{\cdot jk} + \overline{Y}_{\cdot 11} - \overline{Y}_{\cdot j1} - \overline{Y}_{\cdot 1k}$
$\sum_{j=1}^{J} \alpha_j = \sum_{k=1}^{K} \beta_k = 0$ $\sum_{k=1}^{K} \gamma_{j'k} = 0, \forall j'$ $\sum_{j=1}^{J} \gamma_{jk'} = 0, \forall k'$	$\Rightarrow \widehat{\mu} = \overline{Y}_{\cdots}$	$\Rightarrow \widehat{\alpha}_j = \overline{Y}_{\cdot j \cdot} - \overline{Y}_{\cdots}, \forall j \neq 1$ $\Rightarrow \widehat{\alpha}_1 = -\sum_{j=2}^{J} \widehat{\alpha}_j$ $\Rightarrow \widehat{\beta}_k = \overline{Y}_{\cdot \cdot k} - \overline{Y}_{\cdots}, \forall k \neq 1$ $\Rightarrow \widehat{\beta}_1 = -\sum_{k=2}^{K} \widehat{\beta}_k$	$\Rightarrow \widehat{\gamma}_{j1} = 0 \ \forall j$ $\Rightarrow \widehat{\gamma}_{1k} = 0 \ \forall k$ $\Rightarrow \forall j \neq 1$ and $\forall k \neq 1$ $\widehat{\gamma}_{jk} = \overline{Y}_{\cdot jk} + \overline{Y}_{\cdots} - \overline{Y}_{\cdot j \cdot} - \overline{Y}_{\cdot \cdot k}$

Sketch of proof :

First recall that in the the regular model, the OLSE of μ_{jk} is

$$\widehat{\mu}_{jk} = \overline{Y}_{\cdot jk}$$

Then recall that the estimation of $\widehat{Y}_{ijk}$ is unique whatever the used constraints. Then by identification it comes

$$\widehat{\mu}_{jk} = \overline{Y}_{\cdot jk} = \widehat{\mu} + \widehat{\alpha}_j + \widehat{\beta}_k + \widehat{\gamma}_{jk}$$

We conclude by using the constraints.

The results in the previous proposition is only true in the setting of a balanced plan. But it is easy to calculate it in the case of an unbalanced plan. *(Let in exercise)*

Proposition 4

- *The given estimators in proposition 3 are unbiased under the posutlats **[P1]–[P3]**.*
- *With any constraint, an unbiased estimator of σ^2 is*

$$\widehat{\sigma}^2 = \frac{\|Y - P_X Y\|^2}{n - JK} = \frac{\|Y - P_{[\mathbb{1}_n \mid A \mid B \mid C]} Y\|^2}{n - JK} = \frac{1}{n - JK} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (Y_{ijk} - \overline{Y}_{\cdot jk})^2.$$

- *Under the gaussian assumption **[P4]***

$$\frac{(n - JK)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n - JK).$$

Proof : Immediate according to previous Chapters. □

6.4.2. Tests

Let us define the different following models:

- $\mathcal{M}_{\mu} \ : \ Y = \mu \mathbb{1}_n + \varepsilon$
- $\mathcal{M}_{\mu,\alpha} \ : \ Y = \mu \mathbb{1}_n + A\alpha + \varepsilon$
- $\mathcal{M}_{\mu,\beta} \ : \ Y = \mu \mathbb{1}_n + B\beta + \varepsilon$
- $\mathcal{M}_{\mu,\alpha,\beta} \ : \ Y = \mu \mathbb{1}_n + A\alpha + B\beta + \varepsilon$
- $\mathcal{M}_{\mu,\alpha,\beta,\gamma} \ : \ Y = \mu \mathbb{1}_n + A\alpha + B\beta + C\gamma + \varepsilon$

where $\varepsilon \sim \mathcal{N}(0_n, \sigma^2 \mathbb{I}_n)$.

As in the setting of anova single factor, tests can be conducted. **R** also proposes two types of analysis:

- Type I : by the command `anova`$(\mathcal{M}_{\mu,\alpha,\beta,\gamma})$
- Type II : by the command `Anova`$(\mathcal{M}_{\mu,\alpha,\beta,\gamma})$

Line by line, the tests are the following

	Type I	Test Stat. I	Type II	Test Stat. II
Line 1.	$H_0 : \mathcal{M}_{\mu} \quad vs \ H_1 : \mathcal{M}_{\mu,\alpha}$	F^I	$H_0 : \mathcal{M}_{\mu,\beta} \quad vs \ H_1 : \mathcal{M}_{\mu,\alpha,\beta}$	F^{II}
Line 2.	$H_0 : \mathcal{M}_{\mu,\alpha} \quad vs \ H_1 : \mathcal{M}_{\mu,\alpha,\beta}$	$F*$	$H_0 : \mathcal{M}_{\mu,\alpha} \quad vs \ H_1 : \mathcal{M}_{\mu,\alpha,\beta}$	$F*$
Line 3.	$H_0 : \mathcal{M}_{\mu,\alpha,\beta} \ vs \ H_1 : \mathcal{M}_{\mu,\alpha,\beta,\gamma}$	F	$H_0 : \mathcal{M}_{\mu,\alpha,\beta} \ vs \ H_1 : \mathcal{M}_{\mu,\alpha,\beta,\gamma}$	F

Comments:

- Note that only the first line (test) of Type I and II are different. The others tests remain the same whatever the test is of typt I or II.
- From now $\widehat{\sigma}^2$ denote the unbiased estimator calculated from the full model $\mathcal{M}_{\mu,\alpha,\beta,\gamma}$, such that

$$\widehat{\sigma}^2 = \frac{1}{n-JK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk}-\overline{Y}_{\cdot jk})^2 = \frac{\|Y-P_{[\mathbb{1}_n \mid A \mid B \mid C]}Y\|^2}{n-JK} = \frac{\|Y-P_X Y\|^2}{n-JK} = \frac{RSS}{n-JK},$$

  as $X = [\mathbb{1}_n \mid A \mid B \mid C]$.

- We recall that

  - $\text{Rank}(X) = \text{Rank}([\mathbb{1}_n \mid A \mid B \mid C]) = JK$,
  - $\text{Rank}(1_n) = 1$,
  - $\text{Rank}([1_n \mid A]) = J$,
  - $\text{Rank}([1_n \mid B]) = K$,
  - $\text{Rank}([1_n \mid A \mid B]) = J + K - 1$.

Theorem 2 *We consider the model (5).*

• *In the column "Test statistic", we display the associated statistic of test for each test defined in the above table.*

H_0 *vs* H_1	Test statistics	$R = \{F > q_{(DL,1-\alpha}\}$
Line 3. Type I/II	$F = \frac{\|P_{[\mathbb{1}_n \mid A \mid B]}Y - P_X Y\|^2/(J-1)(K-1)}{\widehat{\sigma}^2}$	$DL = ((J-1)(K-1), n-JK)$
Line 2. Type I/II	$F* = \frac{\|P_{[\mathbb{1}_n \mid A]}Y - P_{[\mathbb{1}_n \mid A \mid B]}Y\|^2/(K-1)}{\widehat{\sigma}^2}$	$DL = (K-1, n-JK)$
Line 1. Type I	$F^I = \frac{\|P_{1_n}Y - P_{[\mathbb{1}_n \mid A]}Y\|^2/(J-1)}{\widehat{\sigma}^2}$	$DL = (J-1, n-JK)$
Line 1. Type II	$F^{II} = \frac{\|P_{[\mathbb{1}_n \mid B]}Y - P_{[\mathbb{1}_n \mid A \mid B]}Y\|^2/(K-1)}{\widehat{\sigma}^2}$	$DL = (K-1, n-JK)$

• *Under H_0, every statistic of test follows Fisher law at "DL" degrees of freedom. Therefore,*

$$R = \{F > q_{(DL,1-\alpha}\}$$

is a test of size α for H_0 vs H_1, where $q_{DL,1-\alpha}$ denote the quantile of order $1-\alpha$ of the Fisher law at DL degrees of freedom.

Sketch of proof:

➦ First note that

$$\begin{aligned}
&\text{Rank}(X) - \text{Rank}([\mathbb{1}_n \mid A \mid B]) = JK - (J + K - 1) = (J - 1)(K - 1)\\
&\text{Rank}([\mathbb{1}_n \mid A \mid B) - \text{Rank}([\mathbb{1}_n \mid A]) = (J + K - 1) - J = K - 1\\
&\text{Rank}([\mathbb{1}_n \mid A \mid B]) - \text{Rank}([1_n \mid A]) = (J + K - 1) - J = K - 1\\
&\text{Rank}([\mathbb{1}_n \mid A]) - \text{Rank}(1_n) = J - 1
\end{aligned}$$

➦ By proposition 4

$$\frac{(n - JK)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n - JK).$$

➦ We conclude by the theorem 3 ("donuts" theorem) chapter 2. □

6.5. R example for Anova 2 factor model

6.5.1. The dataset

Consider in this section an Anova two factor model. Consider the example introduce section 6.1. about the influence of the Consumption of alcohol (`alcool`) and the the smoking status (`tabac`) on the thickness of the intima-media (the response `mesure`). We recall that the Consumption of alcohol has 3 modalities that we changed as follows :

"NotDrink"="do not drink", "DrinkOcc"="drink occasionally" and "DrinkReg"="drink regularly"

The dataset

For sake of simplicity in the interpretation, we also change the name of the modalities of the variable `tabac` and declare it as a factor.

```
marqueur$tabac=replace(marqueur$tabac,marqueur$tabac==0,"NotSmoke")
marqueur$tabac=replace(marqueur$tabac,marqueur$tabac==1,"QuitSmoke")
marqueur$tabac=replace(marqueur$tabac,marqueur$tabac==2,"Smoke")
marqueur$tabac=as.factor(marqueur$tabac)
```

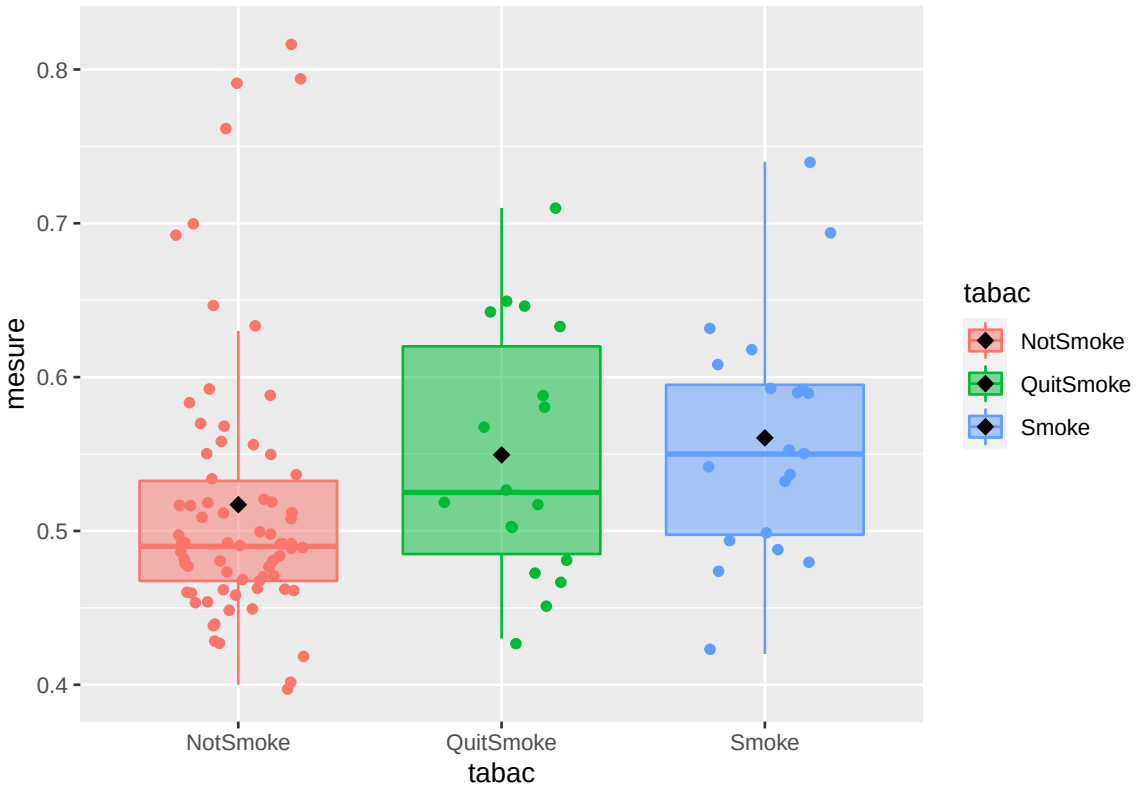
Therefore, the 3 modalities of the factor `tabac` are:

"NotSmoke", "QuitSmoke" and "Smoke"

Plot of the dataset

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

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```



6.5.2 Empirical means

➤➤ Display the number of modalities K of the factor `tabac` and J of the factor `alcool`

```
K =length(levels(marqueur$tabac))
print(paste("K=",K," and J=",J))
```

```
## [1] "K= 3  and J= 3"
```

➤➤ Display the n_{jk} , $j = 1, \cdots, J$ and $k = 1 \cdot, K$ the number of observations of the modality (j, k). Note that, in this dataset, the plan is unbalanced.

```
n_jk =table(marqueur$tabac,marqueur$alcool);n_jk
```

```
##
##              NotDrink DrinkReg DrinkOcc
##    NotSmoke         18        9       45
##    QuitSmoke         3        1       14
##    Smoke             2        6       12
```

Tabac| alcool	$j=1$: NotDrink	$j=2$: DrinkReg	$j=3$: DrinkOcc	$n_{.k}$
$k=1$: NotSmoke	18	9	45	**72**
$k=2$: QuitSmoke	3	1	14	**18**
$k=3$: Smoke	2	6	12	**20**
$n_{j.}$	**23**	**16**	**71**	$n=110$

Table 1: The number n_{jk} of observations by cell (j,k)

➤➤ Note that an easy way to display $\overline{Y}_{.jk}$, the empirical means by cell is the following.

```
Tp=tapply(marqueur$mesure,list(Tabac=marqueur$tabac,Alcool=marqueur$alcool)
       ,mean,na.rm=TRUE);Tp
```

```
##            Alcool
## Tabac        NotDrink  DrinkReg  DrinkOcc
##   NotSmoke  0.4861111 0.5600000 0.5208889
##   QuitSmoke 0.5400000 0.6400000 0.5450000
##   Smoke     0.5800000 0.5766667 0.5491667
```

Tabac| alcool	$j=1$: NotDrink	$j=2$: DrinkReg	$j=3$: DrinkOcc
$k=1$: NotSmoke	0.4861111	0.56	0.5208889
$k=2$: QuitSmoke	0.54	0.64	0.545
$k=3$: Smoke	0.58	0.5766667	0.5491667

Table 2: Empirical means $\overline{Y}_{.jk}=\frac{1}{n_{jk}}\sum_{i=1}^{n_{jk}}Y_{ijk}$

➤➤ Then, to display $\overline{\overline{Y}}_{...}$, the empirical mean of the empirical means $\overline{Y}_{.jk}$

```
mean(Tp)
```

```
## [1] 0.5553148
```

$$\overline{\overline{Y}}_{...}=\frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K}\overline{Y}_{.jk}=0.5553148$$

➤➤ While, $\overline{Y}_{...}$, the general empirical mean of the response variable `mesure` is

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

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

$$\overline{Y}_{...}=\frac{1}{n}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}Y_{ijk}=0.5302727$$

➤➤ We can display $\overline{\overline{Y}}_{.j.}$, the empirical means of the empirical means $\overline{Y}_{.jk}$ having modality j :

```
c(mean(Tp[,1]),mean(Tp[,2]),mean(Tp[,3]))
```

```
## [1] 0.5353704 0.5922222 0.5383519
```

alcool	$j = 1$: NotDrink	$j = 2$: DrinkReg	$j = 3$: DrinkOcc
$\overline{\overline{Y}}_{.j.}$	0.5353704	0.5922222	0.5383519

Table 3: Empirical means of the Empirical means having modality j : $\overline{\overline{Y}}_{.j.} = \frac{1}{K}\sum_{k=1}^{K}\overline{Y}_{.jk}$

➤➤ We can display $\overline{Y}_{.j.}$, the empirical means of the observations $Y_{.ijk}$ having modality j which are note equal to the previous $\overline{\overline{Y}}_{.j.}$ as the plan is unbalanced.

```
tapply(marqueur$mesure,list(Alcool=marqueur$alcool),mean,na.rm=TRUE)
```

```
## Alcool
##  NotDrink  DrinkReg  DrinkOcc
## 0.5013043 0.5712500 0.5304225
```

alcool	$j = 1$: NotDrink	$j = 2$: DrinkReg	$j = 3$: DrinkOcc
$\overline{Y}_{.j.}$	0.5013043	0.5712500	0.5304225

Table 4: Empirical means of the Empirical means having modality j : $\overline{Y}_{.j.} = \frac{1}{n_{j.}}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}Y_{.ijk}$

➤➤ We can also display $\overline{\overline{Y}}_{..k}$, the empirical means of the empirical means $\overline{Y}_{.jk}$ having modality k :

```
c(mean(Tp[1,]),mean(Tp[2,]),mean(Tp[3,]))
```

```
## [1] 0.5223333 0.5750000 0.5686111
```

Tabac	$k = 1$: NotSmoke	$k = 2$: QuitSmoke	$k = 3$: Smoke
$\overline{\overline{Y}}_{..k}$	0.5223333	0.575	5686111

Table 5: Empirical means of the empirical means having modality k: $\overline{\overline{Y}}_{..k} = \frac{1}{J}\sum_{j=1}^{j}\overline{Y}_{.jk}$

➤➤ We can display $\overline{Y}_{\cdot\cdot k}$, the empirical means of the observations $Y_{\cdot ijk}$ having modality k which are note equal to the previous $\overline{\overline{Y}}_{\cdot\cdot k}$ as the plan is unbalanced.

```
tapply(marqueur$mesure,list(Tabac=marqueur$tabac),mean,na.rm=TRUE)
```

```
## Tabac
##  NotSmoke QuitSmoke     Smoke
## 0.5170833 0.5494444 0.5605000
```

Tabac	$k=1$: NotSmoke	$k=2$: QuitSmoke	$k=3$: Smoke
$\overline{Y}_{\cdot\cdot k}$	0.5170833	0.5494444	0.5605000

Table 6: Empirical means of the empirical means having modality k: $\overline{Y}_{\cdot\cdot k} = \frac{1}{n_{\cdot k}} \sum_{j=1}^{J} \sum_{i=1}^{n_{jk}} Y_{\cdot ijk}$

6.5.3. Model anova two factors

➤➤ Let define the following anova 2 factors model

$$Y = \mu \mathbb{1}_n + A\alpha + B\beta + C\gamma + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(O_n, \sigma^2 \mathbb{I}_n),$$

where α is the main effect of the factor `alcool`, β the main effect of the factor `tabac` and γ represents the interaction between the 2 factors. We choose the sums constraints

$$\sum_{j=1}^{J} \alpha_j = \sum_{k=1}^{K} \beta_k = \sum_{j=1}^{J} \gamma_{jk'} = \sum_{k=1}^{K} \gamma_{j'k} = 0, \quad \forall j', k'.$$

```
MOD1=lm(mesure~alcool*tabac,contrasts=list(tabac="contr.sum",
                alcool="contr.sum"),data=marqueur)
summary(MOD1)
```

```
##
## Call:
## lm(formula = mesure ~ alcool * tabac, data = marqueur, contrasts = list(tabac = "co
##     alcool = "contr.sum"))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12917 -0.05089 -0.02003  0.03389  0.29911
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.5553148  0.0145005  38.296   <2e-16 ***
## alcool1        -0.0199444  0.0212016  -0.941   0.3491
## alcool2         0.0369074  0.0235409   1.568   0.1201
```

```
## tabac1          -0.0329815  0.0161587  -2.041   0.0438 *
## tabac2           0.0196852  0.0242560   0.812   0.4190
## alcool1:tabac1 -0.0162778  0.0233496  -0.697   0.4873
## alcool2:tabac1  0.0007593  0.0263577   0.029   0.9771
## alcool1:tabac2 -0.0150556  0.0331171  -0.455   0.6504
## alcool2:tabac2  0.0280926  0.0417098   0.674   0.5022
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08525 on 101 degrees of freedom
## Multiple R-squared:  0.1033, Adjusted R-squared:  0.03224
## F-statistic: 1.454 on 8 and 101 DF,  p-value: 0.1837
```

➤➤ Recall that **R** renames the modality. Therefore, with our notations and if one calculate the OLSE in the setting of unbalanced plan with the constraints sums:

Name	**R** outputs	OLSE1	In terms of empirical means
`Intercept`	0.5553148	$\widehat{\mu}$	$= \overline{\overline{\overline{Y}}}_{\cdots}$
`alcool1`	-0.0199444	$\widehat{\alpha}_1$	$= \overline{\overline{Y}}_{\cdot 1\cdot} - \overline{\overline{\overline{Y}}}_{\cdots}$
`alcool2`	0.0369074	$\widehat{\alpha}_2$	$= \overline{\overline{Y}}_{\cdot 2\cdot} - \overline{\overline{\overline{Y}}}_{\cdots}$
`tabac1`	-0.0329815	$\widehat{\beta}_1$	$= \overline{\overline{Y}}_{\cdot\cdot 1} - \overline{\overline{\overline{Y}}}_{\cdots}$
`tabac2`	0.0196852	$\widehat{\beta}_2$	$= \overline{\overline{Y}}_{\cdot\cdot 2} - \overline{\overline{\overline{Y}}}_{\cdots}$
`alcool1:tabac1`	-0.0162778	$\widehat{\gamma}_{11}$	$= \overline{Y}_{\cdot 11} + \overline{\overline{\overline{Y}}}_{\cdots} - \overline{\overline{Y}}_{\cdot 1\cdot} - \overline{\overline{Y}}_{\cdot\cdot 1}$
`alcool2:tabac1`	0.0007593	$\widehat{\gamma}_{21}$	$= \overline{Y}_{\cdot 21} + \overline{\overline{\overline{Y}}}_{\cdots} - \overline{\overline{Y}}_{\cdot 2\cdot} - \overline{\overline{Y}}_{\cdot\cdot 1}$
`alcool1:tabac2`	0.0150556	$\widehat{\gamma}_{12}$	$= \overline{Y}_{\cdot 12} + \overline{\overline{\overline{Y}}}_{\cdots} - \overline{\overline{Y}}_{\cdot 1\cdot} - \overline{\overline{Y}}_{\cdot\cdot 2}$
`alcool2:tabac2`	0.0280926	$\widehat{\gamma}_{22}$	$= \overline{Y}_{\cdot 22} + \overline{\overline{\overline{Y}}}_{\cdots} - \overline{\overline{Y}}_{\cdot 2\cdot} - \overline{\overline{Y}}_{\cdot\cdot 2}$

Moreover, the other coefficients have to be calculated by hand

$$\widehat{\alpha}_3 = -(\widehat{\alpha}_1 + \widehat{\alpha}_2), \qquad \widehat{\beta}_3 = -(\widehat{\beta}_1 + \widehat{\beta}_2), \qquad \widehat{\gamma}_{13} = (\widehat{\gamma}_{11} + \widehat{\gamma}_{12}) \qquad \text{and} \qquad \widehat{\gamma}_{23} = -(\widehat{\gamma}_{21} + \widehat{\gamma}_{22})$$

➤➤ The other outputs will be discussed in lecture class.

6.5.4. Model selection : commands anova and Anova

➤➤ To highlight the limit of the anova and Anova commands, let's introduce our model in two different ways: change the order of the factor in the lm command

```
MOD=lm(mesure~tabac*alcool,data=marqueur)
MODbis=  lm(mesure~alcool*tabac,data=marqueur)
```

➤➤ Recall that the anova command compares nested models by introducing one by one the factors. When the factor tabac is introduce first, the command concludes that no factor has an impact on the variable mesure.

```
anova(MOD)
```

```
## Analysis of Variance Table
##
## Response: mesure
##               Df  Sum Sq   Mean Sq F value  Pr(>F)
## tabac          2 0.03741 0.0187074  2.5743 0.08120 .
## alcool         2 0.03531 0.0176530  2.4292 0.09324 .
## tabac:alcool   4 0.01180 0.0029509  0.4061 0.80389
## Residuals    101 0.73397 0.0072670
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

➤➤ On the other hand, when the alcool factor is entered first, the command concludes that only the alcool factor has an impact on the variable mesure.

```
anova(MODbis)
```

```
## Analysis of Variance Table
##
## Response: mesure
##               Df  Sum Sq   Mean Sq F value  Pr(>F)
## alcool         2 0.04617 0.0230843  3.1766 0.04593 *
## tabac          2 0.02655 0.0132762  1.8269 0.16619
## alcool:tabac   4 0.01180 0.0029509  0.4061 0.80389
## Residuals    101 0.73397 0.0072670
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

➤➤ Recall that the `Anova` command compares nested models by removing one of the two factors in the models without interaction (the 2 first tests). Whatever the order, in our case the command fails to highlight the impact of the factor `alcool`.

```
Anova(MOD)
```

```
## Anova Table (Type II tests)
##
## Response: mesure
##               Sum Sq  Df F value  Pr(>F)
## tabac        0.02655   2  1.8269 0.16619
## alcool       0.03531   2  2.4292 0.09324 .
## tabac:alcool 0.01180   4  0.4061 0.80389
## Residuals    0.73397 101
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(MODbis)
```

```
## Anova Table (Type II tests)
##
## Response: mesure
##               Sum Sq  Df F value  Pr(>F)
## alcool       0.03531   2  2.4292 0.09324 .
## tabac        0.02655   2  1.8269 0.16619
## alcool:tabac 0.01180   4  0.4061 0.80389
## Residuals    0.73397 101
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

➤➤ We could select the "best" model with step-by-step methods. This is the aim of the next section.

6.5.5. Model selection : Step-by-step method

```
library(MASS)
MOD0=lm(mesure~1,data=marqueur)
MOD=lm(mesure~tabac*alcool,data=marqueur)
```

➤➤ Let's complete our study with a step-by-step model selection.

```
#stepAIC(MOD, ~.,data=marqueur,trace=F,direction=c('backward'))
step(MOD,direction='backward',trace=F)
```

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

```
#stepAIC(MOD0,mesure~tabac*alcool,trace=F,direction=c('forward'))
step(MOD0,mesure~tabac*alcool,direction='forward',trace=F)
```

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

```
#stepAIC(MOD0,mesure~tabac*alcool,trace=F,direction=c('both'))
step(MOD0,mesure~tabac*alcool,direction='both',trace=F)
```

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

➤➤ Every method gives the same final model which is the model anova single factor study in section 6.3. We refer to this section to finish the sudy (validation of the model).

6.6. Illustration under R Ancova Single factor

In many situations, the set of explanatory variables is composed of both quantitative and qualitative variables. We presented only techniques working in one case (quantitative) or the other (qualitative). Since both of these cases are derived from the linear model, it is possible to mix genres: this is called **covariance analysis** (ANCOVA). We will treat here only a simple example when one is in the presence of the case 1 factor and 1 quantitative variable. The generalization will be seen in practice.

The dataset

- `pH` : pH of the wine.
- `Origine`: factor which admits $I = 2$ modalities : `Bordeaux` and `Bourgogne`.
- `Couleur` : factor which admits $I = 2$ modalities : `Blanc` and `Rouge`.
- `Alcool` : It is the alcohol content of the wine.
- `Malique` : Malic acid that reflects greenness / biting wine (green apple).
- `Tartrique` : Tartaric acid that reflects hardness / structure of the wine (the acid most present in the grapes).
- `Citrique` : Citric acid that reflects freshness of the wine (lemony taste).
- `Acetique` : Acetic acid is a natural organic acid, the main constituent of the volatile acidity of a wine.
- `Lactique` : Lactic acid is an organic acid that plays a role in various biochemical processes.
- `AcTot` : Total acidity.

➤➤ First upoload the data set `"CepagesB.csv"` with the function `read.csv2()`

```
Cepages =read.csv2("CepagesB.csv")
names(Cepages)
```

```
##  [1] "Origine"   "Couleur"   "Libelle"   "Alcool"    "pH"        "AcTot"
##  [7] "Tartrique" "Malique"   "Citrique"  "Acetique"  "Lactique"
```

```
Cepages= Cepages[,-(3)] # do not consider the column "Libelle"
```

➤➤ Our aim is to explain Y =`pH`. by the factor `Couleur` and the covariate/regressor `AcTot`.

Resume of dataset

➤➤ We have $n = 36$ observations.

```
dim(Cepages)
```

```
## [1] 36 10
```

➤➤ The plan is balanced.

```
table(Cepages$Couleur)
```

```
##
## Blanc Rouge
##    18    18
```

➤➤ Display the table of empirical means by cell.

```
Tmean=tapply(Cepages$pH,list(Coul=Cepages$Couleur),mean);Tmean
```

```
## Coul
##    Blanc    Rouge
## 3.040556 3.414444
```

➤➤ As the plan is balanced, the empirical mean of the pH and the empirical mean of all the empirical means are equal.

```
c(mean(Tmean),mean(Cepages$pH))
```

```
## [1] 3.2275 3.2275
```

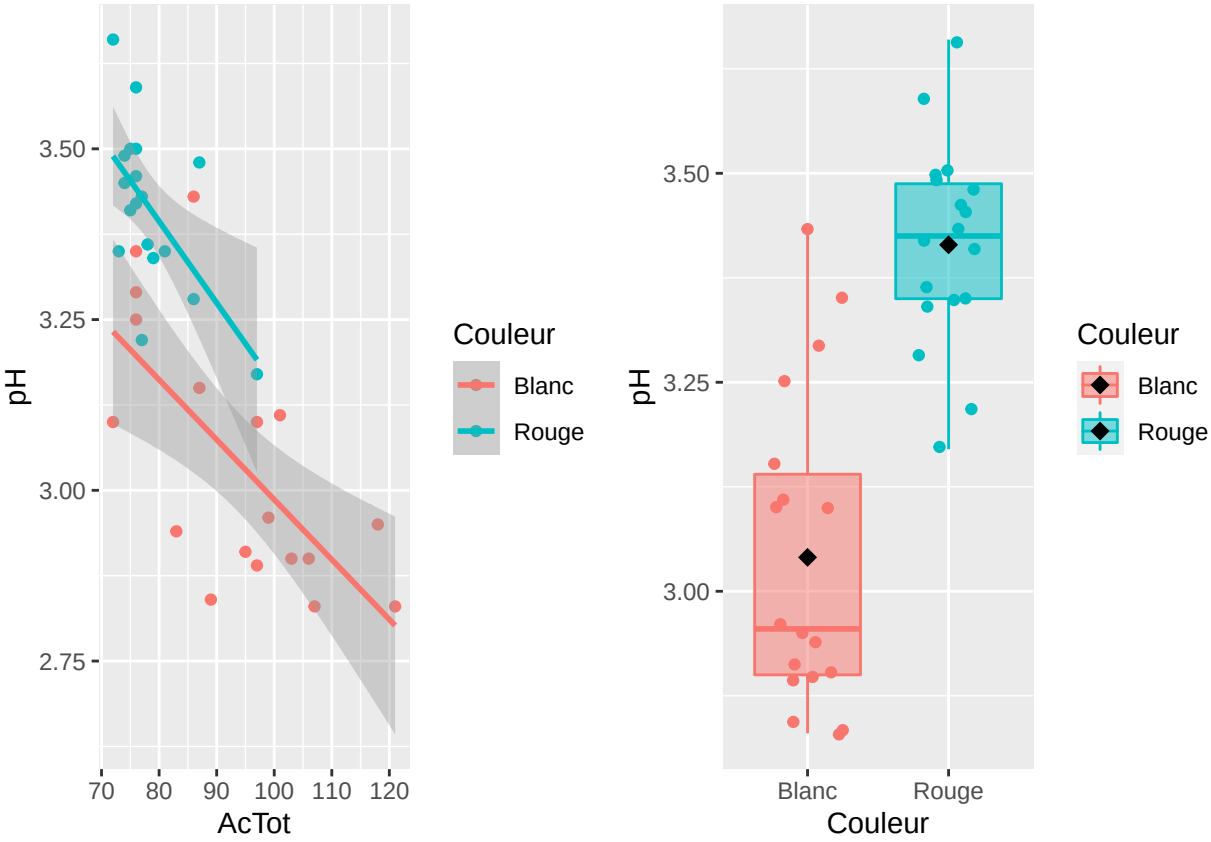
Plot the dataset

```
AcTot=Cepages[,"AcTot"]
pH=Cepages[,"pH"]
Couleur= as.factor(Cepages[,"Couleur"])
library(cowplot)
library(ggplot2)
PlotCouleur1=ggplot(Cepages, aes(x = AcTot, y =pH,color=Couleur)) +
geom_point()+geom_smooth(method = "lm")
```

```
PlotCouleur2=ggplot(Cepages, aes(y=pH, x=Couleur,colour=Couleur ,fill=Couleur))+
geom_boxplot(alpha=0.5, outlier.alpha=0)+geom_jitter(width=0.25)+
stat_summary(fun.y=mean, colour="black", geom="point",shape=18, size=3)
```

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```

```
plot_grid(PlotCouleur1,PlotCouleur2,ncol=2,nrow=1)
```



➤➤ It seems that the `Couleur` factor has an impact on the variable `pH`. The regression lines are dfifferent with respect to the choosen modality.

6.7. Modelisation of an Ancova Single factor

6.7.1. Definition of the model

Treat now the case of a simple example when we are in the presence of the case 1 factor and 1 quantitative variable.

- The factor modeled is supposed to have J possible modalities.
- Therefore, rather than using a single index for the variable to be explained, we write Y_{ij} to denote the observation i having modality j.

- The quantitative explanatory variable, also called covariate, is modeled by x. We denote by x_{ij} to denote the observation i having modality j.
- We define n_j the number of observations Y_{ij} associated with the modality j of the factor such as :

$$\sum_{j=1}^{J} n_j = n.$$

To write the ancova model, we will assume that the regression line differs according to the modalities, that is to say that the y-intercept τ_j and the slope β_j varies according to the modality j.

Definition 3 *the plan is said to be*

- *complete if* $\forall j,\ n_j \geq 1$,
- *imcomplete if* $\exists j,\ n_j = 0$,
- *balanced if* $\forall j,\ n_j = I$.

Regular Model:

Modality 1	$\cdots$	Modality j	$\cdots$	Modality J
$Y_{i1} = \tau_1 + \beta_1 x_{i1} + \varepsilon_{i1}$	$\cdots$	$Y_{ij} = \tau_j + \beta_j x_{ij} + \varepsilon_{ij}$	$\cdots$	$Y_{iJ} = \tau_J + \beta_J x_{iJ} + \varepsilon_{iJ}$

$$Y_{ij} = \tau_j + \beta_j x_{ij} + \varepsilon_{ij}, \quad i \in \{1, \ldots, n_j\},\ j \in \{1, \ldots, J\} \tag{7}$$

where ε_{ij} is the random error and n_j the number of observations Y_{ij} associated to the modality j of the factor. We still assume

$$\varepsilon_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2).$$

Of course, this hypothesis can be verified as we saw in previous Chapters..

Matrix form of the regular model

First define for all $j = 1, \cdots, J$, the vectors $Y^j \in \mathbb{R}^{n_j}$, $\varepsilon^j \in \mathbb{R}^{n_j}$, $\theta^j \in \mathbb{R}^2$ and the X^j matrices of size $n_j \times 2$ such that

$$Y^j = \begin{pmatrix} Y_{1j} \\ \vdots \\ Y_{n_j j} \end{pmatrix}, \qquad \varepsilon^j = \begin{pmatrix} \varepsilon_{1j} \\ \vdots \\ \varepsilon_{n_j j} \end{pmatrix}, \qquad \theta^j = \begin{pmatrix} \tau_j \\ \beta_j \end{pmatrix} \quad \text{and} \quad X^j = \begin{pmatrix} 1 & x_{1j} \\ \vdots & \vdots \\ 1 & x_{n_j j} \end{pmatrix}$$

We can now define $Y \in \mathbb{R}^n$ the response vector, $\varepsilon \in \mathbb{R}^n$ the error vector, $\theta \in \mathbb{R}^{2J}$ unknown parameters vector and X the design matrix of size $n \times 2J$

$$Y = \begin{pmatrix} Y^1 \\ \vdots \\ Y^J \end{pmatrix}, \qquad \varepsilon = \begin{pmatrix} \varepsilon^1 \\ \vdots \\ \varepsilon^J \end{pmatrix}, \qquad \theta = \begin{pmatrix} \theta^1 \\ \vdots \\ \theta^J \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} X^1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & X^J \end{pmatrix}$$

Therefore, the regular model (7) can be written in the following matrix form :

$$Y = X\theta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0_n, \mathbb{I}_n) \tag{8}$$

Singular model:

Consider the following decompostion of

$$\tau_j + \beta_j x_{ij} = (\mu + \alpha_j) + (b + c_j)x_{ij}$$

Modality 1	$\cdots$	Modality J
$Y_{i1} = (\mu + \alpha_1) + (b + c_1)x_{i1} + \varepsilon_{i1}$	$\cdots$	$Y_{iJ} = (\mu + \alpha_J) + (b + c_J)x_{iJ} + \varepsilon_{iJ}$

$$Y_{ij} = \underbrace{(\mu + \alpha_j)}_{\tau_j} + \underbrace{(b + c_j)}_{\beta_j} x_{ij} + \varepsilon_{ij}, \quad i \in \{1, \ldots, n_j\},\ j \in \{1, \ldots, J\} \tag{9}$$

where

$$\varepsilon_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2).$$

Note that the previous parameters represent

- μ : **y-intercept of reference**.
- $\mu + \alpha_j$: **y-intercept of the cell** j.
- b : **the reference slope**.
- $b + c_j$: **the slope of the cell** j.

Matrix form of the regular model

We consider in this paragraph, the same definitions of the vector $Y \in \mathbb{R}^n$ and $\varepsilon \in \mathbb{R}^n$. We define for all $j = 1, \cdots, J$, the vectors $x^j = (x_{1j}, \cdots, x_{n_j,j})^T$. We denote by $x \in \mathbb{R}^n$ the vector of observation x_{ij}, by $\boldsymbol{x}$ the $n \times J$ matrix, by A the $n \times J$ matrix, by c and α deux vectors of size J such that

$$x = \begin{pmatrix} c_1 \\ \vdots \\ c_J \end{pmatrix}, \qquad \boldsymbol{x} = \begin{pmatrix} x^1 & 0 & & \\ 0 & x^2 & & \vdots \\ \vdots & & \ddots & 0 \\ & & & x^J \end{pmatrix}, \qquad A = \begin{pmatrix} \mathbb{1}_{n_1} & \cdots & 0_{n_1} \\ \vdots & \ddots & \vdots \\ 0_{n_J} & \cdots & \mathbb{1}_{n_J} \end{pmatrix}, \qquad c = \begin{pmatrix} c_1 \\ \vdots \\ c_J \end{pmatrix} \quad \text{and} \quad \alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_J \end{pmatrix}$$

Therefore, the regular model (9) can be written in the following matrix form :

$$Y = \mu\mathbb{1}_n + A\alpha + bx + \boldsymbol{x}c + \varepsilon = \boldsymbol{X}\boldsymbol{\theta} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0_n, \mathbb{I}_n) \tag{10}$$

where

$$\boldsymbol{X} = [\mathbb{1}_n \mid A \mid x \mid \boldsymbol{x}] \quad \text{and} \quad \boldsymbol{\theta} = \begin{pmatrix} \mu \\ \alpha \\ b \\ c \end{pmatrix} \in \mathbb{R}^{2J+2}$$

6.7.2. Estimation of the model

In the model (7), the number of parameters to estimate is \$2J \$, the matrix X is assumed of full rank. In the singular model (9), the number of parameters to estimate is $2 + 2J$,the matrix $\boldsymbol{X}$ is not full rank. Therefore, the model is not identifiable. To make the model identifiable the following constraints can be used:

The classic used constraints:

1. $\alpha_1 = c_1 = 0$.
2. $\alpha_k = c_k = 0$ (choice of the cell k as the reference cell).
3. $\sum_{j=1}^{J} \alpha_j = \sum_{j=1}^{J} c_j = 0$.
4. $\sum_{j=1}^{J} n_j\alpha_j = \sum_{j=1}^{J} n_j c_j = 0$. (orthogonality constraint)

Comments:

- ☛ The constraint 1. is the constraint by default under **R** and is called the ***Contrast treatement***.
- ☛ The constraint 2. For $k > 1$, it can be done with the `relevel()` under **R**.
- ☛ The constraint 3. is called the ***Contrast sum***. Under **R**, we declare it at `contr.sum`.
- ☛ The constraint 4. is not coded in **R**, so we have to code it by ourselves.

Some notations:

Empirical	Definition
mean of the observations Y_{ij} having the modality j	$\overline{Y}_{\cdot j} = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij}$
mean of all the observations Y_{ij}	$\overline{Y}_{\cdot\cdot} = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j} Y_{ij} = \frac{1}{n}\sum_{j=1}^{J} n_j \overline{Y}_{\cdot j}$
mean of all the empirical mean $\overline{Y}_{\cdot j}$	$\overline{\overline{Y}}_{\cdot\cdot} = \frac{1}{J}\sum_{j=1}^{J} \overline{Y}_{\cdot j}$
mean of all the observations x_{ij}	$\overline{x}_{\cdot j} = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij}$
mean of order 2 of all the observations x_{ij}	$\overline{x^2}_{\cdot j} = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij}^2$
mean of the observations (x_{ij}, Y_{ij})	$\overline{x_{\cdot j}Y_{\cdot j}} = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij}Y_{ij}$

As in the case of an anova singlefactor, the values of the OLSE depend on the constraint used. Here we will only give the case of constraint 1. For other constraints, a similar calculation is sufficient to find the result.

Proposition 5

	Estimators	
Regular model (7) No constraints	$\widehat{\tau}_j = \frac{\overline{x^2}_{\cdot j}\overline{Y}_{\cdot j} - \overline{x}_{\cdot j}\overline{x_{\cdot j}Y_{\cdot j}}}{\overline{x^2}_{\cdot j} - (\overline{x}_{\cdot j})^2}$	$\widehat{\beta}_i = \frac{\overline{x_{\cdot j}Y_{\cdot j}} - \overline{x}_{\cdot j}\overline{Y}_{\cdot j}}{\overline{x^2}_{\cdot j} - (\overline{x}_{\cdot j})^2}$
Singular model (9) Constr. $\alpha_1 = c_1 = 0$	$\widehat{\alpha}_1 = \widehat{c}_1 = 0$	$\widehat{\alpha}_j = \widehat{\tau}_j - \widehat{\tau}_1, \forall j \geq 2$ $\widehat{c}_j = \widehat{\beta}_j - \widehat{\beta}_1, \forall j \geq 2$

Sketch of proof

- In the model (7) the X matrix is assumed to be of full rank $2J$, so the matrices X^j are also of full rank 2. It follows that the ordinary least squares estimator (OLSE) gives $\widehat{\theta} = (X^T X)^{-1} X^T Y$. Since the X matrix is diagonal by block, we have for all $j = 1, \cdots, J$:

$$\widehat{\theta}_j = \begin{pmatrix} \widehat{\tau}_j \\ \widehat{\beta}_j \end{pmatrix} = \left((X^j)^T X^j\right)^{-1} (X^j)^T Y^j$$

- As $\widehat{Y}_{ij}$ is unique, it comes

$$\widehat{Y}_{ij} = \widehat{\mu} + \widehat{\alpha}_j + (\widehat{b} + \widehat{c}_j)x_{ij} = \widehat{\tau}_j + \widehat{\beta}_j x_{ij}$$

- The result is obtained using the constraint and by identification. □

Proposition 6

- *The given estimators in proposition 5 are unbiased under the posutlats **[P1]–[P3]**.*
- *With any constraint, an unbiased estimator of σ^2 is*

$$\widehat{\sigma}^2 = \frac{\|Y - P_X\|^2}{n - 2J} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(\widehat{Y}_{ij} - Y_{ij})^2}{n - 2J}.$$

- *Under the gaussian assumption **[P4]***

$$\frac{(n - 2J)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n - 2J).$$

Proof : Immediate according to previous Chapters. □

6.7.3. Test

Let us define the different following models:

- $\mathcal{M}_{\mu} \ : \ Y = \mu\mathbb{1}_n + \varepsilon$
- $\mathcal{M}_{\mu,\alpha} \ : \ Y = \mu\mathbb{1}_n + A\alpha + \varepsilon$
- $\mathcal{M}_{\mu,b} \ : \ Y = \mu\mathbb{1}_n + bx + \varepsilon$
- $\mathcal{M}_{\mu,\alpha,b} \ : \ Y = \mu\mathbb{1}_n + A\alpha + bx + \varepsilon$
- $\mathcal{M}_{\mu,b,c} \ : \ Y = \mu\mathbb{1}_n + bx + \boldsymbol{x}c + \varepsilon$
- $\mathcal{M}_{\mu,\alpha,b,c} \ : \ Y = \mu\mathbb{1}_n + A\alpha + bx + \boldsymbol{x}c + \varepsilon$

where $\varepsilon \sim \mathcal{N}(0_n, \sigma^2\mathbb{I}_n)$.

As in the setting of anova two factors, tests can be conducted. **R** proposes two types of analysis:

- Type I : by the command `anova(`$\mathcal{M}_{\mu,\alpha,b,c}$`)`
- Type II : by the command `Anova(`$\mathcal{M}_{\mu,\alpha,b,c}$`)`

We suppose here that we define our model int the following order

```
mod=lm(Y~Factor*Covariate)
```

Line by line, the tests are the following

	Type I	Test Stat. I	Type II	Test Stat. II
Line 1.	$H_0 : \mathcal{M}_{\mu} \quad vs\ H_1 : \mathcal{M}_{\mu,\alpha}$	F^I	$H_0 : \mathcal{M}_{\mu,b} \quad vs\ H_1 : \mathcal{M}_{\mu,\alpha,b}$	F^{II}
Line 2.	$H_0 : \mathcal{M}_{\mu,\alpha} \quad vs\ H_1 : \mathcal{M}_{\mu,\alpha,b}$	$F*$	$H_0 : \mathcal{M}_{\mu,\alpha} \quad vs\ H_1 : \mathcal{M}_{\mu,\alpha,b}$	$F*$
Line 3.	$H_0 : \mathcal{M}_{\mu,\alpha,b} \ vs\ H_1 : \mathcal{M}_{\mu,\alpha,b,c}$	F	$H_0 : \mathcal{M}_{\mu,\alpha,b} \ vs\ H_1 : \mathcal{M}_{\mu,\alpha,b,c}$	F

Comments:

- Note that only the first line (test) of Type I and II are different. The others tests are the same.
- From now $\widehat{\sigma}^2$ denote the unbiased estimator calculated from the full model $\mathcal{M}_{\mu,\alpha,\beta,\gamma}$ and defined in proposition 6

$$\widehat{\sigma}^2 = \frac{\|Y - P_{[\mathbb{1}_n \mid A \mid x \mid \boldsymbol{x}]}Y\|^2}{n - JK} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(\widehat{Y}_{ij} - Y_{ij})^2}{n - 2J} = \frac{RSS}{n - 2J}.$$

  as $X = [\mathbb{1}_n \mid A \mid x \mid \boldsymbol{x}]$.

- We recall that

  - $\text{Rank}(X) = \text{Rank}([\mathbb{1}_n \mid A \mid x \mid \boldsymbol{x}]) = 2J$,
  - $\text{Rank}(1_n) = 1$,
  - $\text{Rank}([1_n \mid A]) = J$,
  - $\text{Rank}([1_n \mid x]) = 2$,
  - $\text{Rank}([1_n \mid A \mid x]) = J + 1$.

Theorem 3 *We consider the model (9).*

• *In the column "Test statistic", we display the associated statistic of test for each test defined in the above table.*

H_0 *vs* H_1	Test statistics	$R = \{F > q_{(DL,1-\alpha}\}$
Line 3. Type I/II	$F = \frac{\|P_{[\mathbb{1}_n \mid A \mid x]}Y - P_X Y\|^2/(J-1)}{\widehat{\sigma}^2}$	$DL = (J-1, n-2J)$
Line 2. Type I/II	$F* = \frac{\|P_{[\mathbb{1}_n \mid A]}Y - P_{[\mathbb{1}_n \mid A \mid x]}Y\|^2/1}{\widehat{\sigma}^2}$	$DL = (1, n-2J)$
Line 1. Type I	$F^I = \frac{\|P_{1_n}Y - P_{[\mathbb{1}_n \mid A]}Y\|^2/(J-1)}{\widehat{\sigma}^2}$	$DL = (J-1, n-2J)$
Line 1. Type II	$F^{II} = \frac{\|P_{[\mathbb{1}_n \mid x]}Y - P_{[\mathbb{1}_n \mid A \mid x]}Y\|^2/(J-1)}{\widehat{\sigma}^2}$	$DL = (J-1, n-2J)$

• *Under H_0, every statistic of test follows Fisher law at "DL" degrees of freedom. Therefore,*

$$R = \{F > q_{(DL,1-\alpha}\}$$

is a test of size α for H_0 vs H_1, where $q_{DL,1-\alpha}$ denote the quantile of order $1-\alpha$ of the Fisher law at DL degrees of freedom.

Sketch of proof:

➦ First note that

$$\begin{aligned}
&\text{Rank}(X) - \text{Rank}([\mathbb{1}_n \mid A \mid x]) = 2J - (J+1) = J-1\\
&\text{Rank}([\mathbb{1}_n \mid A \mid x) - \text{Rank}([\mathbb{1}_n \mid A]) = (J+1) - J = 1\\
&\text{Rank}([\mathbb{1}_n \mid A]) - \text{Rank}(1_n) = J-1\\
&\text{Rank}([\mathbb{1}_n \mid A \mid x]) - \text{Rank}([1_n \mid x]) = (J+1) - 2 = J-1
\end{aligned}$$

➦ By proposition 4

$$\frac{(n-JK)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-JK).$$

➦ We conclude by the theorem 3 ("donuts" theorem) chapter 2. □

6.8. R example : Ancova Single factor model

Come back to the `Cepages` data set sudied in section 6.6. We want to explain Y =`pH` by the factor `Couleur` and the covariate `AcTot`.

➤➤ Let define the following ancova 2 factors model

$$Y = \mu \mathbb{1}_n + A\alpha + bx + c\boldsymbol{x} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(O_n, \sigma^2 \mathbb{I}_n),$$

➤➤ We use here the constraint by default under **R**.

$$\alpha_1 = c_1 = 0$$

```
modancova=lm(pH~Couleur*AcTot)
```

➤➤ We can test the influence of the regressors as follows

```
anova(modancova)
```

```
## Analysis of Variance Table
##
## Response: pH
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## Couleur        1 1.25814 1.25814  80.475 3.015e-10 ***
## AcTot          1 0.35643 0.35643  22.798 3.820e-05 ***
## Couleur:AcTot  1 0.00543 0.00543   0.347    0.5599
## Residuals     32 0.50029 0.01563
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(modancova)
```

```
## Anova Table (Type II tests)
##
## Response: pH
##                Sum Sq Df F value    Pr(>F)
## Couleur       0.31151  1  19.926 9.368e-05 ***
## AcTot         0.35643  1  22.798 3.820e-05 ***
## Couleur:AcTot 0.00543  1   0.347    0.5599
## Residuals     0.50029 32
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

➤➤ Whatever the tests (type I or type II), the interaction have no impact. Then, we select the following model without interaction

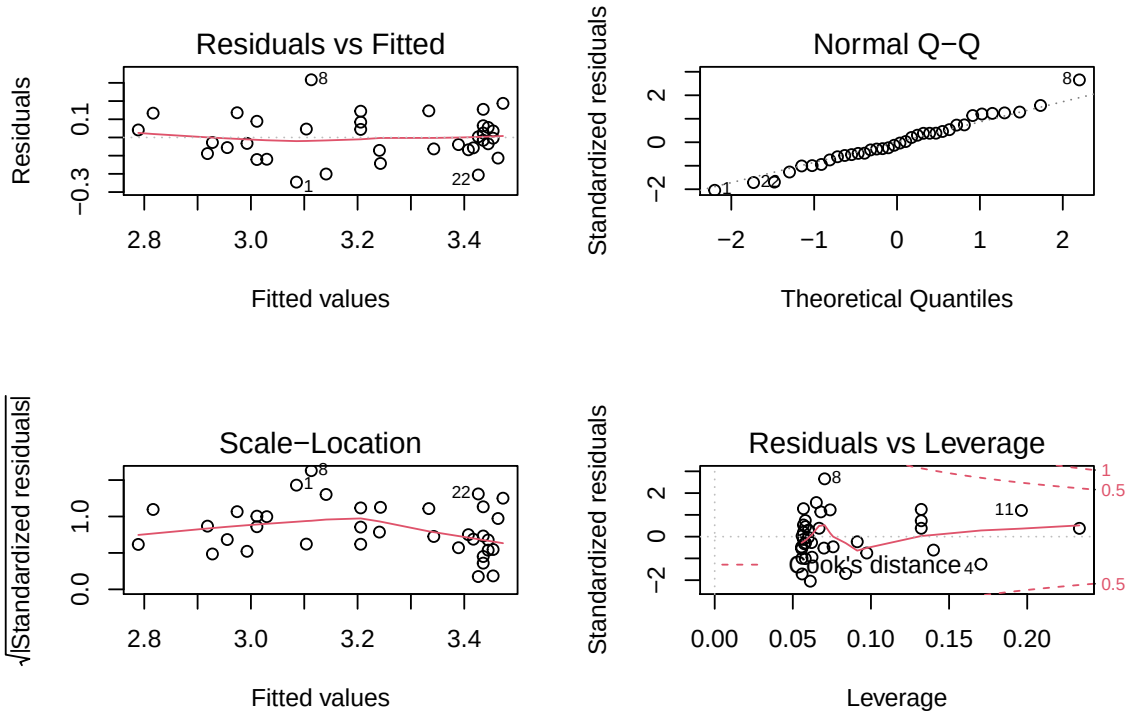
$$Y = \mu \mathbb{1}_n + A\alpha + bx + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(O_n, \sigma^2 \mathbb{I}_n),$$

```
modancovaWI=lm(pH~Couleur+AcTot);summary(modancovaWI)
```

```
##
## Call:
## lm(formula = pH ~ Couleur + AcTot)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24535 -0.06855 -0.00982  0.06938  0.31685
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.910142   0.182656  21.407  < 2e-16 ***
## CouleurRouge  0.229730   0.050953   4.509 7.79e-05 ***
## AcTot        -0.009267   0.001922  -4.823 3.11e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1238 on 33 degrees of freedom
## Multiple R-squared:  0.7615, Adjusted R-squared:  0.747
## F-statistic: 52.68 on 2 and 33 DF,  p-value: 5.357e-11
```

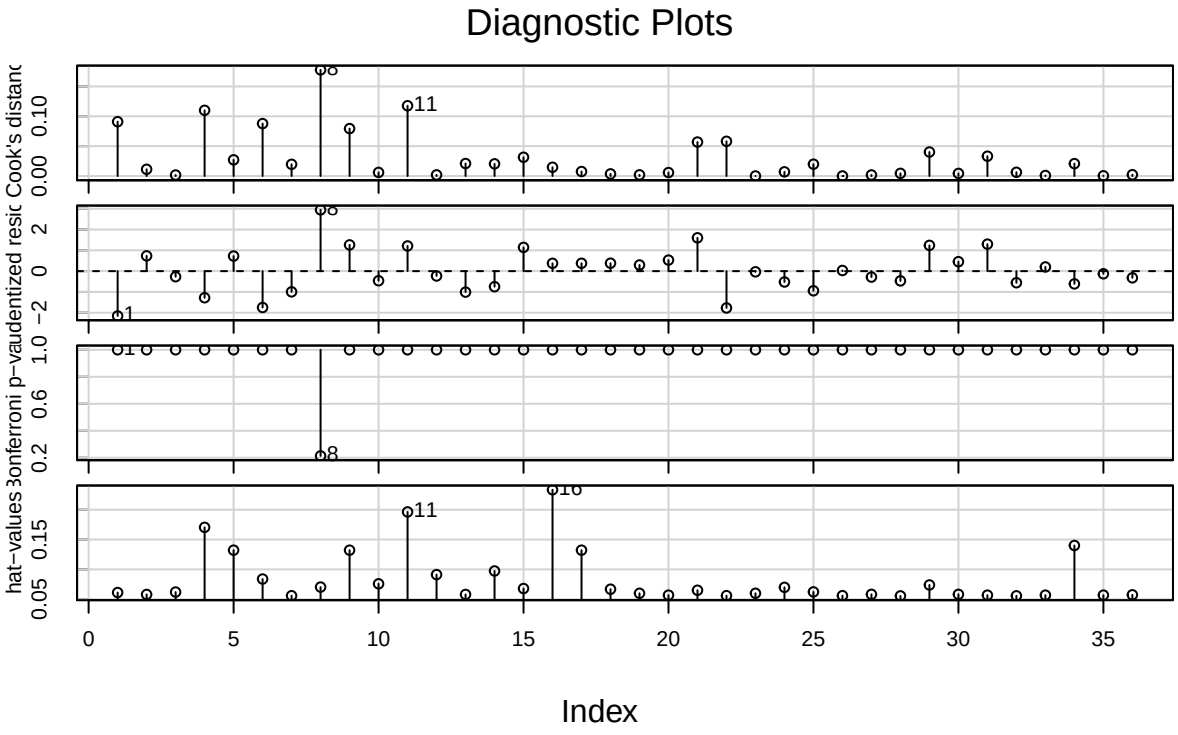
➤➤ We have to validate the model. Graphically, it can be done as follows

```
par(mfrow=c(2,2)); plot(modancovaWI)
```



➤➤ The postulates are validated. We can look for outliers to remove.

```
library(car);library(carData)
influenceIndexPlot(modancovaWI)
```




```
outlierTest(modancovaWI)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##   rstudent unadjusted p-value Bonferroni p
## 8 2.947663          0.0059344      0.21364
```

➤➤ There are no atypical points which need to be removed.