Introduction to Regression -Chapter 1

MAP 535

Contents

Chapter 1 : Introduction

This chapter will allow us to motivate this course. We are interested in the following dataset

https://www.economicswebinstitute.org/data/wagesmicrodata.xls

from the Economics Web Institute. In this dataset, we found datas such as economic characteristics about 534 persons:

- **WAGE :** Wage (dollars per hour).
- **OCCUPATION :** Occupational category (1=Management, 2=Sales, 3=Clerical, 4=Service, 5=Professional, 6=Other).
- **SECTOR :** Sector (0=Other,1=Manufacturing, 2=Construction).
- **UNION :** Indicator variable for union membership (1=Union member, 0=Not union member).
- **EDUCATION :** Number of years of education.
- **EXPERIENCE :** Number of years of work experience.
- **AGE :** Age (years).
- **SEX :** Indicator variable for sex (1=Female, 0=Male) .
- **MARR :** Marital Status (0=Unmarried, 1=Married).
- **RACE :** Race (1=Other,2=Hispanic, 3=White).
- **SOUTH :**Indicator variable for Southern Region (1=Person lives in South, 0=Person lives elsewhere).

An extract of the data is given in the Table~1

ID	WAGE	OC.	SECT.	EDUC.	EXPER.	AGE	SEX	MARR	RACE	SOUTH
1	510.00	6	1	8	21	35	1	1	2	0
2	495.00	6	1	9	42	57	1	1	3	0
3	667.00	6	1	12	1	19	0	0	3	0
4	400.00	6	0	12	4	22	0	0	3	0
5	750.00	6	0	12	17	35	0	1	3	0
6	1307.00	6	0	13	9	28	0	0	3	0

Table 1: Extract from the dataset "Wages"

The purpose of this study (and of the course) is to evaluate the possible effect of socio-demographic characteristics on the salary of employees.

1.1. Description of the data

Before any more elaborate statistical study, a descriptive study of the data must first be carried out. The study depends on the type of the considered variables (qualitative or quantitative).

1.1.1. Quantitative variables

For the quantitative variables, we will calculate for example the mean, standard deviation, median, extremal values…

For our dataset, these statistics are regrouped in the Table~2 or graphically as in Figure~1.

	MOYENNE	ECART TYPE	MINIMUN	MEDIAN	MAXIMUM
WAGE	902.41	513.91	100.00	778.00	4450.00
EDUCATION	13.02	2.62	2.00	12.00	18.00
AGE	36.83	11.73	18.00	35.00	64.00
EXPERIENCE	17.82	12.38	0.00	15.00	55.00

Table 2: Dataset "Wages": statistics for quantitative variables

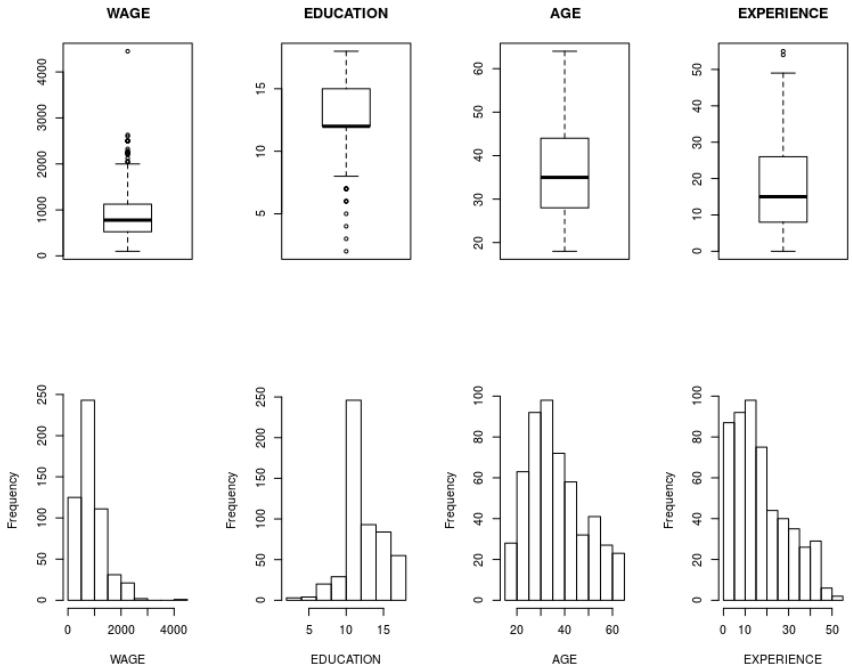


Figure 1: Dataset "Wages": Boxplots and histograms for quantitative variables

1.1.2. Qualitative variables

The qualitative variables are represented in the form of frequency tables (Table~3) or graphically as in Figure~2.

	Modalities	Effectifs	Frequencies (%)
OCCUPATION	1	55	10.30
	2	38	7.12
	3	97	18.16
	4	83	15.54
	5	105	19.66
	6	156	29.21
SECTOR	0	411	76.97
	1	99	18.54
	2	24	4.49
SEX	0	289	54.12
	1	245	45.88

Table 3: Dataset "Wages" : effectifs and frequencies for qualitative variables

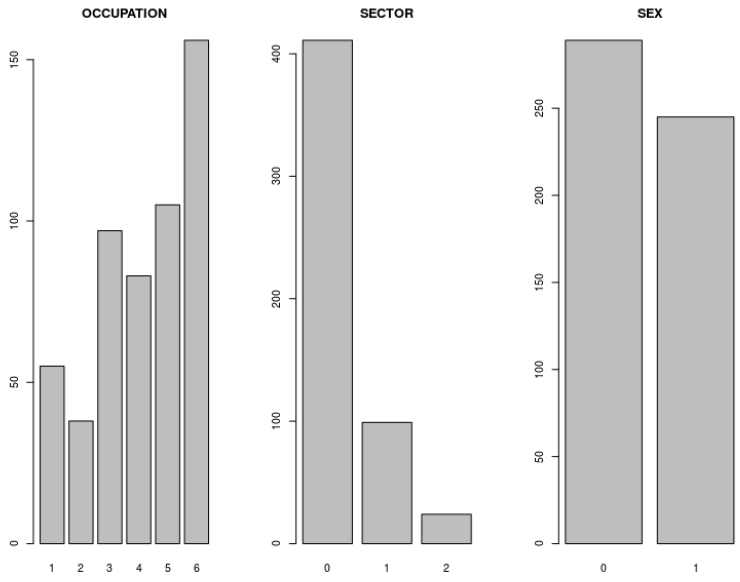


Figure 2: Dataset: "Wages": Barplot for qualitative variables

1.2. Linear regression

Now try to understand the influence of quantitative variables (`EDUCATION`, `AGE`, `EXPERIENCE`) on wages . We plot on the Figure~3 the three corresponding scatter plots.

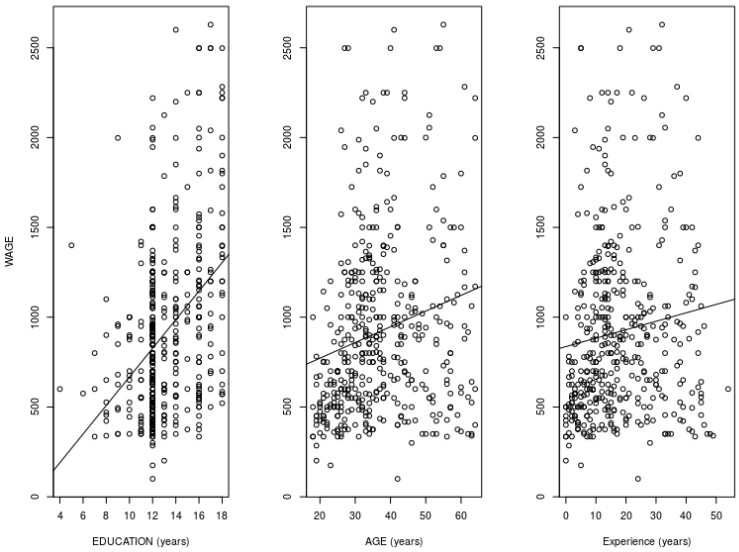


Figure 3: Dataset: "Wages": `WAGE` as functions of `EDUCATION`, `AGE` and `EXPERIENCE`.

Linear correlation coefficient

In order to quantify the linear relationship between two quantitative variables X and Y, the linear correlation coefficient can be calculated r_{XY}:

$$r_{X,Y} = \frac{s_{x,y}}{s_x s_y} = \frac{\frac{1}{n}\sum_{i,j}(y_i - \bar{y})(x_j - \bar{x})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{1}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. By construction, $|r_{XY}| \leq 1$. If the points are perfectly aligned then $|r_{XY}| = 1$. On this dataset, we get

r_{XY}	`EDUCATION`	`AGE`	`EXPERIENCE`
`WAGE`	0.40	0.21	0.12

Question : Is it big? small? Significant?

Comment:

- ☛ Statistical tests can be used (non-parametric tests based on the rank method). Under **R**, these tests are implemented in the function `cor.test` in which to specify the desired test (`method = "kendall" or "spearman"`).

Consider again the scatter plot on the left of Figure~3. If we try to summarize it by a line, called *simple linear regression line*, we write:

$$\text{Wage}_i = \beta_0 + \beta_1 \text{Education}_i + e_i$$

where e_i is an error term between the line of the observation y_i.

In this very simple model, the variable of interest (quantitative) is explained by a quantitative variable called *explonatory variable* (or *covariate* or *predictor* or *regressor*...). The slope (β_1) and the intercept (β_0) of the line are *estimated* from the observations to properly "set" the line. In this course, we will see how to estimate these parameters, what are the properties of such an estimator. In addition, we want to know if the slope is significantly different from 0, *i.e.* we will try to write tests on the parameters of the models.

We can also try to explain the salary as a linear combination of the other quantitative variables:

$$\begin{aligned} \text{Wage}_i &= \beta_0 + \beta_1 \text{Education}_i + \beta_2 \text{Age}_i + \beta_3 \text{ Experience}_i + e_i \\ &= \beta_0 + \sum_{k=1}^{K} \beta_k x_{ik} + e_i \end{aligned}$$

where x_{ik} is the value of the k-th predictor variable of the individual i. Then, we speak of *multiple linear regression*. The same questions as before are: Is β_k significance? How to select the most relevant predictor variables? And so on.

1.3. Analysis of the variance

Analysis of the variance with one factor

It may also be interesting to study the relationship between the `WAGE` (variable of quantitative interest) and the qualitative variables, for example `SEX`, or `OCCUPATION` (type of occupation). Graphically, we can draw boxplots by modality of the qualitative variable as in Figure~4.

In a natural way, to compare the wages within the different populations, one will try to compare the averages within the groups:

OCCUPATION	1	2	3	4	5	6
Mean by OCCUPATION	1254.27	790.68	758.99	591.87	1205.73	879.75

SEX	0	1
Mean by SEX	1032.7	788.47

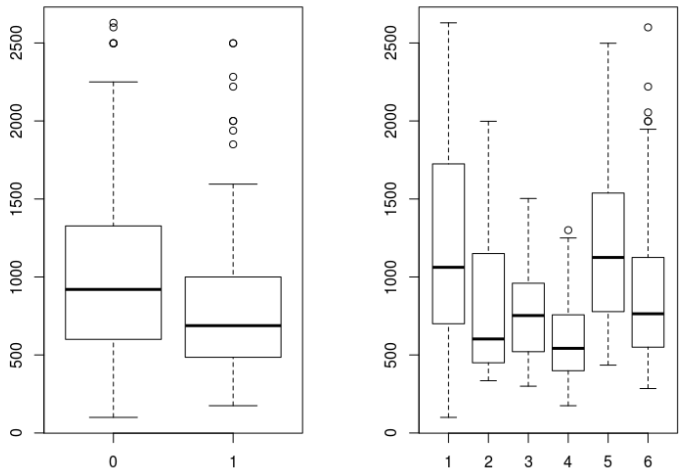


Figure 4: Dataset: "Wages": Wages as functions of `SEX` (0 = Male) or `OCCUPATION` category.

Tests which compare means can then be applied. However, it is actually possible to write a linear model to study salary according to `SEX`:

$$\text{Wage}_i = \underbrace{\beta_0}_{\text{Wage of Male}} \mathbb{1}_{\text{Sex}_i=0} + \underbrace{\beta_1}_{\text{Wage of Female}} \mathbb{1}_{\text{Sex}_i=1} + e_i.$$

Similarly, we can write a linear model to study salary according to `OCCUPATION`:

$$\text{Wage}_i = \sum_{l=1}^{l} \underbrace{\beta_l}_{\text{Wage of Occupation } l} \mathbb{1}_{\text{Occupation}=l} + e_i.$$

These models are called Anova models one factor[1]. We want here to explain the variation of the `WAGE` according to a single factor (`SEX` or `OCCUPATION`).

Analysis of the variance with two factors

One may wonder if there is not a joint effect of the two factors. We will then seek to cross factors, for example by calculating mean of `WAGE` as follows:

We will try to study the influence of each factor on `WAGE` but also their joint influences (possible interactions). We will see in this course how to write and study a linear model.

[1] *Analysis of variance model with with a single factor*

	OCCUPATION	1	2	3	4	5	6	Mean by SEX
SEX	0	1442.11	972.05	795.29	612.52	1277.87	941.92	1032.71
	1	944.88	531.57	750.92	580.74	1132.02	605.17	788.47
Mean by	OCCUPATION	1254.27	790.68	758.99	591.87	1205.73	879.75	919.77

1.4. Analysis of covariance

One can think that the link between WAGE and EDUCATION is not the same depending on whether one is a MALE or a FEMALE. We will then want to write a regression model for each sex (see Figure~5.) In the same way, we will see in this course that it is possible to write a linear model also answering this need.

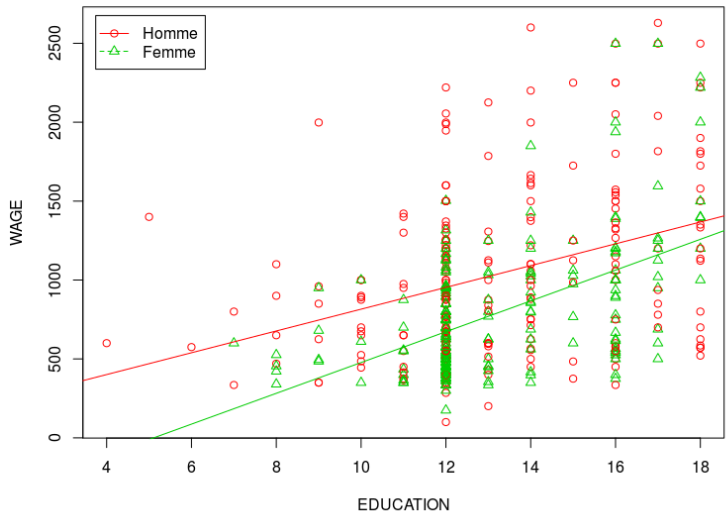


Figure 5: Dataset: "Wages": Wages as functions of SEX and EDUCATION

1.5. Matrix form

Let $y_1, \ldots, y_n$, n independent observations of a quantitative variable. For each observation/individu i, we have p real quantities $(x_{i1}, \ldots, x_{ip})$ (quantitative or indicator). We try to explain **the response** y_i as a linear function of the p **predictors** $(x_{i1}, \ldots, x_{ip})$. So, for all i, we write

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i, \qquad \forall i = 1, \ldots, n \tag{2}$$

where e_i is the error term.

Set:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \underbrace{\begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{i1} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}}_{p+1 \text{ columns}}, \quad \boldsymbol{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}. \tag{3}$$

Then, we can write a matrix version of the equations (2).

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e} \tag{4}$$

$\boldsymbol{y}$ and X are observed. The parameter $\boldsymbol{\beta}$ is unknown and must be estimated. Estimation and statistical tests will be the subject of next chapters.

Introduction to Regression - Chapter 2

MAP 535

Contents

Chapter 2 : Linear regression model

Let formalize the ideas seen in the previous chapter, *i.e.* write a probabilistic model to model the data, study the properties of this model, estimate the parameters and study the properties of these estimators (bias, variance, properties asymptotic ...). Finally we will develop decision tools (confidence intervals, tests).

Modeling the observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ means we suppose that $\boldsymbol{y}$ is the realization of a random variable $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$ whose probability law is described.

2.1. Introduction

The aim of this course is to study the relation between a random variable, denoted Y, with other variables $X_1, \ldots, X_p$ called *predictors* (*regressor* or *explanatory variables*...). We focus on two objectives : **explain** and **predict**. For this, we try to construct a function f such that

$$Y \approx f(X_1, \ldots, X_p).$$

The construction will be based on the n observations $(y_i)_{i=1,\cdots,n}$ of the variable Y and n observations $(X_{i1}, \cdots, X_{ip})_{i=1,\cdots,n}$ of the variables $X_1, \ldots, X_p$.

Example 1 *We try to explain and predict gasoline consumption (in liters per 100 km) of different automobile models based on several variables. For this, we have the following characteristics for 31 different cars:*

- `Consommation` = *Fuel consumption in liters per 100 km.*
- `Prix` = *Vehicle price in Swiss francs.*
- `Cylindree` = *Cylinder capacity in cm3.*
- `Puissance` = *Power in kW.*
- `Poids` = *Weight in kg.*

In this example, Y is the variable `Consommation`*. The variables X_j correspond to the other 4 variables.*

In this course, we are interested in the *linear model*, specifying a linear relation between the observed variable Y and the predictors stocked in the matrix X, so f is a linear function in our course. The simple framework of linear regression makes it possible to obtain very rich results, which justifies its indepth study and its widespread use among practitioners.

2.2. Modelization

In the linear regression setting, the function f is linear. It's mean that there exists $\beta = (\beta_1, \ldots, \beta_p)^T$ such that for all $(X_1, \ldots, X_p)$,

$$f(X_1, \ldots, X_p) = \beta_1 X_1 + \cdots + \beta_p X_p.$$

The vocabulary is the following :

- The variable to explain Y is called the **response variable**,
- The explanatory variables X_j are called the **predictors** or **regressor** or **covariates**,...
- The β_j are the **regression coefficients**,

Regression linear model

We can write the regression linear model as follows:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, \ldots, n.$$

Comments:

- ☛ The ε_i are called the **error**.
- ☛ For each i, we modelize $\varepsilon_i = Y_i - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_p X_{ip}$ as a random variable. This stochastic term models the measurement errors and the impact of all the variables not taken into account by the model. We will see that the hypothesis $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ will often be assumed.
- ☛ The value of coefficients β_j measure the importance of the effect of each predictors X_j on the variable to be explained Y. We will see how to test if a predictor has a significant influence on the variable to explain, we say that the predictor is relevant. But also and above all, this model is constructed to predict the typical values that can take a new observation Y_{n+1} of which we only know the values of the associated predictors.

Matrix form of the regression linear model

Let us denote $X_j = (X_{1j}, \ldots, X_{nj})^T \in \mathbb{R}^n$ and $x_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p$.

Then define the matrix X of size $n \times p$ such that

$$X = \underbrace{(X_1, \cdots, X_p)}_{p \text{ predictors}} = \underbrace{\begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_i^T \\ \vdots \\ x_n^T \end{pmatrix}}_{n \text{ observations}} = \underbrace{\begin{pmatrix} X_{11} & \cdots & X_{1p} \\ X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots \\ X_{i1} & \cdots & X_{ip} \\ \vdots & \vdots & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix}}_{n \text{ lines} \times p \text{ columns}}.$$

When the first column is only composed by 1, the β_1 parameter is called *intercept*

$$X = \begin{pmatrix} 1 & X_{12} & \cdots & X_{1p} \\ 1 & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{i2} & \cdots & X_{ip} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n2} & \cdots & X_{np} \end{pmatrix}$$

By using the matrix forms, we obtain the following definition of the regression linear model.

$$Y = X\beta + \varepsilon,$$

where
- $Y = (Y_1, \ldots, Y_n)^T$ is a random vector in $\mathbb{R}^n$.
- $\beta = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ is the unknown parameters vector.
- $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^n$ is the vectors of errors.
- The matrix X is supposed to be deterministic (not random) in this course.

Identifiability problems : The model is identifiable if the matrix X satisfies the following property: if $\beta \in \mathbb{R}^p$ and $\beta' \in \mathbb{R}^p$ satisafy $X\beta = X\beta'$ then $\beta = \beta'$. We have then, uniqueness of the vector of the parameters.This is a very important hypothesis and is realized as soon as the matrix X is injective, which is equivalent to $\text{rang}(X) = p$. Most of the time, we'll assume $p \leq n$ and $\text{rang}(X) = p$. Therefore, we will make the following hypothesis.

Rank assumption : The matrix X is full rank, $\text{rang}(X) = p$.

Comments:

- ☛ Most of the time, we'll assume ε centered. Other hypotheses (**Postulates**) (gaussianity, homoscedasticity, ...) can also be formulated.
- ☛ In the introductory example, we could have used other predictors, such as the square of the logarithm of one of the predictor, for example. This is possible by involving other variables.
- ☛ Thus, a linear model does not mean that the relationship between the predictors and the response variable is linear, but that the model is linear in the β_j parameters.

Definition 1

• *We define the* ***linear regression model*** *as follow :*

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

• *The matrix form is the following*

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon.$$

✎ *If we considère an intercept in our model, our model is written in its matrix form*

$$Y = \beta_0 \mathbb{1}_n + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon \text{ or } Y = \beta_1 \mathbb{1}_n + \beta_2 X_2 + \cdots + \beta_p, X_p + \varepsilon.$$

where $\mathbb{1}_n = (1, \cdots, 1)^T \in \mathbb{R}^n$.

2.3. Assumed postulates

Note that we speak of *postulates* in the sense that we can not formally show that they are verified by statistical tests. We will use graphical tools to test them.

Postulat [P1] : $\forall i = 1, \cdots, n \quad \mathbb{E}_\beta[\varepsilon_i] = 0.$

Errors are centered. In practice, this means that the model is correct and that we have not forgotten a relevant term : the model is linear.

Postulat [P2] : $\forall i = 1, \cdots, n \quad \mathbb{V}\text{ar}_\beta[\varepsilon_i] = \sigma^2 > 0.$

The errors are of constant variance. We speak of a *homoscedastic* model, as opposed to a *heteroscedastic* model where the error term would not have the same variance for all observations.

Postulat [P3] : $\forall i \neq j \quad \mathbb{C}\text{ov}(\varepsilon_i, \varepsilon_j) = 0.$

Errors are uncorrelated. Thus, the observations are assumed to be uncorrelated, *i.e.* independent sampling or the results of a physical experiment conducted under independent conditions. Problems can arise when time has an importance in the phenomenon.

Postulat [P4] : $\forall i = 1, \cdots, n \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$

The errors are Gaussian. This postulate is the least important as we will be able to do without it if the number of observations is large (larger than 20 or 30 observations). It is difficult to detect the non gaussianity of errors. We will see that **R** propose graphical tools to try to validate or not this postulate.

2.4. Estimation of the model under the assumption of rank

Some notations:

- Let X be the *design* matrix of size $n \times p$ and of full rank p, we denote by

$$[X] := \mathcal{I}m(X)$$

  the space generated by the p columns of X.

- Let denote by P_X the orthogonal projection into $[X]$. Then, $P_{X^\perp} = I - P_X$ is the orthogonal projection matrix into $[X]^\perp$ the orthogonal space at $[X]$.

Definition 2 *We define the **ordinary least square estimator (OLSE)** of β in the model* (1)*, the vector $\widehat{\beta} \in \mathbb{R}^p$ such that*

$$\widehat{\beta} = \arg\min_{u \in \mathbb{R}^p} \|Y - Xu\|^2.$$

✎ *Where $\|u\|^2 = \sum_k u_k^2$ denote the euclidian norm*

Proposition 1 *In the model* (1) *and under the Rank assumption, the design matrix X is injective and*

$$\widehat{\beta} = (X^T X)^{-1} X^T Y.$$

Proof : By the definition of the projected orthogonal, we have

$$\|Y - P_X Y\|^2 = \min_{v \in [X]} \|Y - v\|^2.$$

Thus,

$$P_X Y = \arg\min_{v \in [X]} \|Y - v\|^2.$$

Which implies $\exists\widehat{\beta} \in \mathbb{R}^p$ such that $P_X Y = X\widehat{\beta}$. For all $k \in \{1, \cdots, p\}$, we note X_k the $k-th$ column of X, then it comes

$$\langle X_k, P_X Y - Y\rangle = 0 \Leftrightarrow \langle X_k, X\widehat{\beta} - Y\rangle = 0 \Leftrightarrow X_k^T(X\widehat{\beta} - Y) = 0.$$

Thus,

$$X^T(X\widehat{\beta} - Y) = 0_p \quad \Leftrightarrow \quad X^T X\widehat{\beta} = X^T Y$$

As X is full rank, X^TX$ is inversible and for all $u \in \mathbb{R}^p \Leftrightarrow Xu \in [X]$. As we consider the Euclidean norm, $X\widehat{\beta}$ is the orthogonal projection of Y into $[X]$:

$$X\widehat{\beta} = X(X^T X)^{-1} X^T Y.$$

By injectivity of X, we get the result $\widehat{\beta} = (X^T X)^{-1} X^T Y$. □

Exercise 1 *Proof the result using the calculation of the partial derivatives of*

$$u \longmapsto \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{p} u_j X_{ij})^2.$$

Comments:

- It may be noted that we have shown the relation:

$$P_X Y = X\widehat{\beta} = \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_p X_p.$$

- Note that

$$\widehat{Y} = P_X Y \text{ and } P_X = X(X^T X)^{-1} X^T.$$

Exercise 2 *To which condition(s) the formula $P_{X_j} Y = \widehat{\beta}_j X_j$ is true?*

2.5. Statistical Properties of OLSE under Postulates [P1] - [P3]

Study now the statistical properties of the EMCO:

Proposition 2 *In the model* (1)*, under the Rank assumption and under **[P1]**, the OSLE $\widehat{\beta}$ is an unbiased estimator of β:*

$$\forall \beta \in \mathbb{R}^p, \quad \mathbb{E}_\beta[\widehat{\beta}] = \beta.$$

Proof : As X is a deterministic matrix and $Y = X\beta + \varepsilon$, it comes

$$\mathbb{E}_\beta[\widehat{\beta}] = \mathbb{E}_\beta[(X^T X)^{-1} X^T Y] = \mathbb{E}_\beta[(X^T X)^{-1} X^T (X\beta + \varepsilon)] = \underbrace{(X^T X)^{-1} X^T X}_{identity}\beta + (X^T X)^{-1} X^T \underbrace{\mathbb{E}_\beta[\varepsilon]}_{0_n} = \beta. \quad \square$$

Theorem 1 (Gauss-Markov theorem) *In the model* (1)*, under the Rank assumption and under **[P1]–[P3]**, the OLSE $\widehat{\beta}$ is such that*

$$\mathbb{V}ar_\beta(\widehat{\beta}) = \sigma^2(X^TX)^{-1}$$

*and $\widehat{\beta}$ is the estimator of "minimum variance" **among linear and unbiased estimators.***

Proof :

• Note that $\mathbb{V}\text{ar}_\beta(Y) = \mathbb{V}\text{ar}_\beta(X\beta + \varepsilon) = \mathbb{V}\text{ar}_\beta(\varepsilon) = \sigma^2\mathbb{I}_n$, then

$$\mathbb{V}\text{ar}_\beta(\widehat{\beta}) = \mathbb{V}\text{ar}_\beta\left((X^TX)^{-1}X^TY\right) = (X^TX)^{-1}X^T\underbrace{\mathbb{V}\text{ar}_\beta(Y)}_{\sigma^2\mathbb{I}_n}X(X^TX)^{-1} = \sigma^2\underbrace{(X^TX)^{-1}X^TX}_{\mathbb{I}_p}(X^TX)^{-1} = \sigma^2(X^TX)^{-1}.$$

• Let $\tilde{\beta} = CY$ be any linear and unbiased estimators where C is a matrix of size $p \times n$ then as $\tilde{\beta}$ is unbiased we have:

$$\mathbb{E}_\beta[CY] - \beta = 0 \Leftrightarrow CX\beta - \beta = 0_p \Leftrightarrow (CX - \mathbb{I}_p)\beta = 0_p \Leftrightarrow CX = \mathbb{I}_p \text{ and } \mathrm{X^TC^T} = \mathbb{I}_\mathrm{p}$$

Therefore,

$$\begin{aligned}\mathbb{V}\text{ar}_\beta(\tilde{\beta}) - \mathbb{V}\text{ar}_\beta(\widehat{\beta}) &= \sigma^2(CC^T - (X^TX)^{-1}) = \sigma^2C(\mathbb{I}_n - X(X^TX)^{-1}X^T)C^T \\ &= \sigma^2C(\mathbb{I}_n - P_X)C^T = \sigma^2CP_{X^\perp}C^T = \sigma^2CP_{X^\perp}P_{X^\perp}^TC^T,\end{aligned}$$

$P_{X^\perp}$ is the orthogonal projection into the orhtogonal space at $[X]$. Then, $\forall u \in \mathbb{R}^p$ we get

$$u^T(\mathbb{V}\text{ar}_\beta(\tilde{\beta}) - \mathbb{V}\text{ar}_\beta(\widehat{\beta}))u = \sigma^2u^TCP_{X^\perp}P_{X^\perp}^TC^Tu = \|P_{X^\perp}^TC^Tu\|^2 \geq 0. \quad \square$$

2.6. General Cochran theorem

Theorem 2 (General Cochran theorem) *Let $W \sim \mathcal{N}(m, \sigma^2I_d)$ be a gaussian vector in $\mathbb{R}^d$ and $E_1 \oplus \cdots \oplus E_r$ a decomposition of $\mathbb{R}^d$ into two-by-two orthogonal subspaces of dimension $d_1, \ldots, d_r$. For all $j = 1, \cdots, r$, we define the random vectors $W_{E_1}, \ldots, W_{E_r}$ such that*

$$W_{E_j} = P_{E_j}\left(\frac{W - m}{\sigma}\right)$$

is the orthogonal projection of $\frac{W-m}{\sigma}$ into E_j. Then:

The random vectors W_{E_j} are mutually independent. For all $j = 1, \cdots, r$, the random vectors $\|W_{E_j}\|^2$ are mutually independent and $\|W_{E_j}\|^2 \sim \chi^2_{d_j}$.

Comment:

☛ Cochran's theorem is most often applied with $W \sim \mathcal{N}(0, \sigma^2 I_d)$, $\mathbb{R}^d = E_1 \oplus E_2$ and $E_2 = E_1^{\perp}$. In this case, W_{E_1} and $W_{E_1^{\perp}}$ are indenpendent vectors and

$$\sigma^{-2}\|W_{E_1}\|^2 \sim \chi^2_{d_1}, \quad \sigma^{-2}\|W_{E_1^{\perp}}\|^2 \sim \chi^2_{d-d_1}.$$

2.7. Residuals and variance estimation

Recall first that the vector ε is the vector of errors. It is also called the vector of theoretical residues.

Definition 3 *In the model* (1)*, we define the* ***residuals*** *(or estimated residuals) as follows:*

$$\widehat{\varepsilon} = Y - \widehat{Y} = Y - X\widehat{\beta} = Y - P_X Y = (I - P_X)Y = P_{X^{\perp}} Y = P_{X^{\perp}}\varepsilon.$$

Proposition 3 *In the model* (1)*, under the Rank assumption and under* ***[P1]–[P3]****, we have*

- $\mathbb{E}_\beta[\widehat{\varepsilon}] = 0_n$
- $\mathbb{V}ar_\beta[\widehat{\varepsilon}] = \sigma^2 P_{X^{\perp}}$.
- $\mathbb{C}ov_\beta(\widehat{\varepsilon}, \widehat{Y}) = 0_{n\times n}$.

Proof : By using the fact that $\widehat{\varepsilon} = P_{X^{\perp}} Y$, the two first points are obvious. Then, as ε is centered

$$\mathbb{C}\mathrm{ov}_\beta(\widehat{\varepsilon}, \widehat{Y}) = \mathbb{C}\mathrm{ov}_\beta(P_{X^{\perp}} Y, P_X Y) = 0_{n\times n}. \quad \square$$

Proposition 4 *In the model* (1)*, under the Rank assumption and under* ***[P1]–[P3]****, an unbiased estimator of* σ^2 *is* $\widehat{\sigma}^2$ *defined by*

$$\widehat{\sigma}^2 = \frac{\|\widehat{\varepsilon}\|^2}{n-p} = \frac{\|Y - X\widehat{\beta}\|^2}{n-p} = \frac{\|Y - P_X Y\|^2}{n-p} = \frac{\|\widehat{\varepsilon}\|^2}{n-p} = \frac{\|P_{X^{\perp}} Y\|^2}{n-p}.$$

Proof : Note that $\sigma^{-1}\varepsilon \sim \mathcal{N}(O_n, \mathbb{I}_n)$ and $P_{X^{\perp}}$ is an orthogonal projector of rank $(n-p)$ then by Cochran theorem

$$\|P_{X^{\perp}}(\sigma^{-1}\varepsilon)\|^2 \sim \chi^2_{(n-p)}.$$

Then, it comes :

$$\mathbb{E}_\beta[\|\widehat{\varepsilon}\|^2] = \mathbb{E}_\beta[\|P_{X^{\perp}}\varepsilon\|^2] = \sigma^2 \mathbb{E}_\beta[\|P_{X^{\perp}}(\sigma^{-1}\varepsilon)\|^2] = \sigma^2(n-p)$$

as the expectation of the Chi-2 random variable of $(n-p)$ degree of freedom. $\square$

2.7. Inference in the Gaussian linear regression model

In this section, we assume the Rank assumption and **[P1]–[P4]** satisfied, then

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

In others words the ε_i are i.i.d. This assumption allow us to consider maximum likelihood estimators and to build confidence regions and tests.

Comments:

- ☛ In this section, $\widehat{\beta}$ is the ordinary maximum likelihood estimator which is in this gaussian setting exactly the least squares estimator.
- ☛ Note that the maximum likelihood estimator of σ^2 is $\widehat{\sigma}^2_{ML} = \frac{\|Y - X\widehat{\beta}\|^2}{n} = \frac{\|\widehat{\varepsilon}\|^2}{n}$. which is a biased estimator. We recall that an unbiased estimator of σ^2 is $\widehat{\sigma}^2 = \frac{\|Y - P_X Y\|^2}{n-p} = \frac{\|Y - X\widehat{\beta}\|^2}{n-p} = \frac{\|\widehat{\varepsilon}\|^2}{n-p}$.

2.7.1. Maximum likelihood estimator and its properties

Proposition 5 *the maximum likelihood estimatorof* (β, σ^2) *est le vecteur*

$$(\widehat{\beta}, \widehat{\sigma}^2_{ML}) = \left(X^T X)^{-1} X^T Y, (n-p)/n \times \widehat{\sigma}^2\right) = \left(X^T X)^{-1} X^T Y, \frac{\|Y - X\widehat{\beta}\|^2}{n}\right)$$

Proof : The demonstration follows from the calculation. □

Proposition 6 *In the model* (1)*, under the Rank assumption and under **[P1]–[P4]**, we have*

- $\widehat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1})$
- $\frac{(n-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p)$
- $\widehat{\beta}$ *and* $\widehat{\sigma}^2$ *are indenpendent.*

Proof : The first point is obvious. For the second point, we can write :

$$\frac{(n-p)\widehat{\sigma}^2}{\sigma^2} = \frac{\|Y - P_X Y\|^2}{\sigma^2} = \frac{\|P_{X^\perp} Y\|^2}{\sigma^2} = \left\|P_{X^\perp}(\sigma^{-1}\varepsilon)\right\|^2.$$

We conclude by Cochran theorem. Then, for the last point we show that

$$\widehat{\sigma}^2 = \frac{\|P_{X^\perp}\varepsilon\|^2}{n-p} \quad \text{and} \quad \widehat{\beta} = \beta + (X^T X)^{-1} X^T P_X \varepsilon.$$

Then

$$\mathbb{C}\text{ov}_\beta(\widehat{\sigma}^2, \widehat{\beta}) = \frac{(X^T X)^{-1} X^T}{n-p} \mathbb{C}\text{ov}_\beta\left(\|P_{X^\perp}\varepsilon\|^2, P_X \varepsilon\right) = O_p. \quad \square$$

The following result will be very important for building trust regions and hypothesis testing.

Theorem 3 *In the model* (1)*, under the Rank assumption and under* ***[P1]–[P4]****, let* $\widehat{\sigma}^2$ *and* $\widehat{\beta}$ *the estimator defined in proposition 6.*

• *For all vectors* $c \in \mathbb{R}^p$*, we have*

$$\frac{c^T\widehat{\beta} - c^T\beta}{\widehat{\sigma}\sqrt{c^T(X^TX)^{-1}c}} \sim t_{(n-p)}.$$

• *For all matrices* C *of size* $q \times p$ *and of full rank* q $(q \leq p)$*, we have*

$$\frac{(C\widehat{\beta} - C\beta)^T(C(X^TX)^{-1}C^T)^{-1}(C\widehat{\beta} - C\beta)}{q\widehat{\sigma}^2} \sim \mathcal{F}(q, n-p).$$

• *We set two vector subspaces* V *and* W *where* W *is a vector subspace of* V*. We assume* $q = dim(W) < p = dim(V)$*. If* $X\beta \in W \subset V$*, then*

$$F = \frac{\|P_WY - P_VY\|^2/(p-q)}{\|Y - P_VY\|^2/(n-p)} \sim \mathcal{F}(p-q, n-p),$$

where P_VY *denote the orthogonal projection of* Y *into* V *and* P_WY *denote the orthogonal projection of* Y *into* W*.*

Proof :

• The first point is immediate by applying a probability result

$$\begin{cases} A \sim \mathcal{N}(0,1), \\ B \sim \chi^2(n-p) \\ A \text{ independent of } B \end{cases} \Leftrightarrow \frac{A}{\sqrt{B/(n-p)}} \sim t_{(n-p)}.$$

• For the second point : As the rank of C is $q \leq p$, the matrix, of size $q \times q$ and of rank q, $C(X^TX)^{-1}C^T$ is inversible. There exists a inversible and symetric matrix Δ such that $C(X^TX)^{-1}C^T =: \Delta^2$. Then $\Delta^{-1}(C\widehat{\beta} - C\beta) \sim \mathcal{N}(0, \sigma^2 I)$. And finally by using the following probability result

$$\begin{cases} A \sim \mathcal{N}(0_q, \mathbb{I}_q), \\ B \sim \chi^2(n-p) \\ A \text{ independent of } B \end{cases} \Leftrightarrow \frac{\|A\|^2/q}{B/(n-p)} \sim F(q, n-p).$$

$$\frac{(C\widehat{\beta} - C\beta)^T(C(X^TX)^{-1}C^T)^{-1}(C\widehat{\beta} - C\beta)}{q\widehat{\sigma}^2} \sim \mathcal{F}(q, n-p).$$

• The third point stems from the application of Cochran's theorem which implies that

$$P_WY - P_VY \underbrace{=}_{as\ X\beta \in W} P_W\varepsilon - P_V\varepsilon \in V$$

is independent of $Y - P_VY = P_{V^\perp}\varepsilon$. □

2.7.2. Intervals and regions of confidence

By using the previous theorem, we can determine the intervals and confidence regions of the unknownparameters. We have the following theorem which is based on the Theorem~3 :

Theorem 4 *In the model* (1)*, under the Rank assumption and under **[P1]–[P4]**, let $\widehat{\sigma}^2$ and $\widehat{\beta}$ the estimator defined in proposition 6. Let $\alpha \in]0,1[$.*

• *For all $c \in \mathbb{R}^p$,*

$$\left[c^T\widehat{\beta} - t_{n-p,1-\alpha/2}\widehat{\sigma}\sqrt{c^T(X^TX)^{-1}c}\ ,\ c^T\widehat{\beta} + t_{n-p,1-\alpha/2}\widehat{\sigma}\sqrt{c^T(X^TX)^{-1}c}\right],$$

where $t_{n-p,1-\alpha/2}$ is the quantile of order $1-\alpha/2$ the Student's law at $n-p$ degrees of freedom, is a confidence interval for $c^T\beta$ of exactly $1-\alpha$ level.

• *Let q_1 and q_2 such that if $Z \sim \chi^2(n-p)$, then $P(q_1 \leq Z \leq q_2) = 1-\alpha$. Then*

$$\left[\frac{(n-p)\widehat{\sigma}^2}{q_2}\ ,\ \frac{(n-p)\widehat{\sigma}^2}{q_1}\right]$$

is a confidence interval for σ^2 of exactly $1-\alpha$ level.

• *Let C be a matrix of size $q \times p$ and rank $q \leq p$. Then*

$$R = \left\{a : \quad \frac{(C\widehat{\beta}-a)^T(C(X^TX)^{-1}C^T)^{-1}(C\widehat{\beta}-a)}{q\widehat{\sigma}^2} \leq f_{q,n-p,1-\alpha}\right\}$$

with $f_{q,n-p,1-\alpha}$ the quantile of order $1-\alpha$ of the Fisher law at $(q, n-p)$ degrees of freedom, is a confidence region (set) for $C\beta$ of exactly $1-\alpha$ level.

Proof : Immediate. □

Corollary 1 *In the model* (1)*, under the Rank assumption and under **[P1]–[P4]**, let $\widehat{\sigma}^2$ and $\widehat{\beta}$ the estimator defined in proposition 6. Let $\alpha \in]0,1[$. Then for all $j = 1, \cdots, p$*

$$\left[\widehat{\beta}_j - t_{n-p,1-\alpha/2}\widehat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}\ ,\ \widehat{\beta}_j + t_{n-p,1-\alpha/2}\widehat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}\right],$$

where $t_{n-p,1-\alpha/2}$ is the quantile of order $1-\alpha/2$ the Student's law at $n-p$ degrees of freedom, is a confidence interval for β_j of exactly $1-\alpha$ level.

Proof : Direct application of the first point of the theorem~4 with $c \in \mathbb{R}^p$ such that $c_k = 0$ for all $k \neq j$ et $c_j = 1$. □

Corollary 2 *In the model* (1)*, under the Rank assumption and under **[P1]–[P4]**, let $\widehat{\sigma}^2$ and $\widehat{\beta}$ the estimator defined in proposition 6. Let $\alpha \in]0,1[$. Let $x_0 = (x_{01}, \cdots, x_{0p})^T$, then*

$$\left[x_0^T\widehat{\beta} - t_{n-p,1-\alpha/2}\widehat{\sigma}\sqrt{x_0^T(X^TX)^{-1}x_0}\ ,\ x_0^T\widehat{\beta} + t_{n-p,1-\alpha/2}\widehat{\sigma}\sqrt{x_0^T(X^TX)^{-1}x_0} \right],$$

where $t_{n-p,1-\alpha/2}$ is the quantile of order $1-\alpha/2$ the Student's law at $n-p$ degrees of freedom, is a confidence interval for $x_0\beta = \mathbb{E}_\beta[Y_0]$ of exactly $1-\alpha$ level.

Proof : The proof is left in exercice. □

Exercise 3 *Assume $p \geq 2$ and set $c_{ij} := ((X^TX)^{-1})_{ij}$.*

1. *Determine a confidence interval for β_1 and β_2 of level $1-\alpha$ based on the $\widehat{\beta}$, $\widehat{\sigma}$ and the c_{ij}.*
2. *Deduce a confidence region for the vector (β_1, β_2).*
3. *Answer the previous question but using the 3rd point of the theorem.*

2.7.3. Hypothesis tests

We want to first test

$$H_0 :\ c^T\beta = a \quad \text{vs} \quad H_1 :\ c^T\beta \neq a$$

for $c \in \mathbb{R}^p$ and $a \in \mathbb{R}$. Note that, within this framework, there is the following test

$$H_0 :\ \beta_j = 0 \quad \text{vs} \quad H_1 :\ \beta_j \neq 0, \qquad j \in \{1, \ldots, p\}.$$

Theorem 5 *Let $c \in \mathbb{R}^p$ and $\alpha \in]0,1[$, we set*

$$T := \frac{c^T\widehat{\beta} - a}{\widehat{\sigma}\sqrt{c^T(X^TX)^{-1}c}}$$

and $t_{n-p,1-\alpha/2}$ is the quantile of order $1-\alpha/2$ of the Student law of $n-p$ degrees of freedom. Then, $\phi(Y) = 1_{\{|T| > t_{n-p,1-\alpha/2}\}}$ is a test of size α of

$$H_0 :\ c^T\beta = a \quad vs \quad H_1 :\ c^T\beta \neq a.$$

Proof : Immediate. □

The previous test can be generalized as in the following theorem.

Theorem 6 *Let V and W be two vector subspaces where W is a vector subspace of V. Assume $q = dim(W) < p = dim(V)$. We set*

$$F := \frac{\|P_W Y - P_V Y\|^2/(p-q)}{\|Y - P_V Y\|^2/(n-p)}$$

Where $P_V Y$ is the orthogonal projection of Y into V and $P_W Y$ is the orthogonal projection of Y into W. Then,

$$\phi(Y) = 1_{\{F > f_{p-q,n-p,1-\alpha}\}},$$

where $f_{p-q,n-p,1-\alpha}$ is the quantile of order $1-\alpha$ of the Fisher law at $(p-q, n-p)$degrees of freedom, is a test of sizeα of

$$H_0 : X\beta \in W \subset V \quad vs \quad H_1 : X\beta \in V \setminus W.$$

Proof : Immediate. □

Comments:

- The idea of the previous theorem is to select the smallest model (the smallest number of needed predictors): Let denote by $\widehat{Y}_0$ the projection of Y under H_0 (the smaller than the model under H_1) and $\widehat{Y}_1$ the projection of Y under H_1. If $\widehat{Y}_0$ is "closed to" we keep $\widehat{Y}_0$ ("sparse selection"). "Close" in the sense of the euclidian distance $\|\widehat{Y}_0 - \widehat{Y}_1\|^2$ standardized by the estimation error $\|Y - \widehat{Y}_1\|^2$.
- The previous test is justified because if H_0 is true then the numerator is expected to be small.
- We call the **Global Fisher test** the following test

Global Fisher test $H_0 : Y = \beta_0 \mathbb{1}_n + \varepsilon$ vs $H_1 : Y = \beta_0 \mathbb{1}_n + \sum_{j=1}^{p} \beta_j X_j + \varepsilon.$

  We test under H_0 the model reduced to the intercept aigainst the full model under H_1.

Exercise 4 (Test between nested models) *Let $1 \leq q < p$. How to test if regressors $X_{q+1}, \ldots, X_p$ are unrelevant ?*

Exercise 5 (Global Fisher test) *Let $p \geq 2$. Assume $X_1 = (1, \ldots, 1)^T$. We want to test if at least one predictor variable comes into play in the model. Prove that the Fisher test answers the problem and that in this case the variable F is written:*

$$F = \frac{\|\widehat{Y} - \overline{Y}\mathbb{1}_n\|^2/(p-1)}{\widehat{\sigma}^2} = \frac{\|\widehat{Y} - \overline{Y}\mathbb{1}_n\|^2/(p-1)}{\|Y - \widehat{Y}\|^2/(n-p)}$$

with $\widehat{Y} = P_X Y$. What is the law of F under H_0 ?

2.7.4. Confidence interval and bootstrap

The aim of this section is to present the regression bootstrap method in order to obtain a confidence interval for β without additional assumptions about the distribution of ε. The algorithm is as follows:

1. The method is built from the formulas seen previously. We estimate (β, ε) by $(\widehat{\beta}, \widehat{\varepsilon})$. We note $\widehat{F}_n$ the empirical distribution of $\widehat{\varepsilon}_i$:

$$\widehat{F}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\widehat{\varepsilon}_i}.$$

2. We sample n estimated residuals (say $\widehat{\varepsilon}_i^*$) with replacement from estimated residuals $\widehat{\varepsilon}_i$.

3. Based on these $\widehat{\varepsilon}_i^*$, we construct

$$Y^* = X\widehat{\beta} + \widehat{\varepsilon}^* \in \mathbb{R}^n$$

4. We estimate β from Y^*

$$\widehat{\beta}^* = (X^T X)^{-1} X^T Y^*.$$

The bootsrap theory gives

$$\sqrt{n}(\widehat{\beta}^* - \widehat{\beta}) \overset{loi}{\approx} \sqrt{n}(\widehat{\beta} - \beta).$$

To estimate the distribution of $\sqrt{n}(\widehat{\beta}^* - \widehat{\beta})$, we repeat M times the previous operation. For $k = 1, \ldots, M$, we have at our disposal a vector $(\widehat{\beta}^*(k), \widehat{\varepsilon}^*(k))$. We obtain the distribution of $\widehat{\beta}_j$ using the histogram of the $(\widehat{\beta}_j^*(k))_{k=1,\ldots,M}$. We deduce the approximate quantiles.

2.8. Prediction

Consider a new individual $n + 1$ such that

$$x_{n+1} = (x_{n+1,1}, \ldots, x_{n+1,p})^T$$

are the values of the regressors associated with this individual. We want to predict the value of Y_{n+1}. We set:

$$Y_{n+1} = x_{n+1}^T \beta + \varepsilon_{n+1}$$

where ε_{n+1} is a random variable independent of all the $(\varepsilon_i)_{1 \leq i \leq n}$. Moreover ε_{n+1} has the same distribution as each ε_i. We estimate β by $\widehat{\beta}$ which is a function of $Y_1, \ldots, Y_n$. Then we estimate Y_{n+1} by

$$\widehat{Y}_{n+1} = x_{n+1}^T \widehat{\beta}.$$

We get :

$$\mathbb{E}_\beta[\widehat{Y}_{n+1}] = x_{n+1}^T \beta, \quad \mathbb{V}\mathrm{ar}_\beta(\widehat{Y}_{n+1}) = \sigma^2 (x_{n+1}^T (X^T X)^{-1} x_{n+1}).$$

Theorem 7 *In the model* (1)*, under the Rank assumption and under **[P1]–[P4]**. Assume* $\varepsilon = (\varepsilon_i)_{1\leq i\leq n} \sim \mathcal{N}(0, \sigma^2 I)$*. Un intervalle de confiance pour* Y_{n+1} *est donné par :*

$$\left[x_{n+1}^T\widehat{\beta} - t_{n-p,1-\alpha/2}\widehat{\sigma}\sqrt{1 + x_{n+1}^T(X^TX)^{-1}x_{n+1}}\ ,\ x_{n+1}^T\widehat{\beta} + t_{n-p,1-\alpha/2}\widehat{\sigma}\sqrt{1 + x_{n+1}^T(X^TX)^{-1}x_{n+1}}\right],$$

où $t_{n-p,1-\alpha/2}$ *est le quantile d'ordre* $1-\alpha/2$ *de la loi de Student à* $n-p$ *degrés de liberté.*

Proof : We show that

$$Y_{n+1} - \widehat{Y}_{n+1} \sim \mathcal{N}\left(0, \sigma^2(1 + x_{n+1}^T(X^TX)^{-1}x_{n+1})\right). \quad \square$$

2.9. R^2-coefficient

In this section, we consider the model~(1), under the Rank assumption and under **[P1]–[P3]**. We also assume an intercept, *i.e* the first column of the matrix X is the one vector $X_1 := \mathbb{1}_n := (1, 1, \ldots, 1)^T \in \mathbb{R}^n$.

Comment:

- ☛ First note that the projection of $\left(Y - \overline{Y}\mathbb{1}_n\right)$ into $[X] = \mathcal{I}m(X)$ is such that

$$P_X\left(Y - \overline{Y}\mathbb{1}_n\right) = P_XY - P_X\overline{Y}\mathbb{1}_n = \widehat{Y} - \overline{Y}\mathbb{1}_n.$$

- ☛ Then as $\left(Y - \widehat{Y}\right) \in [X]^\perp$ and $\left(\widehat{Y} - \overline{Y}\mathbb{1}_n\right) \in [X]$ are orhtogonal vectors, we get by Pythagoras formula

$$\|Y - \overline{Y}\mathbb{1}_n\|^2 = \|Y - \widehat{Y} + \widehat{Y} - \overline{Y}\mathbb{1}_n\|^2 = \|Y - \widehat{Y}\|^2 + \|\widehat{Y} - \overline{Y}\mathbb{1}_n\|^2 \tag{2}$$

In this section $\overline{y} = n^{-1}\sum_{i=1}^n y_i$ and $\widehat{y} = X\widehat{\beta}$, where the y_i are the observations.

Definition 4 *We denote by* TSS *(*SCT *in french) the total sum of squares*

$$TSS := \|y - \overline{y}\mathbb{1}_n\|^2 = \sum_{i=1}^n (y_i - \overline{y})^2.$$

Definition 5 *We denote by* RSS *(*SCR *in french) the residual sum of squares*

$$RSS := \|y - \widehat{y}\|^2 = \sum_{i=1}^n (y_i - \widehat{y}_i)^2.$$

Definition 6 *We denote by MSS ($SCEM$ or SCM or SCE in french) model sum of squares by the model*

$$MSS := \|\widehat{y} - \overline{y}\mathbb{1}_n\|^2 = \sum_{i=1}^{n} (\widehat{y_i} - \overline{y})^2.$$

Comments:

- The MSS corresponds to the variability (variance) explained by the model.
- The RSS corresponds to the variability (variance) not explained by the model.
- The TSS corresponds to the total variability (variance).

Definition 7 *From* (2)*, we have the following relation*

$$TSS = MSS + RSS$$

Comment:

- If the model is "good" enough, the part of total variability must be explained by the model and inversely the part of variability not explained by the model shall be small. It is the role of the determination coefficient to quantify this notion.

Definition 8 *We define the* ***determination coefficient*** R^2 *as follows*

$$R^2 = \frac{MSS}{TSS} = \frac{\|\widehat{y} - \overline{y}\mathbb{1}_n\|^2}{\|y - \overline{y}\mathbb{1}_n\|^2}.$$

We can rewrite the R^2-coefficient

$$R^2 = 1 - \frac{RSS}{TSS}.$$

Comments:

- First note that $R^2 \in [0, 1]$.
- Note that in the simplest model reduced to the intercept $Y = \beta\mathbf{1} + \varepsilon$, The least square estimator of the real parameter β is $\widehat{\beta} = \overline{Y}$.
- The linear model is interesting only if the errors "$y_i - \widehat{y_i}$" are small relative to the errors "$y_i - \overline{y}$" that we would make if we took a model without regressors, i.e. a model reduced to the intercept. The linear model has an interest if

$$\frac{RSS}{TSS} = \frac{\|y - \widehat{y}\|^2}{\|y - \overline{y}\mathbb{1}_n\|^2} \to 0.$$

This is equivalent to

$$R^2 \to 1.$$

- ☛ Note that $R^2 \in [0,1]$. R^2-coefficient close to 1 "means" the predictors $X_2, \ldots, X_p$ "well" explain the model.
- ☛ Note that for the global Fisher test,

$$F = \frac{\|P_X Y - \overline{Y}\mathbb{1}_n\|^2/(p-1)}{\widehat{\sigma}^2} = \frac{(n-p)R^2}{(p-1)(1-R^2)}$$

  Thus, H_0 is rejected for large values of R^2.
- ☛ The R^2-coefficient is not a good criterion as it depends on the dimension p. Indeed R^2 increase with p or equivalentely RSS decrease with p. Take two models, the first one with p predictors and the second one with same predictors plus one, so a model of size $p+1$. Denote by RSS_p and RSS_{p+1} the RSS in the first and the second models, it comes as $\widehat{Y} = X\widehat{\beta}$

$$RSS(p+1) = \|Y - X\widehat{\beta}\|^2 = \min_{\beta \in \mathbb{R}^{p+1}} \|Y - X\beta\|^2 = \min_{\beta \in \mathbb{R}^{p+1}} \|Y - \sum_{j=1}^{p+1} X_j \beta_j\|^2 \leq \min_{\beta \in \mathbb{R}^p} \|Y - \sum_{j=1}^{p} X_j \beta_j\|^2 = RSS(p)$$

  So the R^2 coefficient is increasing with p. If we want to get rid of the dimension p, we use the adjusted R^2 define bellow.

Definition 9 *We define the **adjusted determination coefficient** R_a^2 as follows*

$$R_a^2 = 1 - \frac{(n-1)\|Y - \widehat{Y}\|^2}{(n-p)\|Y - \overline{Y}\mathbb{1}_n\|^2} = 1 - \frac{(n-1)\|\widehat{\varepsilon}\|^2}{(n-p)\|Y - \overline{Y}\mathbb{1}_n\|^2}.$$

Introduction to Regression - Chapter 3

MAP 535

Contents

Chapter 3 : Atypical points and model validation

For some atypical observations, the values of the response variable Y and/or predictors X_j appear to behave differently from the majority of observations, these points are called ***outliers***. Moreover, observations that do not follow the same linear regression model that most data are called ***regression outliers***.

Underline that it is important to first check the most obvious reasons of such of points : measurement errors, data transcription errors, and so on. For example : a boat passengers is written to be 400 years, a 1 year old baby running the 100m in 10 seconds...

Then, the remaining *outliers* are not necessarily wrong. Indeed, *outliers* sometimes reveal a particular phenomenon that may be different from the model followed by the majority of observations. Keep in mind that even if the aim of a model is to explain as well as possible a general phenomenon, it can have its own limits. Thus, *outliers* can suggest to track for more elaborate models (missing regressor,...).

In the regression setting, an typical values (*outliers*) can occur in three main ways :

- in the response Y but not in the predictors X_j,
- in the predictors X_j but not in the response variable Y,
- in both Y and X directions.

3.1. Atypical points

Let's place ourselves in the context of the Iinear regression and consider an *outlier* in the Y-direction (an atypical value in a response Y_i) but not in the predictors X_{ij}. We detect them easily by an univariate detection. Let us consider a toy example. A simple boxplot reveal the *outliers* (see figure 1, right).

According to the scatter plot (figure 1, left), a linear model can be considered. We set

$$Y_i = ax_i + \varepsilon_i, \qquad \varepsilon_i \text{ i.i.d. } \mathcal{N}(0, \sigma^2).$$

By simple calculation, the ordinary least square estimator (the same as the maximum likelihood estimator in our setting) is such that

$$\widehat{a} = \frac{s_{x,y}}{s_x^2} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - (\overline{x})^2} \quad \text{and} \quad \widehat{b} = \overline{y} - \widehat{a}\,\overline{x} \quad \Rightarrow \quad y = \widehat{a}x + \widehat{b}.$$

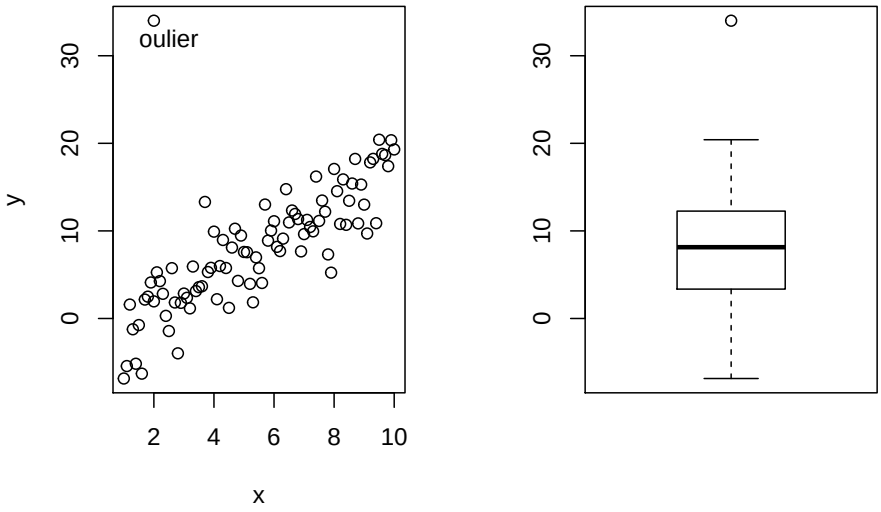


Figure 1: Scatter plot of the toy dataset/Boxplot of the toy dataset

Plot now (figure 2, right) the two least square lines with and without the outlier.

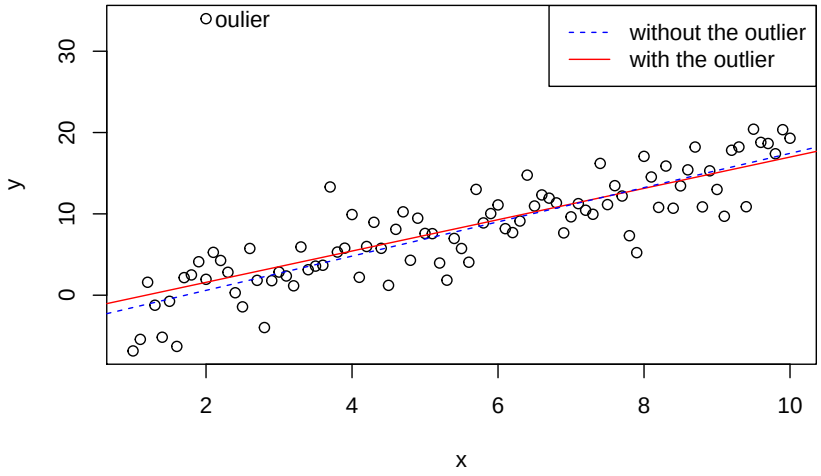


Figure 2: Scatter plot of the toy dataset/The least square lines with and without the outlier

Comments:

- Here, the outlier has only a small effect on the estimation. Indeed, removing this point slightly changes the regression line (least squares line).
- This type of atypical observations (*outliers*) has an impact on the estimation of σ^2 so on the residuals $\widehat{\varepsilon} = Y - \widehat{Y}$.
- **The *regression outliers* can be detected by a residuals analysis.**

3.2. Isolated observations

An *isolated observation* has atypical values in the predictors X_{ij}. It means that the values $\left(X_{ij}\right)_j$ of the observation i are relatively far from all the value $\left(X_{i'j}\right)_j$ of the other observations $i' \neq i$. Let us consider an other toy example. According to the scatter plot (figure~**??**), a linear model can be considered. We set

$$Y_i = ax_i + \varepsilon_i, \qquad \varepsilon_i \text{ i.i.d. } \mathcal{N}(0, \sigma^2).$$

By simple calculation, the ordinary least square estimator (the same as the maximum likelihood estimator in our setting) is such that

$$\widehat{a} = \frac{s_{x,y}}{s_x^2} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - (\overline{x})^2} \quad \text{and} \quad \widehat{b} = \overline{y} - \widehat{a}\,\overline{x} \quad \Rightarrow \quad y = \widehat{a}x + \widehat{b}.$$

Plot now (figure~3) the two least square lines with and without the outlier.

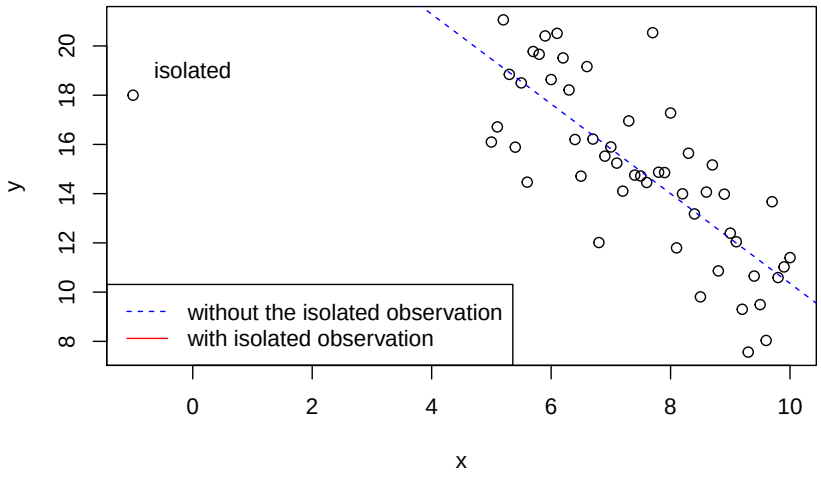


Figure 3: Scatter plot of the toy dataset/The least square lines with and without the isolated observation

Comments:

- Here, the *isolated observation* has a real impact on the estimation. Indeed, removing this point significantly changes the regression line (least squares line).
- This type of atypical observations (*isolated observations*) has an impact on the estimation of β. Here, the*isolated observation* influe on the estimation of β.
- Such of points which influe the estimation of β, are called *leverage point*.
- ***Leverage points*** **can be detected by a multivariate detection study of the "leverage effect".**
- Note that in this example the response Y_i of the *isolated observation* is quite far from the regression line. It does not follow the general linear trend of the majority of observations.

3.3. Leverage effect

Atypical points can be **leverage points** (and/or **regression outliers**). An analysis of the influence (leverage effect) of an observation is based on the idea of comparing the adjustment with and without this observation. Note that it should to be done for each of the observations in the dataset. To this end, let us introduce the calculation of the estimator of β without the observation $i : (x_i, y_i) \in \mathbb{R}^P \times \mathbb{R}$. The index "$(-i)$" means "without the observation i". For example, the matrix $X_{(-i)}$ is the $(n-1) \times p$ matrix corresponding to the matrix X without the *i-th* line. Therefore, calculating the least squares estimator without the observation (x_i^T, Y_i) gives:

$$\widehat{\beta}_{(-i)} = \left(X_{(-i)}^T X_{(-i)}\right)^{-1} X_{(-i)}^T Y_{(-i)}.$$

• Then, the predictive Y_i for the observation x_i in this setting is noted $\widehat{Y}_i^P = x_i^T \widehat{\beta}_{(-i)}$.

• The associated prediction error is $Y_i - \widehat{Y}_i^P$.

Intuitively, if the i observation is not too influential, the estimation error $(Y_i - \widehat{Y}_i)$ and prediction error $(Y_i - \widehat{Y}_i^P)$ will be relatively close. If not, the i observation deserves special attention. These two quantities are related to the projector P_X

$$P_X = X(X^T X)^{-1} X^T.$$

Recall that $\widehat{Y} = P_X Y$, then we deduce :

$$\widehat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j = h_{ii} Y_i + \sum_{j \neq i} h_{ij} Y_j.$$

Proposition 1 *Note $h_{ij} = (P_X)_{ij}$, the entries of P_X. The trace of P_X is equal to :*

$$Tr(P_X) = \sum_{i=1}^{n} h_{ii} = p.$$

Moreover, for all $i = i, \cdots, n$ and for all $j \neq i$,

1. $0 \leq h_{ii} \leq 1, \quad -\frac{1}{2} \leq h_{ij} \leq \frac{1}{2}$.
2. *If $h_{ii} = 1$ then $h_{ij} = 0$.*

Theorem 1 *In the linear regression model, under the Rank assumption and under **[P1]–[P4]**, we have for all* $i = 1, \cdots, n$

$$Y_i - \widehat{Y}_i = (1 - h_{ii)}(Y_i - \widehat{Y}_i^P),$$

where h_{ii} *denote the i-th diagonal element of* P_X.

Comments:

- From the proposition 1, it follows the fact that $\widehat{Y}_i$ is entirely determined by Y_i as soon as $h_{ii} = 1$. If $h_{ii} = 0$, Y_i has no influence on $\widehat{Y}_i$.
- The theorem 1 suggests that the estimation error $(Y_i - \widehat{Y}_i)$ and prediction error $(Y_i - \widehat{Y}_i^P)$ are equal for $h_{ii} = 0$.

Definition 1 *An observation i is called a **leverage point** if* $h_{ii} > s$, *where*

- $s = 2p/n$ *according to Hoaglin & Welsch (1978),*
- $s = 3p/n$ *for* $p > 6$ *and* $(n - p) > 12$ *according to Velleman & Welsch (1981),*
- $s = 1/2$ *according to Huber& Welsch (1981).*

Comment:

- It is possible to prove that h_{ii} corresponds in a certain way to the distance of the point x_i to the gravity center $\bar{x}$ of the scatter plot x_i. **In other words, the h_{ii} tells us precisely which are the isolated observations of the sample**.
- If an observation is such that $h_{ii} > s$, influences its own estimate. But it does not necessarily affect the overall model, that is, the estimate of β.
- Without being necessarily *regression outlier* (residuals analysis), leverage points are atypical points in explanatory variables. Without systematically eliminating them, it is important to detect and analyze them: do they come from measurement errors or from a population of a different nature? Do they impact the estimation of β (cook distance) ?

3.4. Residuals analysis

3.4.1. The different residuals

Recall first that , the $\varepsilon = Y - X\beta$ is the vector of the theoritical errors/residuals such that $\mathbb{E}_\beta[\varepsilon] = 0_n$ and $\mathbb{V}\text{ar}_\beta[\varepsilon] = \sigma^2\mathbb{I}_n$. We give in definition~**??** a first estimation of the ε : the estimated residuals

$$\widehat{\varepsilon} = Y - \widehat{Y} = Y - X\widehat{\beta} = Y - P_X Y = (I - P_X)Y = P_{X^\perp}Y = P_{X^\perp}\varepsilon.$$

Moreover, according to proposition~**??**

$$\mathbb{E}_\beta[\widehat{\varepsilon}] = 0_n \text{ and } \mathbb{V}\text{ar}_\beta[\widehat{\varepsilon}] = \sigma^2 P_{X^\perp}.$$

The postulat **Postulat [P2]** is not satisfied by the estimated residuals. To fix it, we consider the standardized residuals $t = (t_1, \cdots, t_n)^T$ such that

$$t_i = \frac{\widehat{\varepsilon_i}}{\widehat{\sigma}\sqrt{1-h_{ii}}}.$$

But the standardized residuals do not satisfy the postulate **Postulat [P2]** on uncorrelation. Introduce then, the studentized residuals $t^* = (t_1^*, \cdots, t_n^*)^T$ such that

$$t_i^* = \frac{\widehat{\varepsilon_i}}{\widehat{\sigma}_{(-i)}\sqrt{1-h_{ii}}},$$

where $\widehat{\sigma}^2_{(-i)}$ is the estimation of σ^2 in the model deprived of the observation i (by *cross validation*):

$$\widehat{\sigma}^2_{(-i)} = \frac{\|Y_{(-i)} - X_{(-i)}\widehat{\beta}_{(-i)}\|^2}{n-1-p}$$

Theorem 2 *In the regression linear model, under **[P1]–[P4]**, if rank$(X_{(-i)} = p$ then the **studentized residuals** are such that*

$$t_i^* = \frac{\widehat{\varepsilon_i}}{\widehat{\sigma}_{(-i)}\sqrt{1-h_{ii}}} \sim t_{n-1-p},$$

where t_{n-1-p} denotes the student law of $(n-1-p)$ degrees of freedom.

Proof : The demonstration is left as exercise. □

3.4.2. Residuals analysis for outliers detection

To analyze the fit quality of an observation, that is, if the model explains the observation, we look at the associated residual $\widehat{\varepsilon_i} = Y_i - \widehat{Y_i}$. If its standardized residual (or studentized residual) is large, then the observation is a *regression outlier*.

Definition 2 *A **regression outlier** is an observation (x_i^T, Y_i) such that the associated studentized residual t_i^* is high :*

$$|t_i^*| > t_{n-p-1,1-\alpha/2}.$$

Comments:

- Note that in theory, $\alpha\%$ of the datas are outliers.
- In practice, we use $\alpha = 5\%$, then for a large enough sample (larger than $30+p$), $t_{n-p-1,1-\alpha/2} \approx 2$.
- We are actually looking for (x_i^T, Y_i) for which t_i^* is well outside the confidence band in the $i \longmapsto t_i^*$ plot. In figure 4, only the point "52" is an outlier.
- Explaining the presence of these outliers can be difficult. They can be caused by measurement errors or be the result of a population change. It is recommended to pay attention to these points and check if they do not have too much influence on the calculation of $\widehat{\beta}$ and $\widehat{\sigma}^2$.

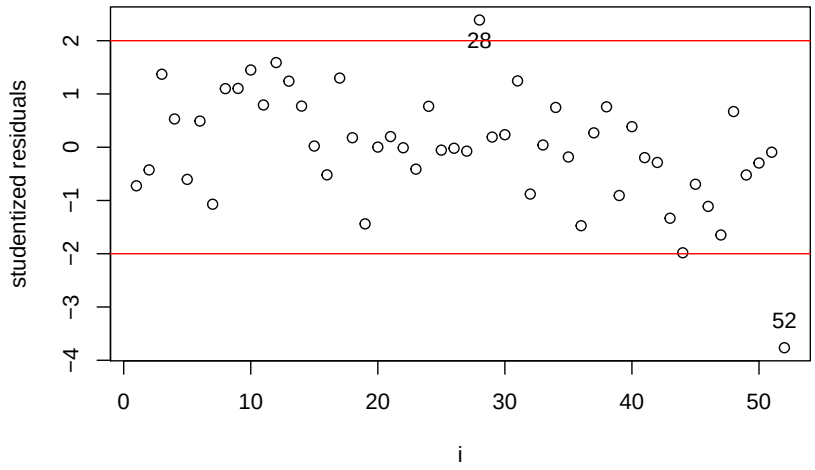


Figure 4: Scatter plot/Studentized residuals plot

3.4.2. Validation of the postulates

We recall that we assume in the regression linear model, the Rank assumption (easy to check) and the postulates **[P1]**–**[P4]** with

- **[P1]**: Errors are centered : $\forall i = 1, \cdots, n \quad \mathbb{E}_\beta[\varepsilon_i] = 0$. In practice, this means that the model is correct (the model is linear).
- **[P2]**: Errors have homosedasctic variance : $\forall i = 1, \cdots, n \quad \mathbb{V}\text{ar}_\beta[\varepsilon_i] = \sigma^2 > 0$.
- **[P3]**: Errors are uncorrelated: $\forall i \neq j \quad \mathbb{C}\text{ov}(\varepsilon_i, \varepsilon_j) = 0$.
- **[P4]**: Errors are gaussian : $\forall i = 1, \cdots, n \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

The simplest way to validate postulates is graphically.

Validation of the postulate [P1]: Errors are centered

The linearity assumption (the centered postulat) can be checked by inspecting the *Residuals vs Fitted*-plot (or $(\widehat{Y}_i, t_i^*)$ plot). Ideally, figure~5 shows no fitted pattern. (figure~**??**) That is, the red line should be approximately horizontal at zero. The presence of a pattern may indicate a problem with some aspect of the linear model.

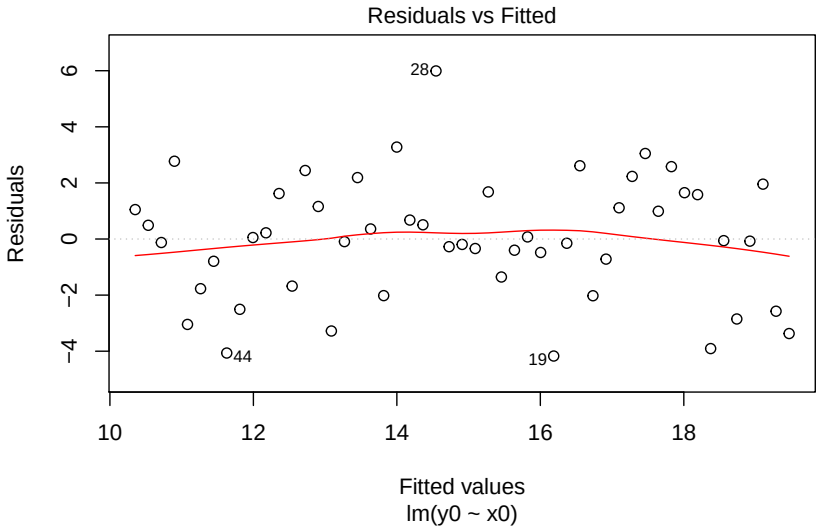


Figure 5: Residuals vs Fitted plot

Validation of the postulate [P2]: Errors have homosedasctic variance

This assumption can be checked by examining the *Residuals vs Fitted*-plot and the *Scale-location*-plot (plot of the points $(\widehat{Y}_i,\ \sqrt{t_i})$), also known as the *spread-location* plot. This last plot shows if residuals are spread equally along the ranges of predictors. It's good if you see a horizontal line with equally spread points. In our example (figure~6), this is the case.

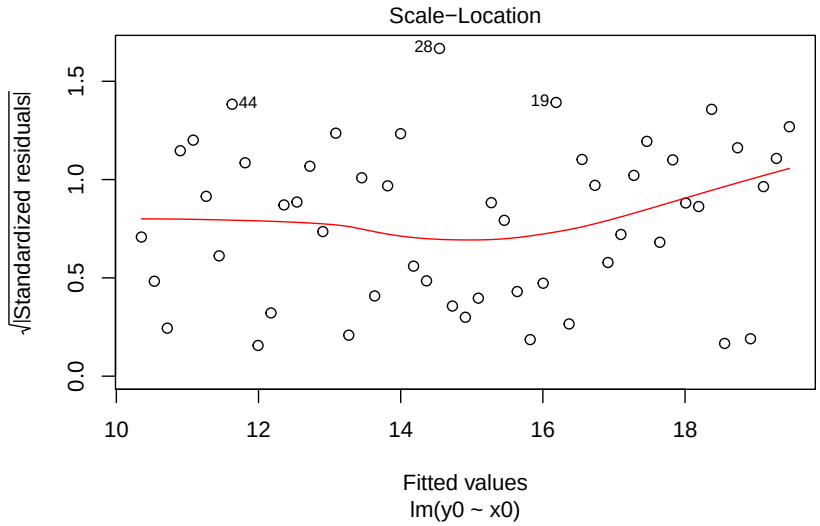


Figure 6: Scale-location plot

Comment:

- ☛ If there is a doubt of heteroscedasticity, we advise to make a test. A possible solution to reduce the heteroscedasticity problem is to use a log or square root transformation of the outcome variable Y.

Validation of the postulate [P3]: Errors are uncorrelated

Under **R**, we can represent the auto-correlation of the residuals using the command `acf()`. The vertical lines (igure~7) represent the correlation coefficients between the residues of each point and those of the points of the following line (lag = 1), or those separated by two lines (lag = 2) and so on. Its interpretation is simple. If a bar, exept the first one, exceeds dashed thresholds, uncorrelation isn't satified. In figure~7, the postulate is validated. We will see in the next chapter how to use it.

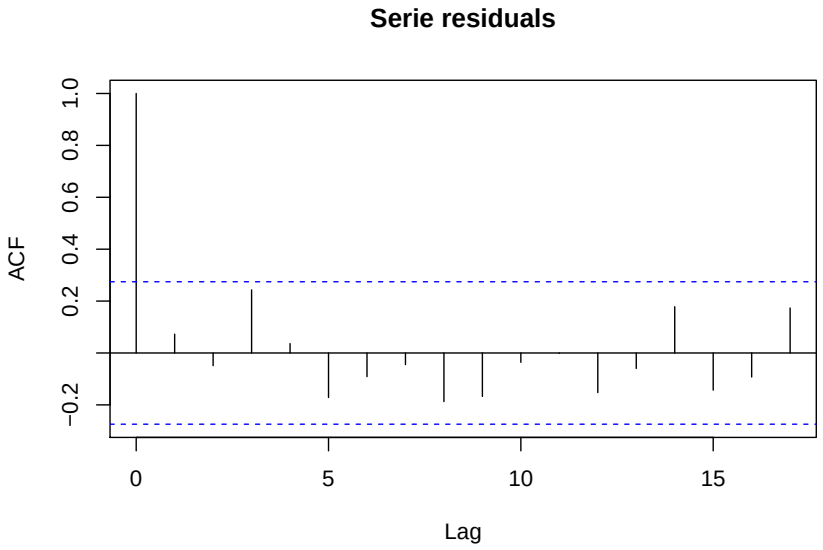


Figure 7: Autocorrelation-plot

Validation of the postulate [P4]: Errors are gaussian

To analyze the normality, we use the Q-Q plot. It consists in comparing the t_i to the theoritical quantiles of the reduced normal centered law (for n large enough, the standard normal is similar to the student law). If all the points fall approximately along this reference line, then the postulate is validated as in figure~8.

Comments:

- ☛ In general, it is often recognized that the normality assumption plays a minor role in regression analysis.
- ☛ The normality assumption is useful for inference purposes, especially for small samples. However, it should be noted that in the presence of small samples, non-normality may be particularly difficult to diagnose by residue examination.

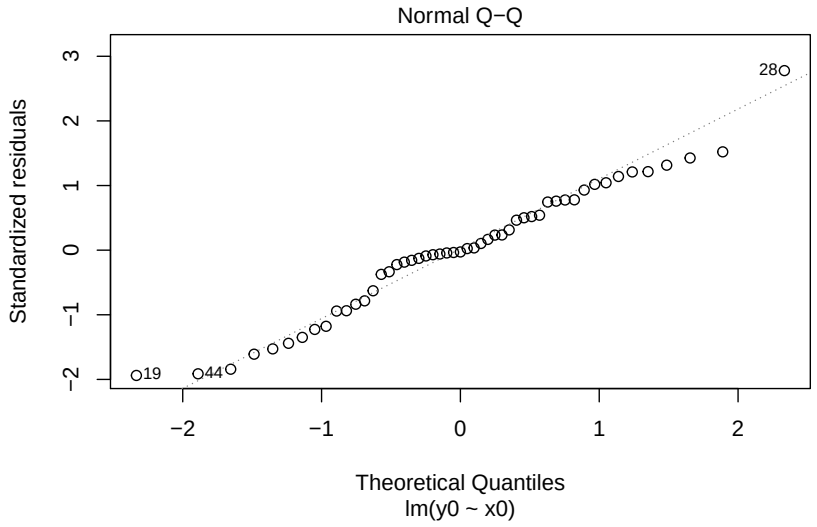


Figure 8: QQ-normal-plot

3.5. Cook's distance

Residuals analysis allow to identify atypical values related to the explained variable ; the analysis of the orthogonal projector allows to detect atypical values related to predictors. In this section, we try to combine these two analyzes. For that, we introduce Cook's distance.

3.5.1 Definition

Definition 3 *For all i, the Cook's distance of the observation* (x_i^T, Y_i) *is given by the following formula :*

$$D_i = \frac{1}{p\widehat{\sigma^2}}(\widehat{\beta}_{(-i)} - \widehat{\beta})^T (X^T X)(\widehat{\beta}_{(-i)} - \widehat{\beta})$$

where $\widehat{\beta}_{(-i)}$ *is the estimation of* β *in the model without the i-th observation.*

Comments:

- ☛ The Cook distance is essentially a standardized distance measure that describes the change in the β estimator when we remove the observation i.
- ☛ A high value of Cook's distance suggests that observation i has a high influence. In practice, Cook's distances are often compared with 1. A value much lower than 1 suggests that the impact of observation i does not seem very important. In contrast, a Cook distance greater than one suggests that observation i has a large impact.
- ☛

Proposition 2 *The Cook's distance of the observation* (x_i^T, Y_i) *satisfies*

$$D_i = \frac{h_{ii}}{p\widehat{\sigma^2}(1-h_{ii})^2}(Y_i - \widehat{Y_i})^2 = \frac{h_{ii}}{p(1-h_{ii})}t_i^2$$

where h_{ii} *is the i-th diagonal element of the orthogonal projector* P_X *and* t_i *is the standardized residual associated to the observation i.*

Proof : Let as exercice.

Comments:

- Recall that the standardized residuals measures the adequacy of the observation Y_i to the estimated model $\widehat{Y_i}$ while the quantity $\frac{h_{ii}}{1-h_{ii}}$ measure the sensitivity of the estimator $\widehat{\beta}$ to the observation i. Indeed, they are such that

$$\frac{h_{ii}}{1-h_{ii}} = \frac{\mathbb{V}\mathrm{ar}_\beta(Y_i)}{\mathbb{V}\mathrm{ar}_\beta(\widehat{\varepsilon_i})}.$$

  The $\frac{h_{ii}}{1-h_{ii}}$'s are called the **levers**.

- It can be seen that the Cook's distance for fixed p, can be large if the standardized residues are large or if the levers are large (or if both are large).

- Thus Cook's distance can be seen as a criterion measuring both the outlier (aberrant) character of an observation (measured by the standardized residual) and its leverage effect. Points with high Cook's distances (greater than 1) will be outliers, or levers, or both. It is strongly recommended to delete points with a high Cook distance. Nevertheless, if we want to keep these points, we have to make sure that they do not change too much the estimation of β and the interpretations.

3.5.2 Examples

Example 1

Consider here the toy example (figure~9 top/left) where the 52-*th* point $(-1, 18)$ is an *isolated point*. If we look at the *Studentized residals*-plot (figure~9 top/right), it comes out that the 52-*th* point is a *regression outlier* as $t^*_{52} > 2$. Moreover, the 28-*th* observation is also a *regression outlier* as $t^*_{28} > 2$. Are they *leverage points* ?

In the h_{ii}-plot (figure~9 bottom/left), we see that all the $h_{ii} < 0.5$, so none point is influent on its own estimation. Nevertheless, according to the *Residuals vs leverage*-plot (figure~9 bottom/right), it turns out that the 52-*th* point has a high Cook's distance (larger than 1). It has a large impact on the estimation of β, this point may be removed.

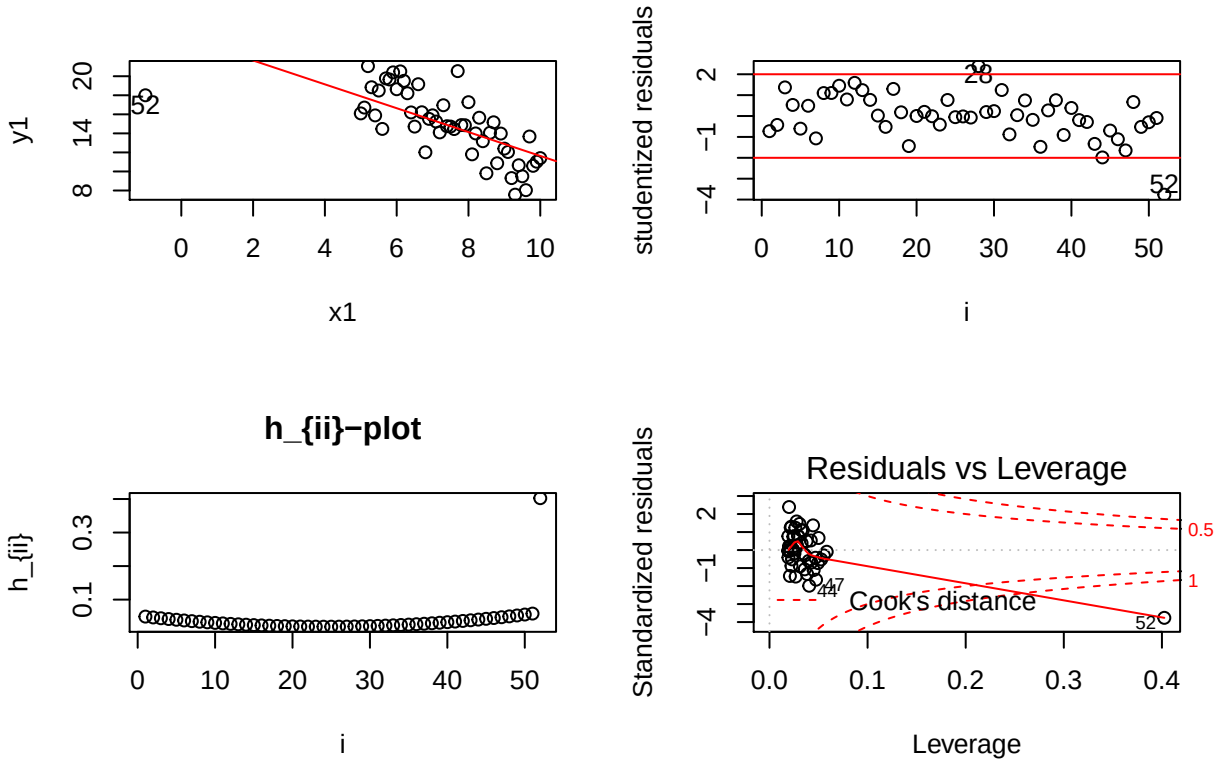


Figure 9: Some Plots for the toy dataset of example 1

Example 2

Consider now, the toy example (figure~10 top/left) where the 52-*th* point $(-10, 50)$ is an *isolated point* and an *outlier*. Note that the point follows the model as it is close to the least square line. The *Studentized residals*-plot (figure~10 top/right) indicates that this point is not a *regression outlier* as $t^*_{52} < 2$. On the other hand, it appears that the 28-*th* observation is an *regression outlier* as $t^*_{28} > 2$.

In the h_{ii}-plot (figure~10 bottom/left), the only *leverage point* is the $52 - th$ point as its $h_{ii} > 0.5$. Moreover, according to the *Residuals vs leverage*-plot (figure~10 bottom/right), it turns out that the 52-*th* point has a Cook's distance larger than 1. It has a large impact on the estimation of β, this point is a *leverage point and a regression outlier*, it may be removed.

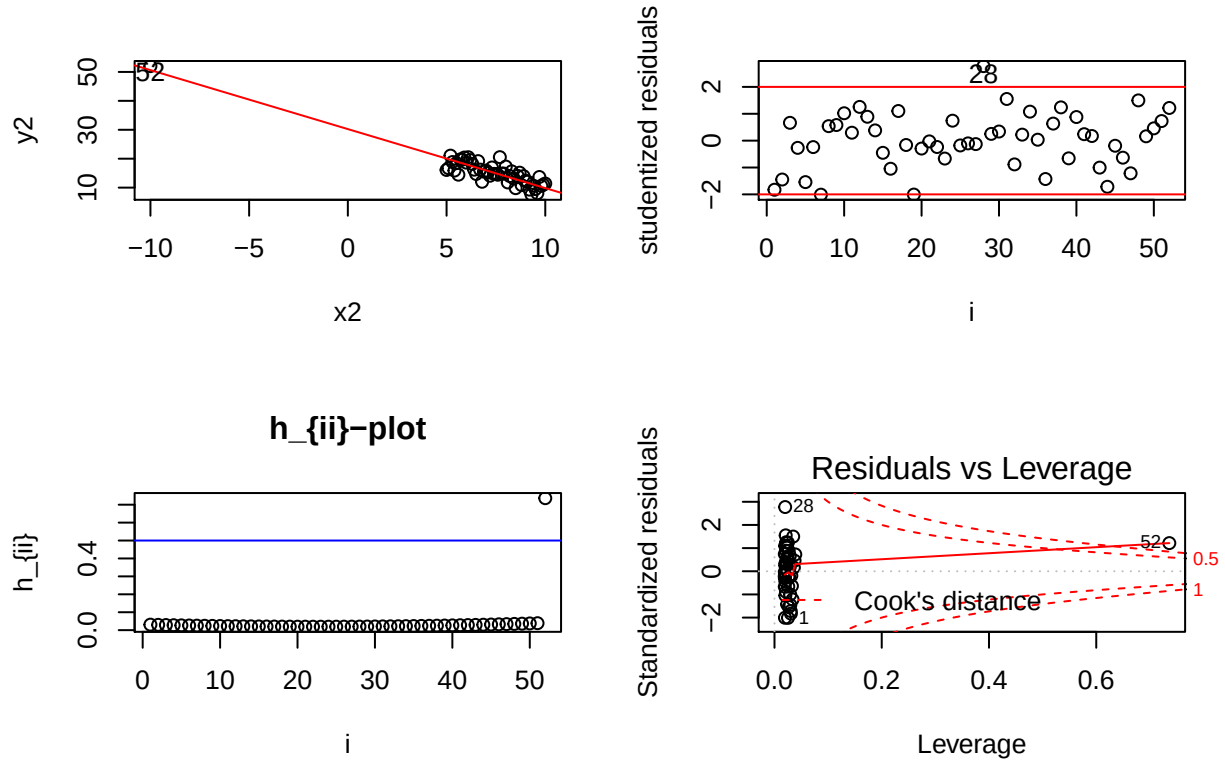


Figure 10: Some Plots for the toy dataset of example 2

Example 3

Consider here the toy example (figure~11 top/left) where the $52-th$ point $(7, 40)$ is an *outlier*.The *Studentized residals*-plot (figure~**??** top/right) indicates that this point is a *regression outlier* as $t^*_{52} > 2$.

In the h_{ii}-plot (figure~11 bottom/left), so none point is influent on its own estimation as for each observation $h_{ii} < 0.5$. Moreover, according to the *Residuals vs leverage*-plot (figure~11 bottom/right), it turns out that the $52-th$ point has a Cook's distance smaller than 1. It has not a big influence on the estimation of β, this point is a *regression outlier* but not a *leverage point*, it may be keeped.

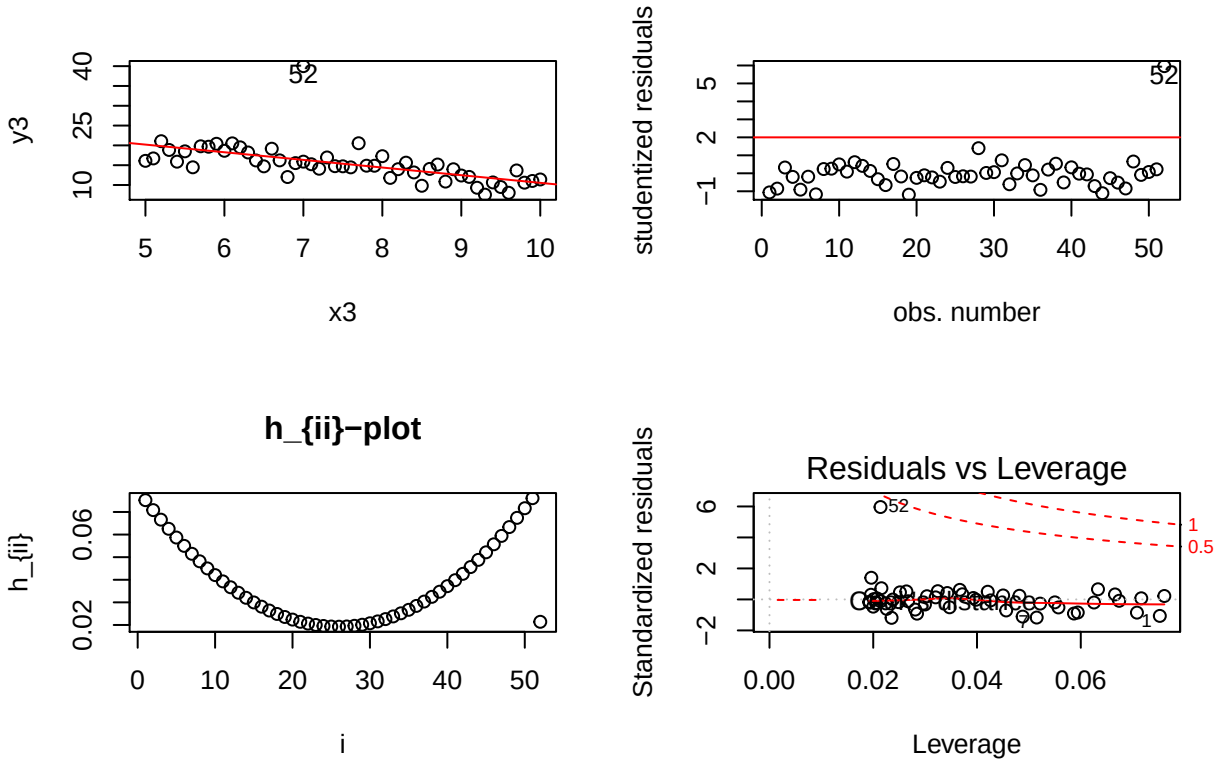


Figure 11: Some Plots for the toy dataset of example 3

3.6. Conclusion

The statistics can not in any case determine which variables have been forgotten,which can be one of the reasons why the postulate on homoscedsticity is not verified. On the other hand, it can determine the function f_j such that Y and $ f_j (X_j) $ has a linear relation. Most of the time, we will leave a model that includes a maximum of variables that can explain Y, even if we remove one using the Fisher or Student tests. In the next Chapter, we will go further into the issue of choice of models. First, start with an example under **R** with a real dataset.

Introduction to Regression - Chapter 4

MAP 535

Contents

Chapter 4: Model estimation/validation through R example

In this Chapter, we will see through an example how to estimate and validate a model. Moreover, we see how to detect atypical points.

4.1. The data

To illustrate the method, we consider the example~**??**.

> We try to explain and predict gasoline consumption (in liters per 100 km) of different automobile models based on several variables. For this, we have the following characteristics for 31 different cars:
>
> - `Type` = Type of the vehicle.
> - `Consommation` = Fuel consumption in liters per 100 km.
> - `Prix` = Vehicle price in Swiss francs.
> - `Cylindree` = Cylinder capacity in cm3.
> - `Puissance` = Power in kW.
> - `Poids` = Weight in kg.
>
> In this example, Y is the variable `Consommation`. The variables X_j correspond to the other 4 variables.

We first download the dataset `conso.txt` with the command `read.table`.

```
conso_voit = read.table("conso.txt", header=TRUE, sep="\t", dec=",",row.names=1)
conso_voit_complet = read.table("conso.txt", header=TRUE, sep="\t", dec=",")
```

To print the names of the variables is possible with the command `names`.

```
names(conso_voit_complet)
```

```
## [1] "Type"         "Prix"         "Cylindree"    "Puissance"
## [5] "Poids"        "Consommation"
```

To display the data in a table, we use the `knitr` library and the function `kable`.

```
library(knitr)
```

```
kable(conso_voit_complet)
```

Type	Prix	Cylindree	Puissance	Poids	Consommation
Daihatsu Cuore	11600	846	32	650	5.7
Suzuki Swift 1.0 GLS	12490	993	39	790	5.8
Fiat Panda Mambo L	10450	899	29	730	6.1
VW Polo 1.4 60	17140	1390	44	955	6.5
Opel Corsa 1.2i Eco	14825	1195	33	895	6.8
Subaru Vivio 4WD	13730	658	32	740	6.8
Toyota Corolla	19490	1331	55	1010	7.1
Ferrari 456 GT	285000	5474	325	1690	21.3
Mercedes S 600	183900	5987	300	2250	18.7
Maserati Ghibli GT	92500	2789	209	1485	14.5
Opel Astra 1.6i 16V	25000	1597	74	1080	7.4
Peugeot 306 XS 108	22350	1761	74	1100	9.0
Renault Safrane 2.2. V	36600	2165	101	1500	11.7
Seat Ibiza 2.0 GTI	22500	1983	85	1075	9.5
VW Golt 2.0 GTI	31580	1984	85	1155	9.5
Citroen ZX Volcane	28750	1998	89	1140	8.8
Fiat Tempra 1.6 Liberty	22600	1580	65	1080	9.3
Fort Escort 1.4i PT	20300	1390	54	1110	8.6
Honda Civic Joker 1.4	19900	1396	66	1140	7.7
Volvo 850 2.5	39800	2435	106	1370	10.8
Ford Fiesta 1.2 Zetec	19740	1242	55	940	6.6
Hyundai Sonata 3000	38990	2972	107	1400	11.7
Lancia K 3.0 LS	50800	2958	150	1550	11.9
Mazda Hachtback V	36200	2497	122	1330	10.8
Mitsubishi Galant	31990	1998	66	1300	7.6
Opel Omega 2.5i V6	47700	2496	125	1670	11.3
Peugeot 806 2.0	36950	1998	89	1560	10.8
Nissan Primera 2.0	26950	1997	92	1240	9.2
Seat Alhambra 2.0	36400	1984	85	1635	11.6
Toyota Previa salon	50900	2438	97	1800	12.8
Volvo 960 Kombi aut	49300	2473	125	1570	12.7

The `dim` command displays the size of the data (number of lines, number of columns). Here $n = 31$ and we have 5 variables.

```
dim(conso_voit_complet)
```

```
## [1] 31  6
```

The command `head` allows you to view the first 6 lines of the data. The command `tail` allows you to view the last 6 lines of the data.

```
head(conso_voit_complet)
```

```
##                   Type  Prix Cylindree Puissance Poids Consommation
## 1       Daihatsu Cuore 11600       846        32   650           5.7
## 2 Suzuki Swift 1.0 GLS 12490       993        39   790           5.8
## 3   Fiat Panda Mambo L 10450       899        29   730           6.1
## 4       VW Polo 1.4 60 17140      1390        44   955           6.5
## 5  Opel Corsa 1.2i Eco 14825      1195        33   895           6.8
## 6     Subaru Vivio 4WD 13730       658        32   740           6.8
```

The command `str` allows us to chek if the nature of each variable is well determined by **R**. Here, there is no mistake.

```
str(conso_voit_complet)
```

```
## 'data.frame':    31 obs. of  6 variables:
##  $ Type        : Factor w/ 31 levels "Citroen ZX Volcane",..: 2 25 4 31 17 24 26 3
##  $ Prix        : int  11600 12490 10450 17140 14825 13730 19490 285000 183900 92500
##  $ Cylindree   : int  846 993 899 1390 1195 658 1331 5474 5987 2789 ...
##  $ Puissance   : int  32 39 29 44 33 32 55 325 300 209 ...
##  $ Poids       : int  650 790 730 955 895 740 1010 1690 2250 1485 ...
##  $ Consommation: num  5.7 5.8 6.1 6.5 6.8 6.8 7.1 21.3 18.7 14.5 ...
```

4.2. Descriptive dataset analysis

A very useful command is the function `summary`. Here this command gives us a summary of the datas (average, quantiles, . . .)

```
summary(conso_voit)
```

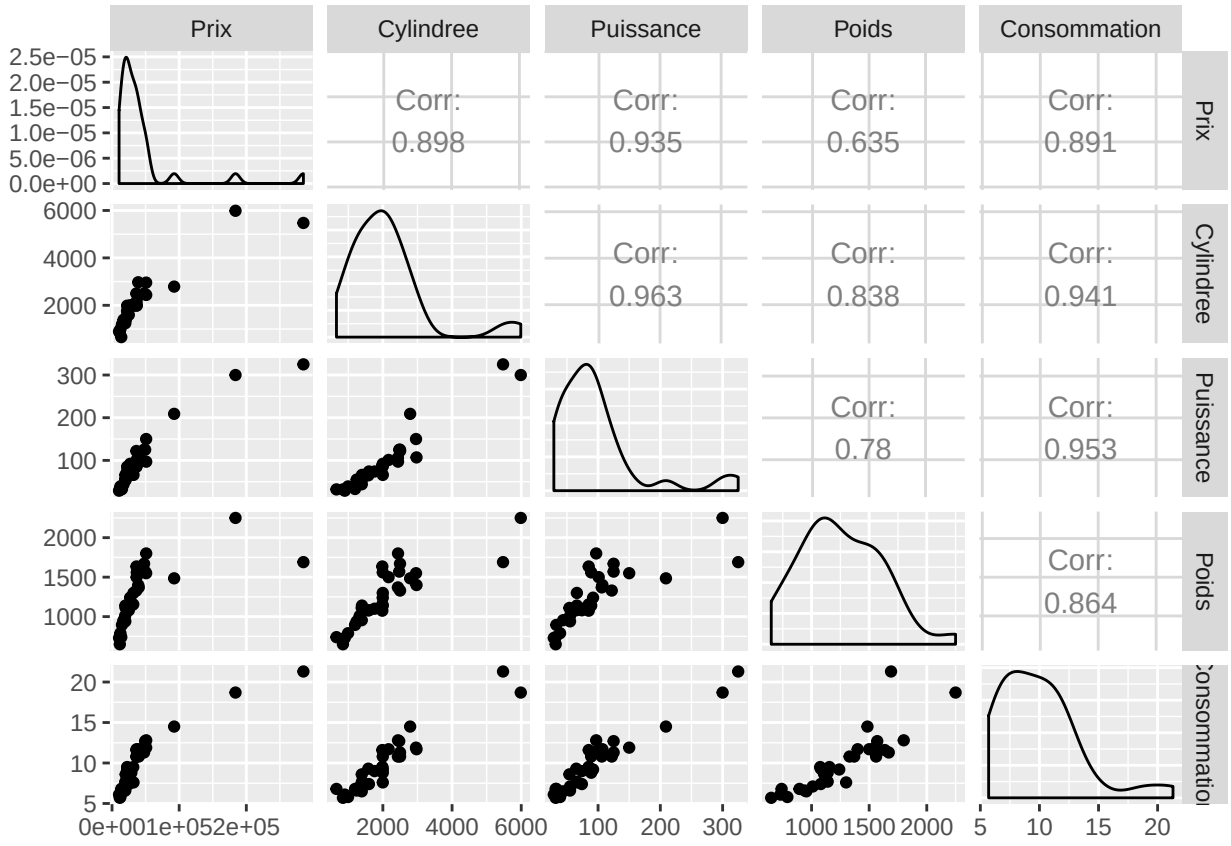
```
##       Prix          Cylindree      Puissance         Poids
##  Min.   : 10450   Min.   : 658   Min.   : 29.0   Min.   : 650
##  1st Qu.: 19820   1st Qu.:1390   1st Qu.: 55.0   1st Qu.:1042
##  Median : 28750   Median :1984   Median : 85.0   Median :1155
##  Mean   : 43756   Mean   :2094   Mean   : 97.1   Mean   :1256
##  3rd Qu.: 39395   3rd Qu.:2456   3rd Qu.:106.5   3rd Qu.:1525
##  Max.   :285000   Max.   :5987   Max.   :325.0   Max.   :2250
##   Consommation
##  Min.   : 5.700
##  1st Qu.: 7.250
##  Median : 9.300
##  Mean   : 9.955
##  3rd Qu.:11.650
##  Max.   :21.300
```

Some useful packages.

```
library(ggplot2)
library(dplyr)
library(GGally)
```

To visualize the relation between each pairs of variables, we use the commande `ggpairs`.

```
ggpairs(conso_voit)
```



It seems to be a linear link between variable `Consommation` and the others variables. On the diagonal, there are plot of the estimated density of each variable. Above the diagonal, it is given the value of the correlations between variables. The command `cor` also gives the correlations.

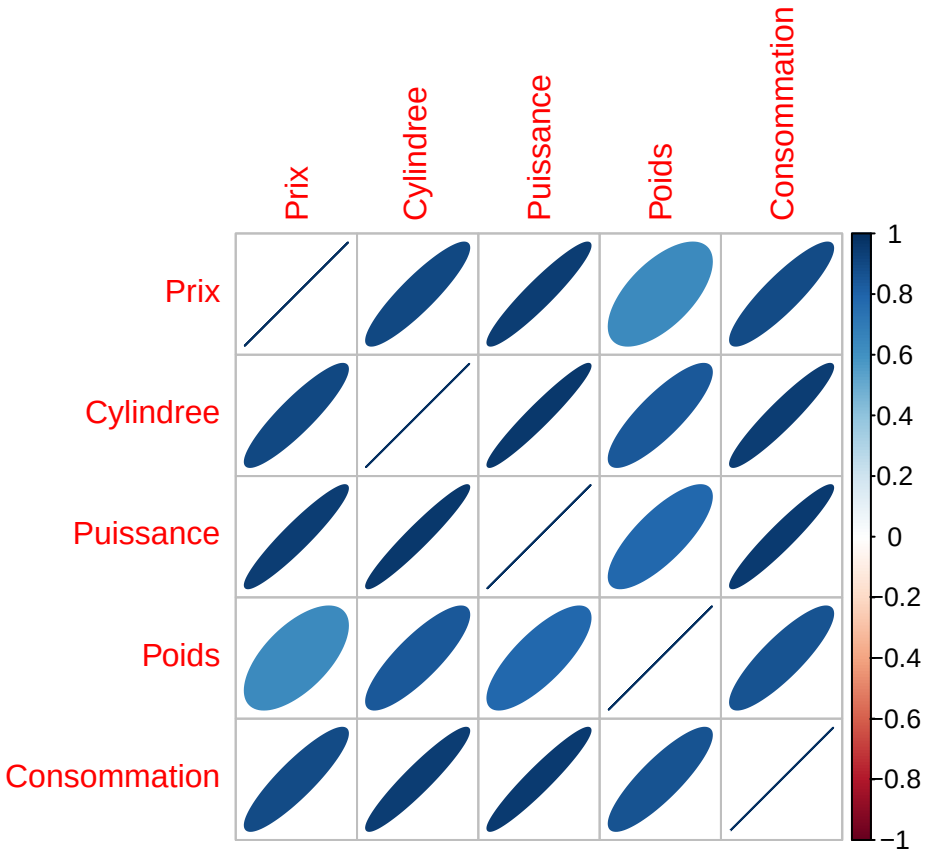
```
library(corrplot)
```

```
cor(conso_voit)
```

```
##                    Prix Cylindree Puissance     Poids Consommation
## Prix          1.0000000 0.8977790 0.9351708 0.6349611    0.8911104
## Cylindree     0.8977790 1.0000000 0.9625134 0.8378676    0.9409920
## Puissance     0.9351708 0.9625134 1.0000000 0.7798228    0.9526249
## Poids         0.6349611 0.8378676 0.7798228 1.0000000    0.8638623
## Consommation  0.8911104 0.9409920 0.9526249 0.8638623    1.0000000
```

A graphical visualization of the correlations is easier to interpret graphically by using the function `corrplot` of the library `corrplot`).

```
r=round(cor(conso_voit),2)
corrplot(r,method="ellipse")
```



The `Consommation` is very correlated with the 4 others variables. Note that the variable `Cylindree` and `Puissance` are highly correlated.

4.3. The linear regression model

The linear regression of the `Consommation` variable on the other variables is done using the `lm` function.

```
reg = lm(Consommation~Prix+Cylindree+Puissance+Poids,data=conso_voit)
```

Comments:

- ☛ By default **R** adds an intercept (a column of 1). Here, the design matrix X is a matrix of size $n \times p$ with $p = 5$.
- ☛ The order in which variables are entered gives the indice j of the regressor X_j
  - Y =`Consommation` = Fuel consumption in liters per 100 km.

- X_1 =`Prix` = Vehicle price in Swiss francs.
- X_2 =`Cylindree` = Cylinder capacity in cm3.
- X_3 =`Puissance` = Power in kW.
- X_4 =`Poids` = Weight in kg.

☛ Then, the linear model defined here is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i, \ \forall i = 1, \cdots, n. \tag{1}$$

☛ We assume that the postulates **[P1]**–**[P4]** are statisfied.

We then visualize the results using the function `summary`.

```
summary(reg)
```

```
##
## Call:
## lm(formula = Consommation ~ Prix + Cylindree + Puissance + Poids,
##     data = conso_voit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5677 -0.6704  0.1183  0.5283  1.4361
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.456e+00  6.268e-01   3.919 0.000578 ***
## Prix         2.042e-05  8.731e-06   2.339 0.027297 *
## Cylindree   -5.006e-04  5.748e-04  -0.871 0.391797
## Puissance    2.499e-02  9.992e-03   2.501 0.018993 *
## Poids        4.161e-03  8.788e-04   4.734 6.77e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8172 on 26 degrees of freedom
## Multiple R-squared:  0.9546, Adjusted R-squared:  0.9476
## F-statistic: 136.5 on 4 and 26 DF,  p-value: < 2.2e-16
```

Interpretation of R outputs

➤ **`Call`** : A reminder of the formula used.

➤ **`Residuals`** : A summary descriptive analysis of residues $\widehat{\varepsilon_i}$.

➤ **`Coefficients`** : This table includes in columns:

- **`Estimate`** : The value of $\widehat{\beta}_j$ the least square estimator $\widehat{\beta}$ (which is the maximum likelihood estimator under **[P1]**–**[P4]**). Here

$$\widehat{\beta}_0 = 2.456e{+}00, \quad \widehat{\beta}_1 = 2.042e{-}05, \quad \widehat{\beta}_2 = -5.006e{-}04, \quad \widehat{\beta}_3 = 2.499e{-}02, \quad \widehat{\beta}_4 = 4.161e{-}03$$

- **`Std. Error`** : The value of $\widehat{\sigma}_j = \widehat{\mathbb{V}\mathrm{ar}_\beta(\widehat{\beta}_j)}$, estimator of the standard deviation of $\widehat{\beta}_j$.
- **`t value`** : Here we test

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0.$$

  The `t value` is the value of the Student test statistic T, such that under H_0

$$T = \frac{\widehat{\beta}_j}{\widehat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}} \sim t_{n-p}$$

  with $(X^TX)^{-1}_{jj}$ the $j-th$ diagonal element of the matrix $(X^TX)-1$.
- **`Pr(>|t|)`** : The $p-value$ of the previous Student tests.

➤ **`Signif. codes`** : Significance level symbols.

➤ **`Residual standard error`** : The value of $\widehat{\sigma}$ and the number of degrees of freedom : $(n-p)$ (here $31-5=26$). Here

$$\widehat{\sigma}^2 = 0.8172^2$$

➤ **`Multiple R-squared`** : The value of $R^2 = 0.9546$.

➤ **`Adjusted R-squared`** : The value of the adjusted R^2 : $R^2_a = 0.9476$.

➤ **`F-statistic`** : Here we test

$$H_0 : Y_i = \beta_0 + \varepsilon_i \quad \text{vs} \quad H_1 : Y_i = \beta_0 + \sum_{j=1}^{4} \beta_j X_{ij} + \varepsilon_i.$$

The `F-statistic` is the value of the Fisher's global test statistic F such that under H_O

$$F = \frac{\|P_X Y - \overline{Y}\mathbf{1}\|^2/(p-1)}{\widehat{\sigma}^2} \sim F_{(p-1,n-p)}.$$

In this example, $F = 136.5$ and the associated degrees of freedom $(p-1, n-p) = (4, 26)$. The $p-value < 2.2e-16$ is very small so we reject H_0, the test is meaningful.

4.4. Model validation

We recall that we assume the model~(1), under the Rank assumption and under **[P1]**–**[P4]** where

- **[P1]** : Errors are centered/(the model is linear) : $\forall i = 1, \cdots, n \quad \mathbb{E}_\beta[\varepsilon_i] = 0$.
- **[P2]** : Errors have homoscedastic variance : $\forall i = 1, \cdots, n \quad \mathbb{V}\text{ar}_\beta[\varepsilon_i] = \sigma^2 > 0$.
- **[P3]** : Errors are uncorrelated: $\forall i \neq j \quad \mathbb{C}\text{ov}(\varepsilon_i, \varepsilon_j)$.
- **[P4]** : Errors are gaussian : $\forall i = 1, \cdots, n \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

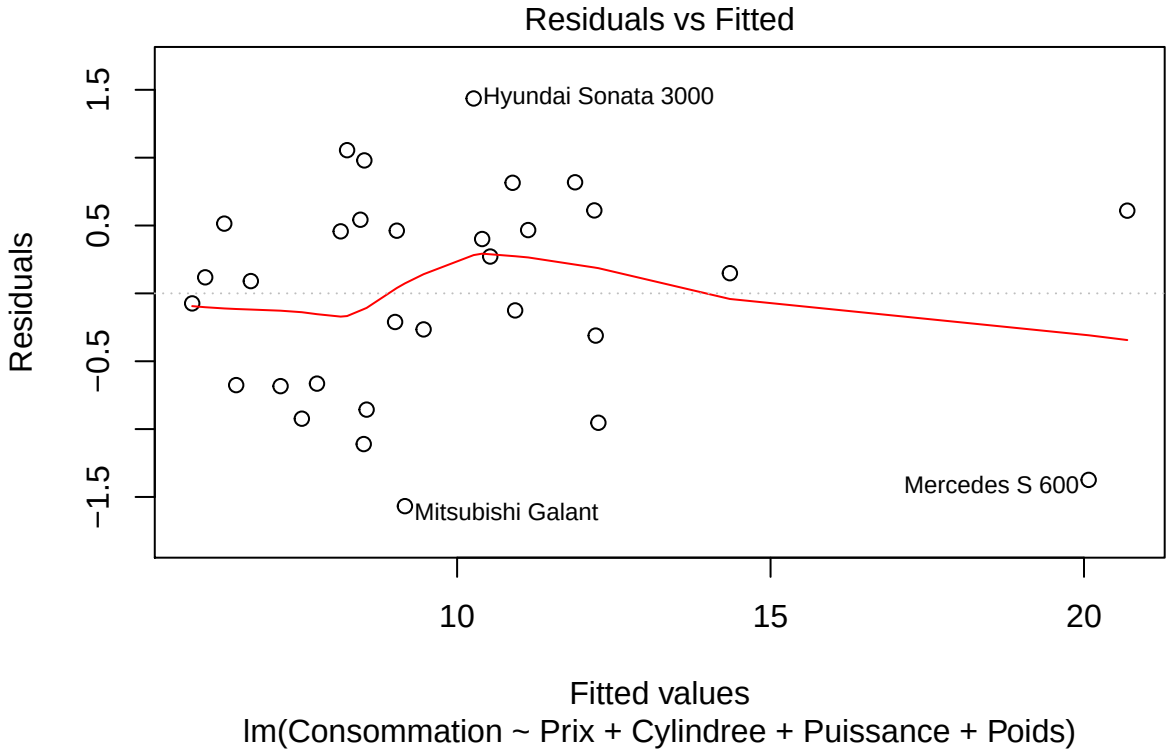
The rank hypothesis is easily verifiable with a simple calculation. For the other assumptions **[Pi]**, this requires an analysis of the residus. First upload the needed **R** library.

```
library(MASS)
library(carData)
library(car)
```

Validation of the postulates [P1]: Errors are centered

The centered postulat (the linearity assumption in practice) can be assesed by inspecting the *Residuals vs Fitted*-plot (or the *Studentized residuals*-plots). The command for the *Residuals vsFitted*-plot is `plot( ,which=1)`.

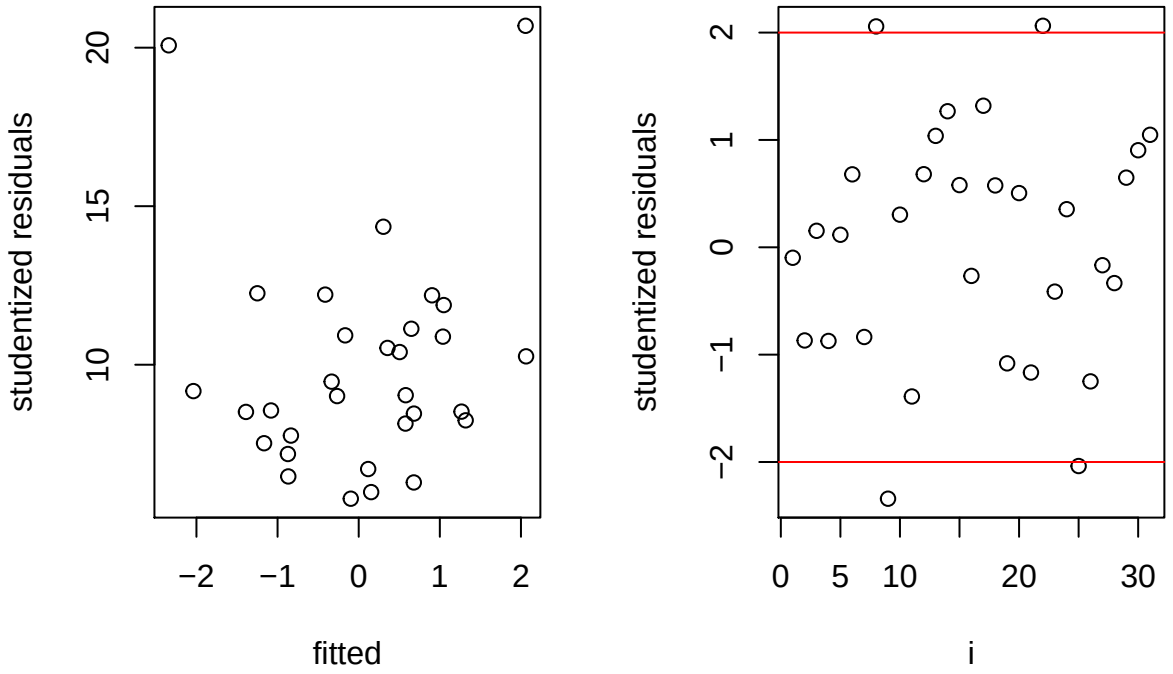
```
plot(reg,which=1)
```



The plot shows that when the responses predicted by the model (fitted values) increase, the residuals remain globally uniformly distributed on both sides of 0. The red line is approximately horizontal at zero.

The command `stdres()` displays the *Studentized residuals* $(t_i^*)_i$, needed to draw the *Studentized residuals*-plot.

```
par(mfrow=c(1,2))
SR=stdres(reg)
plot(SR,fitted(reg),xlab="fitted",ylab="studentized residuals")
plot(SR,xlab="i",ylab="studentized residuals")
abline(h=2,col="red")
abline(h=-2,col="red")
```

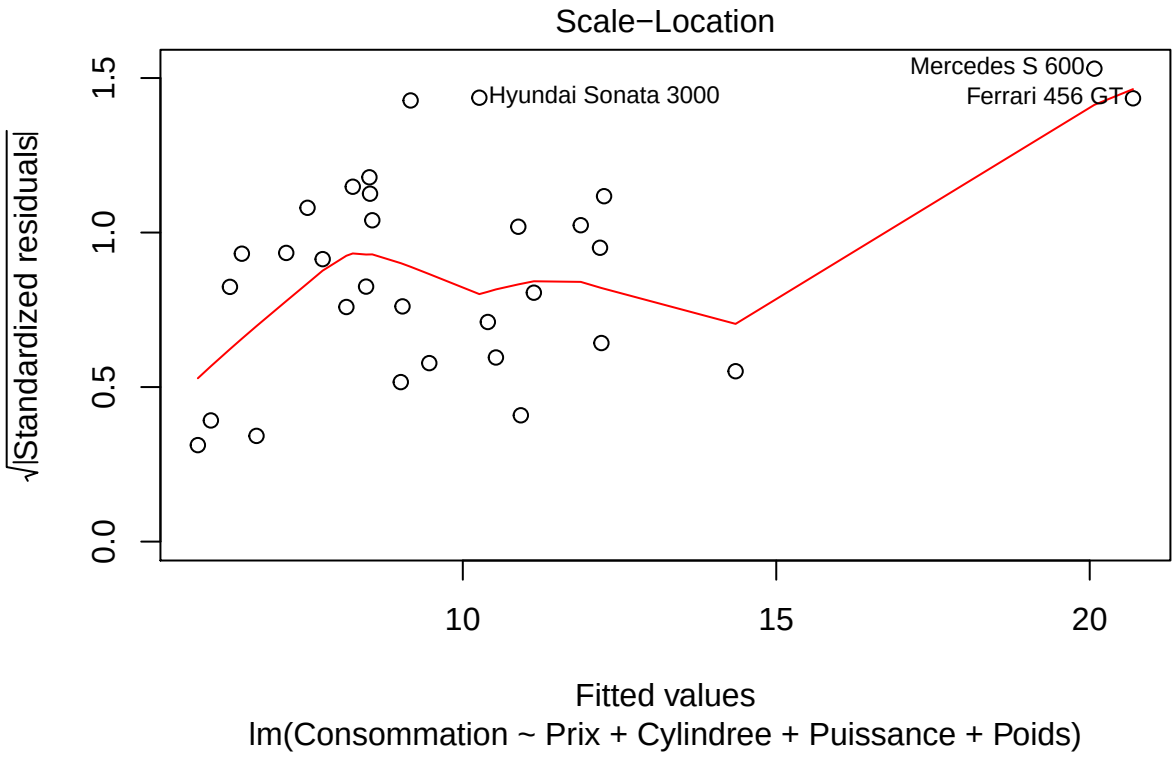


The plots show no fitted pattern, the residuals remain globally (reasonably) uniformly distributed. therefore that the assumption of linearity is acceptable. Thus, we validate the postulate.

Validation of the postulates [P2]: Errors have homoscedastic variance

The homoscedastic assumption can be checked by examining the *Scale-location*-plot (the command `plot( ,which=3)`). The postulate is validated if we see a horizontal line with equally spread points. In our example, it seems difficult to validate the postulat. So, let us make a Breush-Pagan test (H_O : homoscedasticity) to assess it.

```
plot(reg,which=3)
```



The command for the Breush-Pagan test is `ncvTest`. The homoscedasticity is rejected if the *p-value* is less than 0.05. Here, *p-value*= 0.38455 > 0.05, the posutlate is validated.

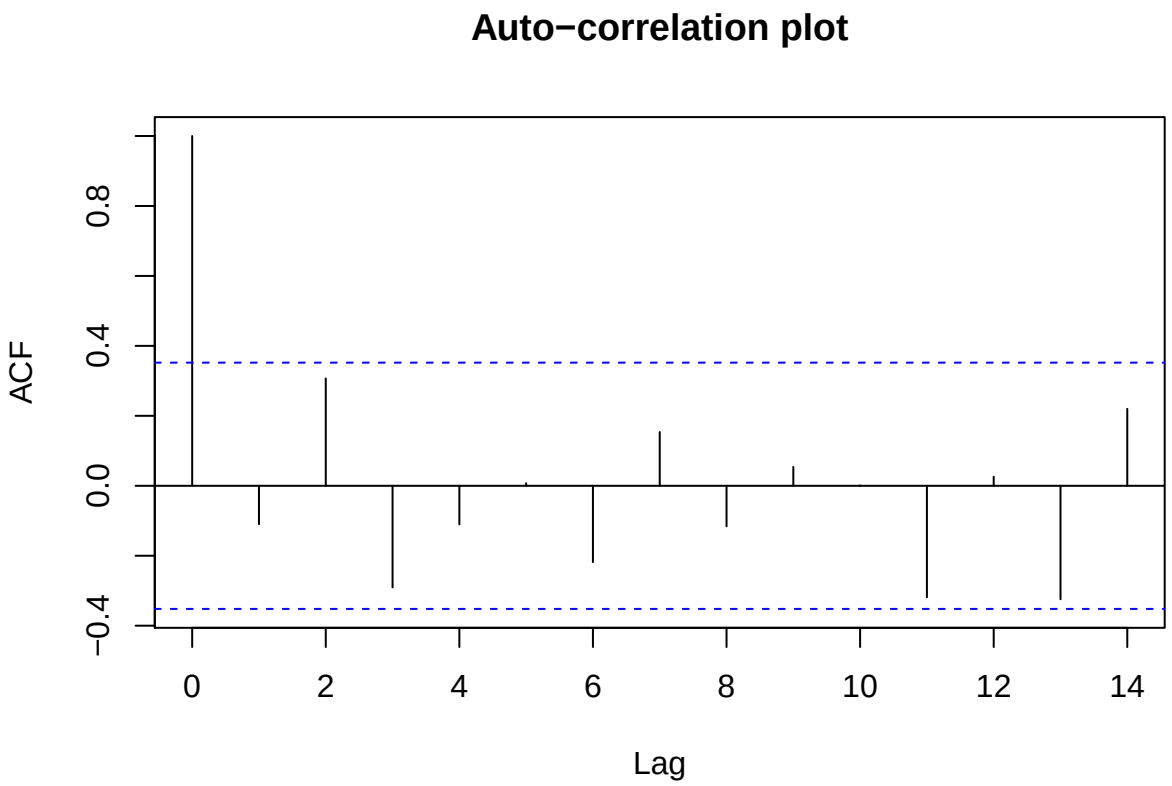
```
ncvTest(reg)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.7560996, Df = 1, p = 0.38455
```

Validation of the postulate [P3]: Errors are uncorrelated

Under **R**, we can represent the auto-correlation of the residuals using the command `acf()`. In our example, exept the first one, none exceeds dashed thresholds thus uncorrelation is satisfied.

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



The Durbin-Watson test can be also used to validate this assumption. The command is `durbinWatsonTest`. Under the null hypothesis the residual are considered auto-uncorrelated.

```
durbinWatsonTest(reg)
```

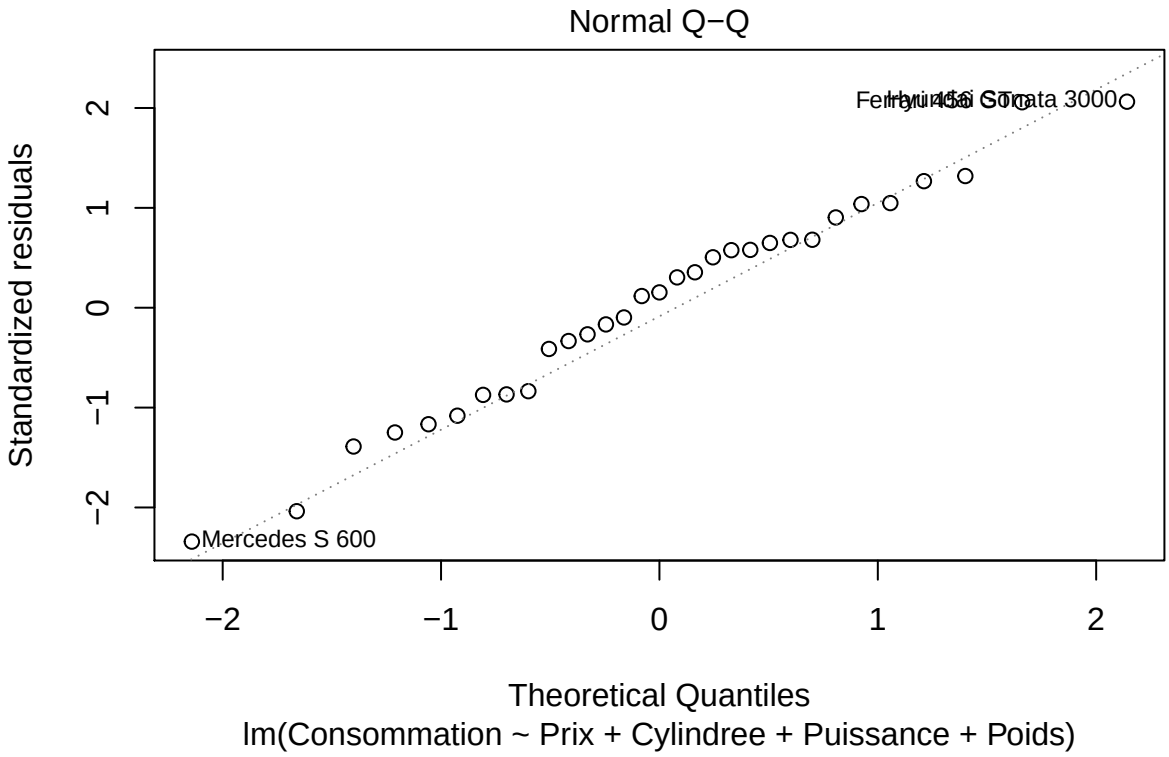
```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.1096954      2.180495   0.764
##  Alternative hypothesis: rho != 0
```

Here, the *p-value*= 0.782 > 0.05 thus we can't reject H_0, the posutlate is validated.

Validation of the postulate [P4]: Errors are gaussian

To analyze the normality, we use the Q-Q plot with the command `plot( ,which=2)`. The points appear reasonably aligned along the reference line even the sample size $n = 31$ is small, then the postulate is validated.

```
plot(reg,which=2)
```



The Shapiro-Wilk test can also be used to assess the normality of residuals. The normality assumption is rejected if the *p-value* is less than 0.05. Here, *p-value*> 0.05, the posutlate is unvalidated. But as the sample size is small, it was expected. However, we assume that the postulate is verified.

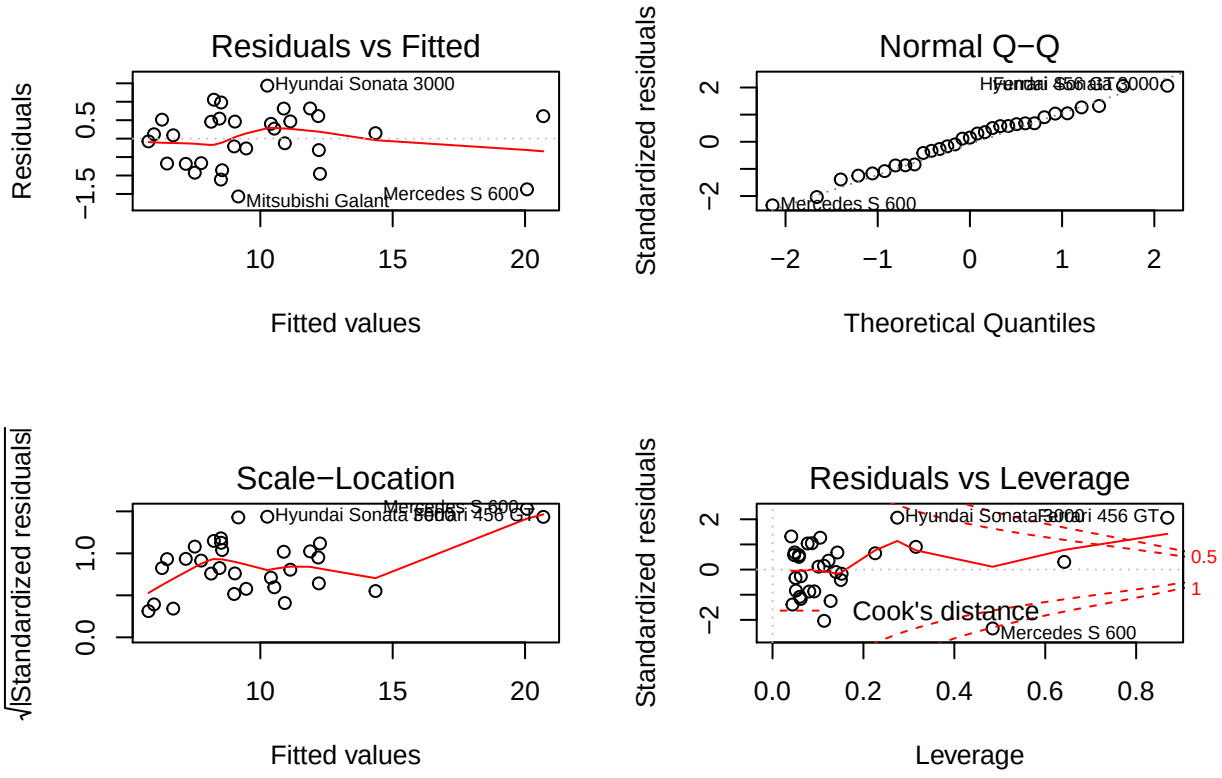
```
shapiro.test(residuals((reg)))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals((reg))
## W = 0.9709, p-value = 0.5442
```

Command `plot`: to resume

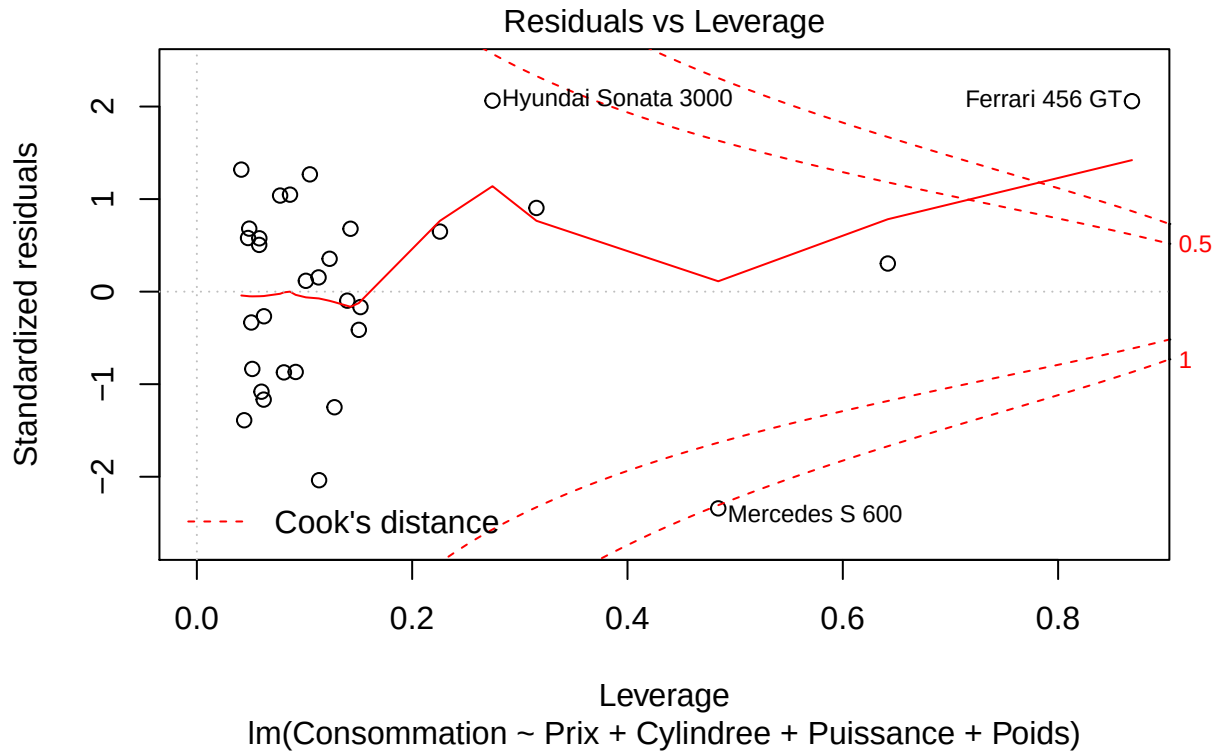
Note that by default, the command `plot)` gives 4 different plots in the linear regression setting. Three of them have been seen and used to validate the postulat. The last one is the `Residuals vs Leverage`-plot.

```
par(mfrow=c(2,2))
plot(reg)
```



The `Residuals vs Leverage`-plot can be called alone as follows :

```
plot(reg,which=5)
```

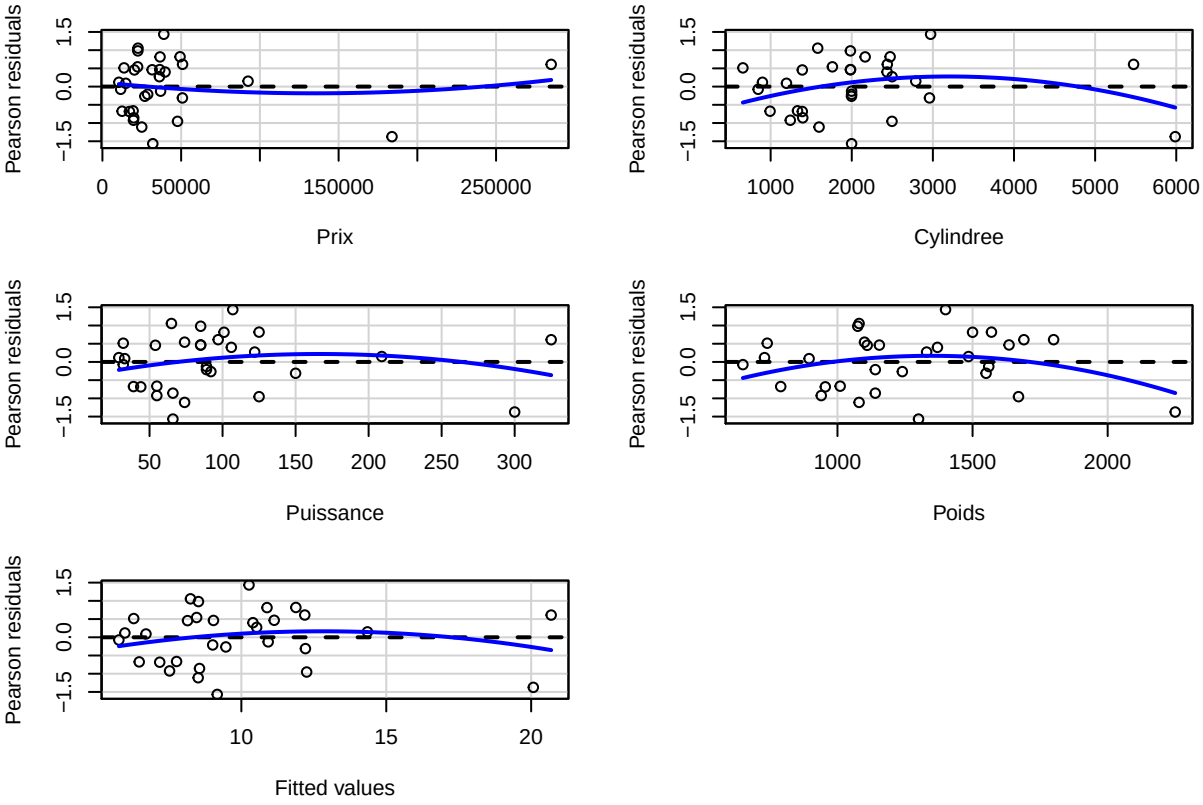


It appears in this plot, that 2 observations have a Cook's distance larger than 1. They are *ouliers* : *regression ouliers* or *leverage points* or both. A sudy of ouliers is done in a further section.

For going further...

For going further, to see the contribution of each variable, we can use the `residualPlots` function from the `cars` library. (it will be discussed in classroom)?

```
residualPlots(reg)
```




```
##             Test stat Pr(>|Test stat|)
## Prix           1.1523          0.26009
## Cylindree     -2.2748          0.03176 *
## Puissance     -2.4246          0.02289 *
## Poids         -1.6631          0.10878
## Tukey test    -2.1976          0.02798 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.5 Confidence Interval

The `confint` command easily displays confidence intervals for the parameters β_j of the model. They are based on a Student's law, if the postulate **[P4]** is satisfied. If not, these intervals are biased.

```
cbind(confint(reg),coef(reg))
```

```
##                      2.5 %       97.5 %
## (Intercept)  1.167851e+00 3.744737e+00  2.456294e+00
## Prix         2.474392e-06 3.836669e-05  2.042054e-05
## Cylindree   -1.682157e-03 6.809703e-04 -5.005933e-04
## Puissance    4.455929e-03 4.553302e-02  2.499448e-02
## Poids        2.354210e-03 5.966955e-03  4.160583e-03
```

The `predict( ,interval = "confidence")` command displays confidence interval for $x_i^T\beta$ (say, for *estimation*); while the `predict( ,interval = "prediction")` command displays confidence interval for Y_i (say, for *prediction*). Note that $x_i^T\beta$ and Y_i are both estimate by $x_i^T\widehat{\beta}$, but the bound of the confidence interval are differents.

```
ICconf = predict(reg, interval = "confidence", level = 0.95)
head(ICconf)
```

```
##                           fit      lwr      upr
## Daihatsu Cuore       5.773872 5.145857 6.401888
## Suzuki Swift 1.0 GLS 6.475902 5.966890 6.984914
## Fiat Panda Mambo L   5.981720 5.416875 6.546566
## VW Polo 1.4 60       7.183591 6.705853 7.661329
## Opel Corsa 1.2i Eco  6.709359 6.174759 7.243959
## Subaru Vivio 4WD     6.285932 5.651261 6.920603
```

```
ICpred= predict(reg, interval = "prediction", level = 0.95)
```

```
## Warning in predict.lm(reg, interval = "prediction", level = 0.95): predictions on c
```

```
head(ICpred)
```

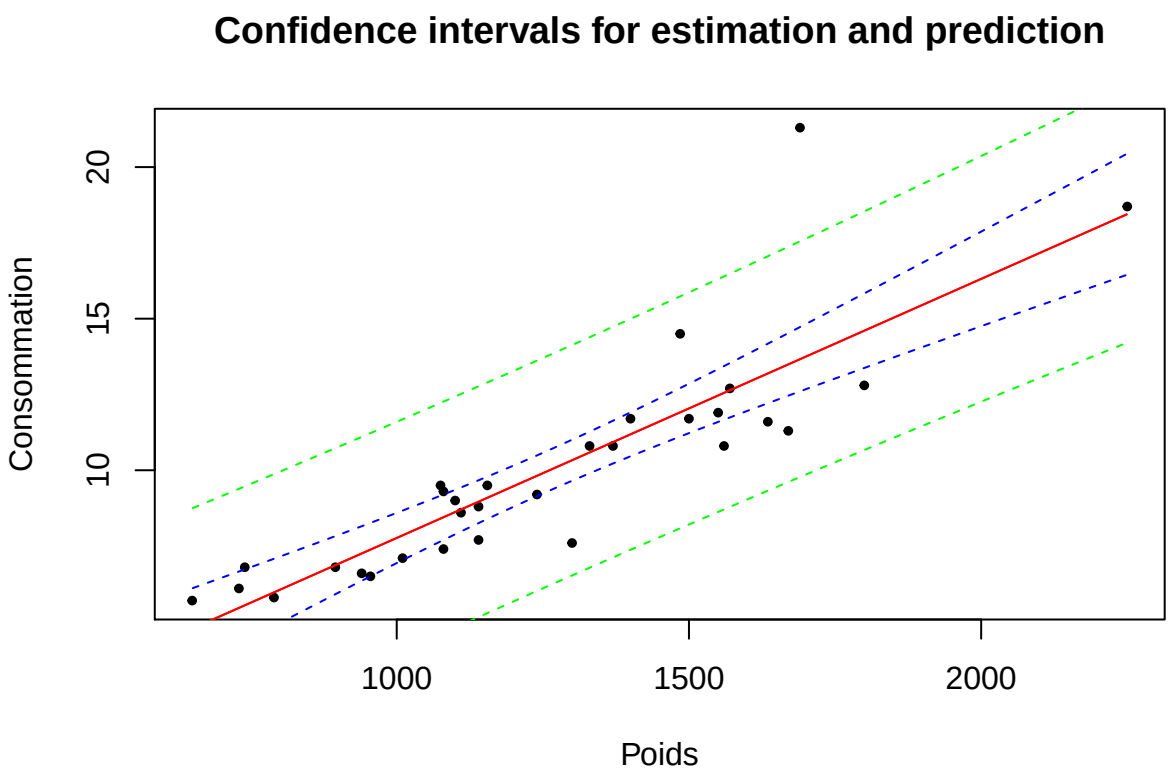
```
##                           fit      lwr      upr
## Daihatsu Cuore       5.773872 3.980461 7.567284
## Suzuki Swift 1.0 GLS 6.475902 4.720620 8.231184
## Fiat Panda Mambo L   5.981720 4.209442 7.753999
## VW Polo 1.4 60       7.183591 5.437121 8.930060
## Opel Corsa 1.2i Eco  6.709359 4.946486 8.472231
## Subaru Vivio 4WD     6.285932 4.490179 8.081685
```

Let us consider the simple regression of the variable `Consommation` on the predictor `Poids`, and consider new data `Poids` (called `newgrille` in our program). The command `predict.lm( ,interval = 'confidence')` displays prediction of the new $x_{i_{new}}^T\beta$ and the command `predict.lm( ,interval = 'prediction')` displays prediction of the new $Y_{i_{new}}$.

```
regP=lm(Consommation~Poids,data=conso_voit)
newgrille=data.frame(Poids=seq(min(conso_voit$Poids)+1,max(conso_voit$Poids)-1),2)
predicgrille=predict.lm(regP,newgrille, interval='confidence',level=0.95)
head(predicgrille)
```

```
##        fit      lwr      upr
## 1 4.783165 3.455416 6.110914
## 2 4.791711 3.465595 6.117828
## 3 4.800257 3.475773 6.124742
## 4 4.808804 3.485950 6.131658
## 5 4.817350 3.496126 6.138574
## 6 4.825897 3.506302 6.145491
```

```
plot(conso_voit$Poids,conso_voit$Consommation,ylab="Consommation",xlab="Poids",pch=20,
#abline(regP,col='red')
matlines(newgrille$Poids,predicgrille,lty=c(1,2,2),col=c("red","blue","blue"))
predicgrille=predict.lm(regP,newgrille, interval='prediction',level=0.95)
matlines(newgrille$Poids,predicgrille,lty=c(1,2,2),col=c("red","green","green"))
```

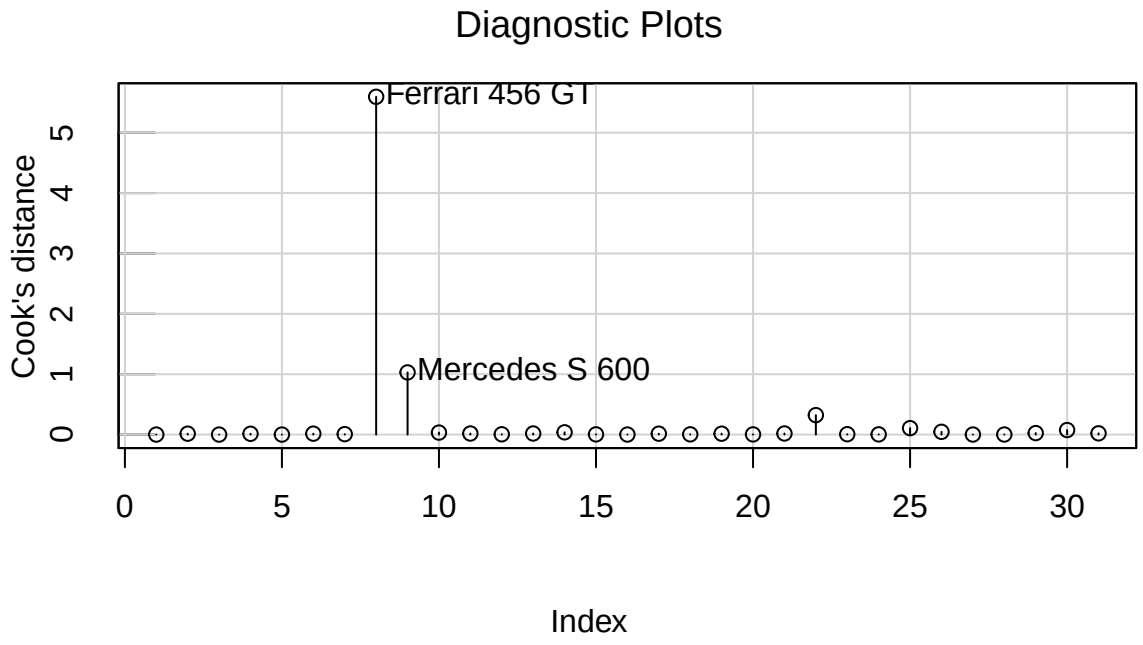


4.6. Outliers and leverage points

The library `car` offers an easy way to detect graphically atypical observations and to assess about their nature by tests. The command `influenceIndexPlot` is an important one.

Cook's distance plot

```
influenceIndexPlot(reg,vars="Cook")
```

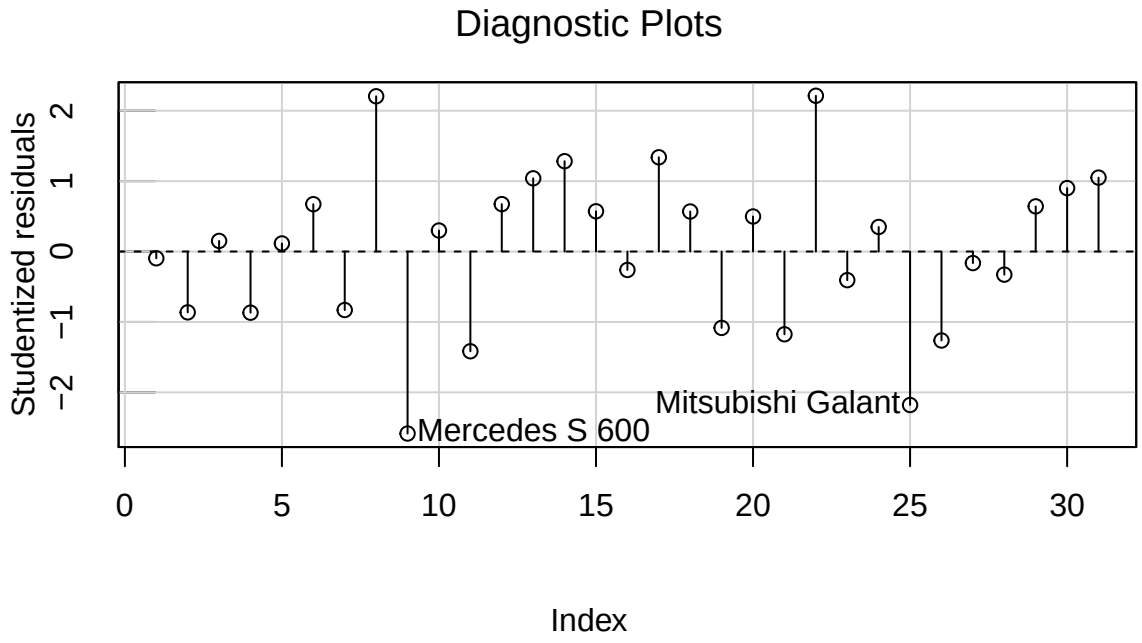


The Cook's distance plot highlight the influence of each observation on the estimation of the model (on β). As seen on the previous chapter, we compare the Cook's distance with 1. Here, two observations have a Cook's distance larger than 1 :

`Ferrari 456 GT` and `Mercedes S 600`

Studentized plot

```
influenceIndexPlot(reg,vars="Studentized")
```

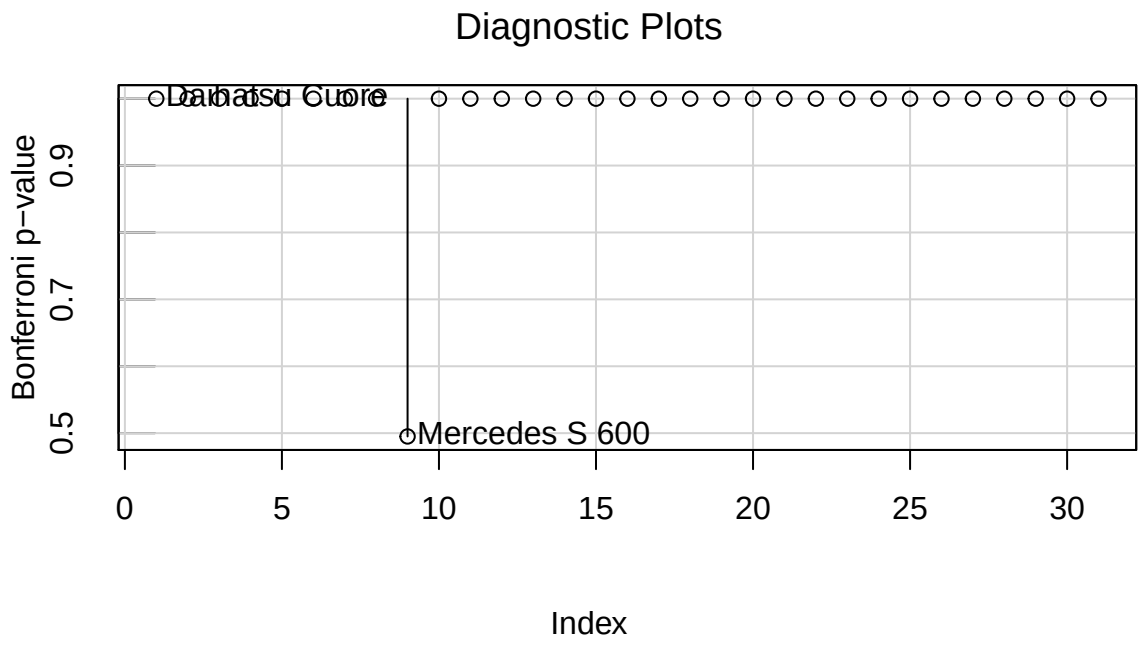


This plot, that of studentized residuals also makes it possible to highlight `outliers`. Here, two observations seems doubtful :

`Mitsubishi Galant` and `Mercedes S 600`

Bonferroni plot

```
influenceIndexPlot(reg,vars="Bonf")
```

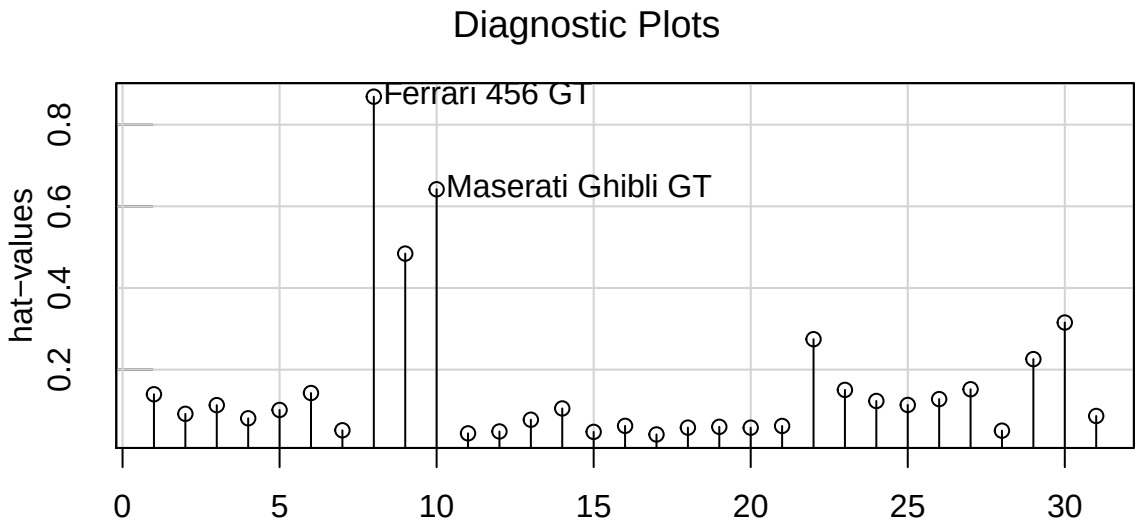


This is the plot of the *p-value* Bonferroni. Is considered as an outlier is a observation with a p-value less than 0.05. Here, the plot detetects one observation :

```
Mercedes S 600
```

Hat plot

```
influenceIndexPlot(reg,vars="hat")
```



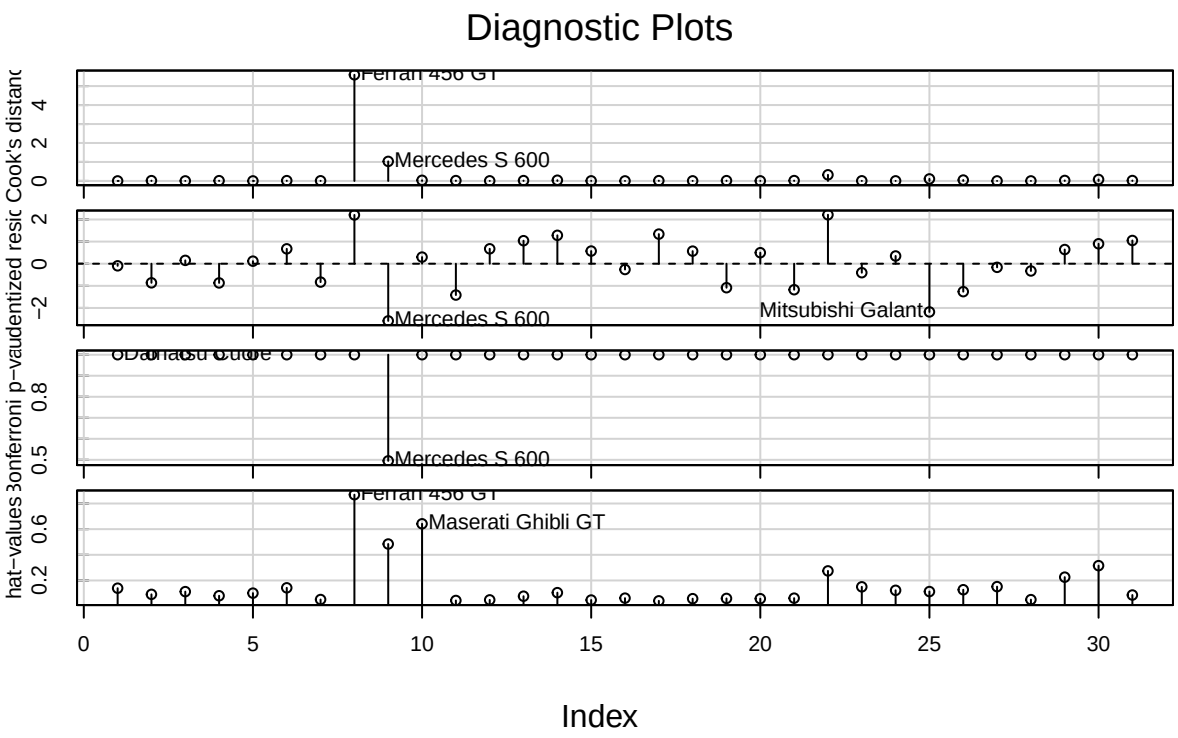
This plot called the *hat value*-plot, reflects the leverage (h_{ii}) of each observation on its own estimate. An observation is considered to be a *leverage point* when this value is less than 0.05. Here, the plot detetects two observations :

`Ferrari 456 GT` and `Maserati Ghibli GT`

The all plots

It's better to display these four graphs in parallel to have a better vision of the atypical points found in the various plot. It can be done as follows :

```
influenceIndexPlot(reg)
```



Here, the doubtful observations are :

`Mercedes S 600` and `Ferrari 456 GT`

We can access to Bonferroni's with the `outlierTest` command.

```
outlierTest(reg)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##                 rstudent unadjusted p-value Bonferroni p
## Mercedes S 600 -2.584781           0.01597      0.49506
```

The adjusted *p-value* by the Bonferonni method is equal to 0.49506 and is very far from the threshold of 0.05. The `Mercedes S 600` observation can not be considered as `outlier`.

To assess if the observations `Ferrari 456 GT` and `Mercedes S 600` really affect the estimation ouf our model (say β), we can compare the results of the estimation of β with and without these observation. We can done this with the command `comparCoefs`.

```
regbis = lm(Consommation~Prix + Cylindree + Puissance + Poids, data=
               conso_voit[-c(which(conso_voit_complet$Type=="Ferrari 456 GT"),
                             which(conso_voit_complet$Type=="Mercedes S 600")),])
compareCoefs(reg ,regbis)
```

```
## Calls:
## 1: lm(formula = Consommation ~ Prix + Cylindree + Puissance + Poids,
##   data = conso_voit)
## 2: lm(formula = Consommation ~ Prix + Cylindree + Puissance + Poids,
##   data = conso_voit[-c(which(conso_voit_complet$Type ==
##   "Ferrari 456 GT"), which(conso_voit_complet$Type == "Mercedes S 600")),
##    ])
##
##                Model 1   Model 2
## (Intercept)      2.456     2.071
## SE               0.627     0.664
##
## Prix          2.04e-05  2.56e-05
## SE            8.73e-06  3.03e-05
##
## Cylindree    -0.000501  0.000327
## SE            0.000575  0.000700
##
## Puissance      0.02499   0.01777
## SE             0.00999   0.01554
##
## Poids         0.004161  0.003598
## SE            0.000879  0.001054
##
```

It comes out that, that the observations `Ferrari 456 GT` and `Mercedes S 600` have little influence on the coefficients of the model parameters, as well as on their standard error, since the values do not vary much, even if they have a Cook's distance larger than 1.

```
#summary(regbis)
#plot(regbis)
```

Introduction to Regression - Chapter 4

MAP 535

Contents

Chapter 4: Model estimation/validation through R example

In this Chapter, we will see through an example how to estimate and validate a model. Moreover, we see how to detect atypical points.

4.1. The data

To illustrate the method, we consider the example~**??**.

> We try to explain and predict gasoline consumption (in liters per 100 km) of different automobile models based on several variables. For this, we have the following characteristics for 31 different cars:
>
> - `Type` = Type of the vehicle.
> - `Consommation` = Fuel consumption in liters per 100 km.
> - `Prix` = Vehicle price in Swiss francs.
> - `Cylindree` = Cylinder capacity in cm3.
> - `Puissance` = Power in kW.
> - `Poids` = Weight in kg.
>
> In this example, Y is the variable `Consommation`. The variables X_j correspond to the other 4 variables.

We first download the dataset `conso.txt` with the command `read.table`.

```
conso_voit = read.table("conso.txt", header=TRUE, sep="\t", dec=",",row.names=1)
conso_voit_complet = read.table("conso.txt", header=TRUE, sep="\t", dec=",")
```

To print the names of the variables is possible with the command `names`.

```
names(conso_voit_complet)
```

```
## [1] "Type"         "Prix"         "Cylindree"    "Puissance"
## [5] "Poids"        "Consommation"
```

To display the data in a table, we use the `knitr` library and the function `kable`.

```
library(knitr)
```

```
kable(conso_voit_complet)
```

Type	Prix	Cylindree	Puissance	Poids	Consommation
Daihatsu Cuore	11600	846	32	650	5.7
Suzuki Swift 1.0 GLS	12490	993	39	790	5.8
Fiat Panda Mambo L	10450	899	29	730	6.1
VW Polo 1.4 60	17140	1390	44	955	6.5
Opel Corsa 1.2i Eco	14825	1195	33	895	6.8
Subaru Vivio 4WD	13730	658	32	740	6.8
Toyota Corolla	19490	1331	55	1010	7.1
Ferrari 456 GT	285000	5474	325	1690	21.3
Mercedes S 600	183900	5987	300	2250	18.7
Maserati Ghibli GT	92500	2789	209	1485	14.5
Opel Astra 1.6i 16V	25000	1597	74	1080	7.4
Peugeot 306 XS 108	22350	1761	74	1100	9.0
Renault Safrane 2.2. V	36600	2165	101	1500	11.7
Seat Ibiza 2.0 GTI	22500	1983	85	1075	9.5
VW Golt 2.0 GTI	31580	1984	85	1155	9.5
Citroen ZX Volcane	28750	1998	89	1140	8.8
Fiat Tempra 1.6 Liberty	22600	1580	65	1080	9.3
Fort Escort 1.4i PT	20300	1390	54	1110	8.6
Honda Civic Joker 1.4	19900	1396	66	1140	7.7
Volvo 850 2.5	39800	2435	106	1370	10.8
Ford Fiesta 1.2 Zetec	19740	1242	55	940	6.6
Hyundai Sonata 3000	38990	2972	107	1400	11.7
Lancia K 3.0 LS	50800	2958	150	1550	11.9
Mazda Hachtback V	36200	2497	122	1330	10.8
Mitsubishi Galant	31990	1998	66	1300	7.6
Opel Omega 2.5i V6	47700	2496	125	1670	11.3
Peugeot 806 2.0	36950	1998	89	1560	10.8
Nissan Primera 2.0	26950	1997	92	1240	9.2
Seat Alhambra 2.0	36400	1984	85	1635	11.6
Toyota Previa salon	50900	2438	97	1800	12.8
Volvo 960 Kombi aut	49300	2473	125	1570	12.7

The `dim` command displays the size of the data (number of lines, number of columns). Here $n = 31$ and we have 5 variables.

```
dim(conso_voit_complet)
```

```
## [1] 31  6
```

The command `head` allows you to view the first 6 lines of the data. The command `tail` allows you to view the last 6 lines of the data.

```
head(conso_voit_complet)
```

```
##                   Type  Prix Cylindree Puissance Poids Consommation
## 1       Daihatsu Cuore 11600       846        32   650           5.7
## 2 Suzuki Swift 1.0 GLS 12490       993        39   790           5.8
## 3   Fiat Panda Mambo L 10450       899        29   730           6.1
## 4       VW Polo 1.4 60 17140      1390        44   955           6.5
## 5  Opel Corsa 1.2i Eco 14825      1195        33   895           6.8
## 6     Subaru Vivio 4WD 13730       658        32   740           6.8
```

The command `str` allows us to chek if the nature of each variable is well determined by **R**. Here, there is no mistake.

```
str(conso_voit_complet)
```

```
## 'data.frame':    31 obs. of  6 variables:
##  $ Type        : Factor w/ 31 levels "Citroen ZX Volcane",..: 2 25 4 31 17 24 26 3
##  $ Prix        : int  11600 12490 10450 17140 14825 13730 19490 285000 183900 92500
##  $ Cylindree   : int  846 993 899 1390 1195 658 1331 5474 5987 2789 ...
##  $ Puissance   : int  32 39 29 44 33 32 55 325 300 209 ...
##  $ Poids       : int  650 790 730 955 895 740 1010 1690 2250 1485 ...
##  $ Consommation: num  5.7 5.8 6.1 6.5 6.8 6.8 7.1 21.3 18.7 14.5 ...
```

4.2. Descriptive dataset analysis

A very useful command is the function `summary`. Here this command gives us a summary of the datas (average, quantiles, . . .)

```
summary(conso_voit)
```

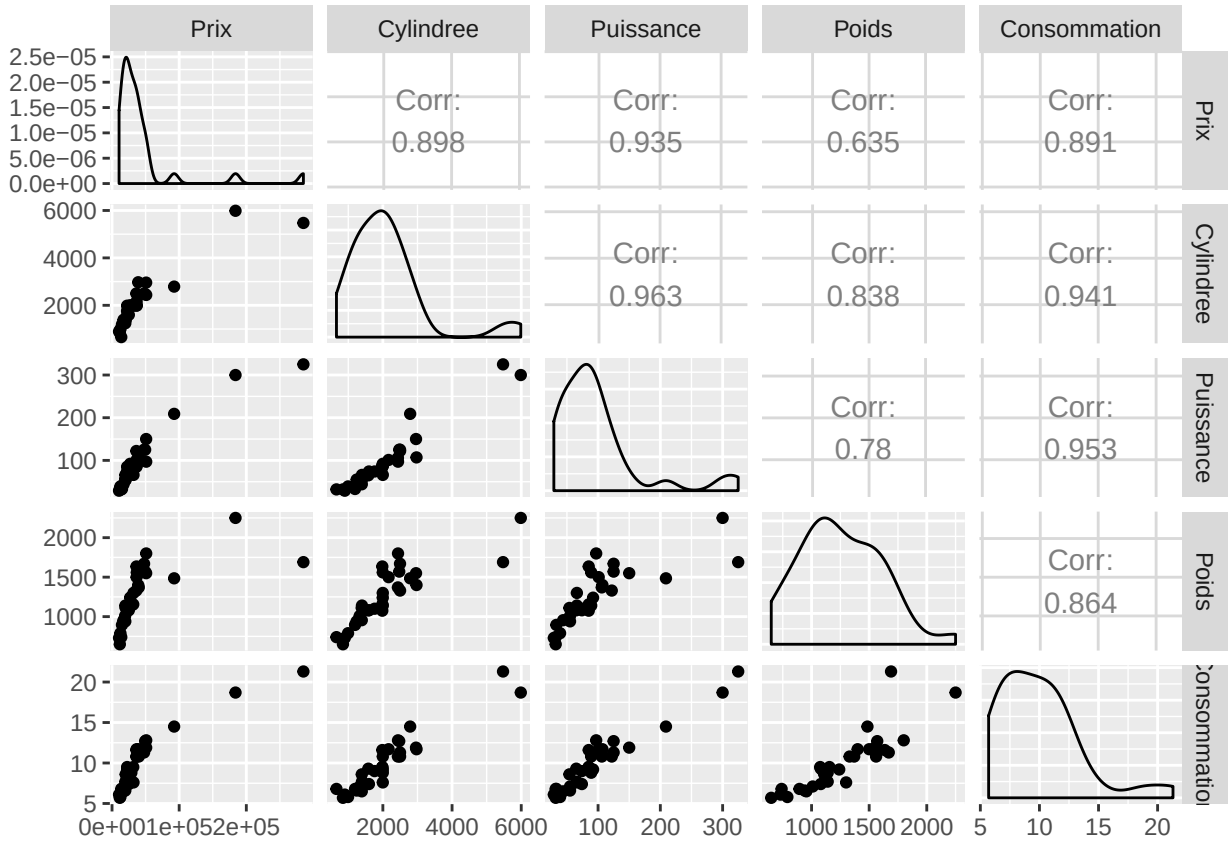
```
##       Prix          Cylindree      Puissance         Poids
##  Min.   : 10450   Min.   : 658   Min.   : 29.0   Min.   : 650
##  1st Qu.: 19820   1st Qu.:1390   1st Qu.: 55.0   1st Qu.:1042
##  Median : 28750   Median :1984   Median : 85.0   Median :1155
##  Mean   : 43756   Mean   :2094   Mean   : 97.1   Mean   :1256
##  3rd Qu.: 39395   3rd Qu.:2456   3rd Qu.:106.5   3rd Qu.:1525
##  Max.   :285000   Max.   :5987   Max.   :325.0   Max.   :2250
##   Consommation
##  Min.   : 5.700
##  1st Qu.: 7.250
##  Median : 9.300
##  Mean   : 9.955
##  3rd Qu.:11.650
##  Max.   :21.300
```

Some useful packages.

```
library(ggplot2)
library(dplyr)
library(GGally)
```

To visualize the relation between each pairs of variables, we use the commande `ggpairs`.

```
ggpairs(conso_voit)
```



It seems to be a linear link between variable `Consommation` and the others variables. On the diagonal, there are plot of the estimated density of each variable. Above the diagonal, it is given the value of the correlations between variables. The command `cor` also gives the correlations.

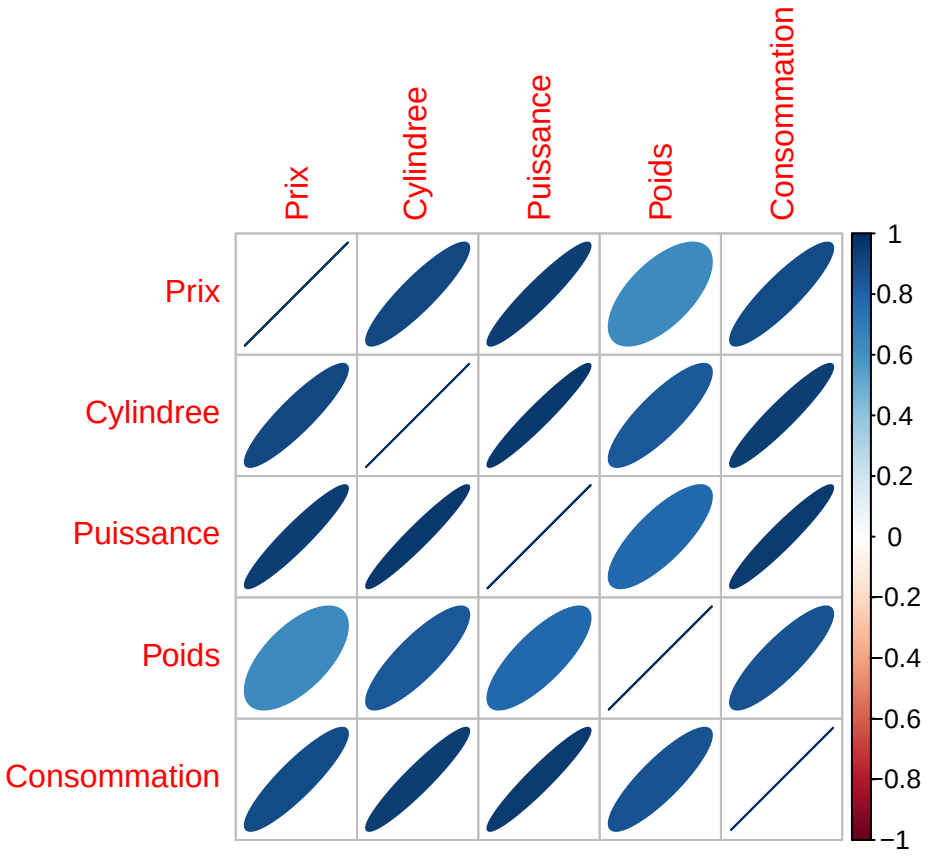
```
library(corrplot)
```

```
cor(conso_voit)
```

```
##                    Prix Cylindree Puissance     Poids Consommation
## Prix          1.0000000 0.8977790 0.9351708 0.6349611    0.8911104
## Cylindree     0.8977790 1.0000000 0.9625134 0.8378676    0.9409920
## Puissance     0.9351708 0.9625134 1.0000000 0.7798228    0.9526249
## Poids         0.6349611 0.8378676 0.7798228 1.0000000    0.8638623
## Consommation  0.8911104 0.9409920 0.9526249 0.8638623    1.0000000
```

A graphical visualization of the correlations is easier to interpret graphically by using the function `corrplot` of the library `corrplot`).

```
r=round(cor(conso_voit),2)
corrplot(r,method="ellipse")
```



The `Consommation` is very correlated with the 4 others variables. Note that the variable `Cylindree` and `Puissance` are highly correlated.

4.3. The linear regression model

The linear regression of the `Consommation` variable on the other variables is done using the `lm` function.

```
reg = lm(Consommation~Prix+Cylindree+Puissance+Poids,data=conso_voit)
```

Comments:

- ☛ By default **R** adds an intercept (a column of 1). Here, the design matrix X is a matrix of size $n \times p$ with $p = 5$.
- ☛ The order in which variables are entered gives the indice j of the regressor X_j
  - Y =`Consommation` = Fuel consumption in liters per 100 km.

- X_1 =`Prix` = Vehicle price in Swiss francs.
- X_2 =`Cylindree` = Cylinder capacity in cm3.
- X_3 =`Puissance` = Power in kW.
- X_4 =`Poids` = Weight in kg.

☛ Then, the linear model defined here is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i, \ \forall i = 1, \cdots, n. \tag{1}$$

☛ We assume that the postulates **[P1]**–**[P4]** are statisfied.

We then visualize the results using the function `summary`.

```
summary(reg)
```

```
##
## Call:
## lm(formula = Consommation ~ Prix + Cylindree + Puissance + Poids,
##     data = conso_voit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5677 -0.6704  0.1183  0.5283  1.4361
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.456e+00  6.268e-01   3.919 0.000578 ***
## Prix         2.042e-05  8.731e-06   2.339 0.027297 *
## Cylindree   -5.006e-04  5.748e-04  -0.871 0.391797
## Puissance    2.499e-02  9.992e-03   2.501 0.018993 *
## Poids        4.161e-03  8.788e-04   4.734 6.77e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8172 on 26 degrees of freedom
## Multiple R-squared:  0.9546, Adjusted R-squared:  0.9476
## F-statistic: 136.5 on 4 and 26 DF,  p-value: < 2.2e-16
```

Interpretation of R outputs

➤ **`Call`** : A reminder of the formula used.

➤ **`Residuals`** : A summary descriptive analysis of residues $\widehat{\varepsilon_i}$.

➤ **`Coefficients`** : This table includes in columns:

- **`Estimate`** : The value of $\widehat{\beta}_j$ the least square estimator $\widehat{\beta}$ (which is the maximum likelihood estimator under **[P1]**–**[P4]**). Here

$$\widehat{\beta}_0 = 2.456e{+}00, \quad \widehat{\beta}_1 = 2.042e{-}05, \quad \widehat{\beta}_2 = -5.006e{-}04, \quad \widehat{\beta}_3 = 2.499e{-}02, \quad \widehat{\beta}_4 = 4.161e{-}03$$

- **`Std. Error`** : The value of $\widehat{\sigma}_j = \widehat{\mathbb{V}\mathrm{ar}_\beta(\widehat{\beta}_j)}$, estimator of the standard deviation of $\widehat{\beta}_j$.
- **`t value`** : Here we test

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0.$$

  The `t value` is the value of the Student test statistic T, such that under H_0

$$T = \frac{\widehat{\beta}_j}{\widehat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}} \sim t_{n-p}$$

  with $(X^TX)^{-1}_{jj}$ the $j-th$ diagonal element of the matrix $(X^TX)-1$.
- **`Pr(>|t|)`** : The $p-value$ of the previous Student tests.

➤ **`Signif. codes`** : Significance level symbols.

➤ **`Residual standard error`** : The value of $\widehat{\sigma}$ and the number of degrees of freedom : $(n-p)$ (here $31-5=26$). Here

$$\widehat{\sigma}^2 = 0.8172^2$$

➤ **`Multiple R-squared`** : The value of $R^2 = 0.9546$.

➤ **`Adjusted R-squared`** : The value of the adjusted R^2 : $R^2_a = 0.9476$.

➤ **`F-statistic`** : Here we test

$$H_0 : Y_i = \beta_0 + \varepsilon_i \quad \text{vs} \quad H_1 : Y_i = \beta_0 + \sum_{j=1}^{4} \beta_j X_{ij} + \varepsilon_i.$$

The `F-statistic` is the value of the Fisher's global test statistic F such that under H_O

$$F = \frac{\|P_X Y - \overline{Y}\mathbf{1}\|^2/(p-1)}{\widehat{\sigma}^2} \sim F_{(p-1,n-p)}.$$

In this example, $F = 136.5$ and the associated degrees of freedom $(p-1, n-p) = (4, 26)$. The $p-value < 2.2e-16$ is very small so we reject H_0, the test is meaningful.

4.4. Model validation

We recall that we assume the model~(1), under the Rank assumption and under **[P1]**–**[P4]** where

- **[P1]** : Errors are centered/(the model is linear) : $\forall i = 1, \cdots, n \quad \mathbb{E}_\beta[\varepsilon_i] = 0$.
- **[P2]** : Errors have homoscedastic variance : $\forall i = 1, \cdots, n \quad \mathbb{V}\text{ar}_\beta[\varepsilon_i] = \sigma^2 > 0$.
- **[P3]** : Errors are uncorrelated: $\forall i \neq j \quad \mathbb{C}\text{ov}(\varepsilon_i, \varepsilon_j)$.
- **[P4]** : Errors are gaussian : $\forall i = 1, \cdots, n \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

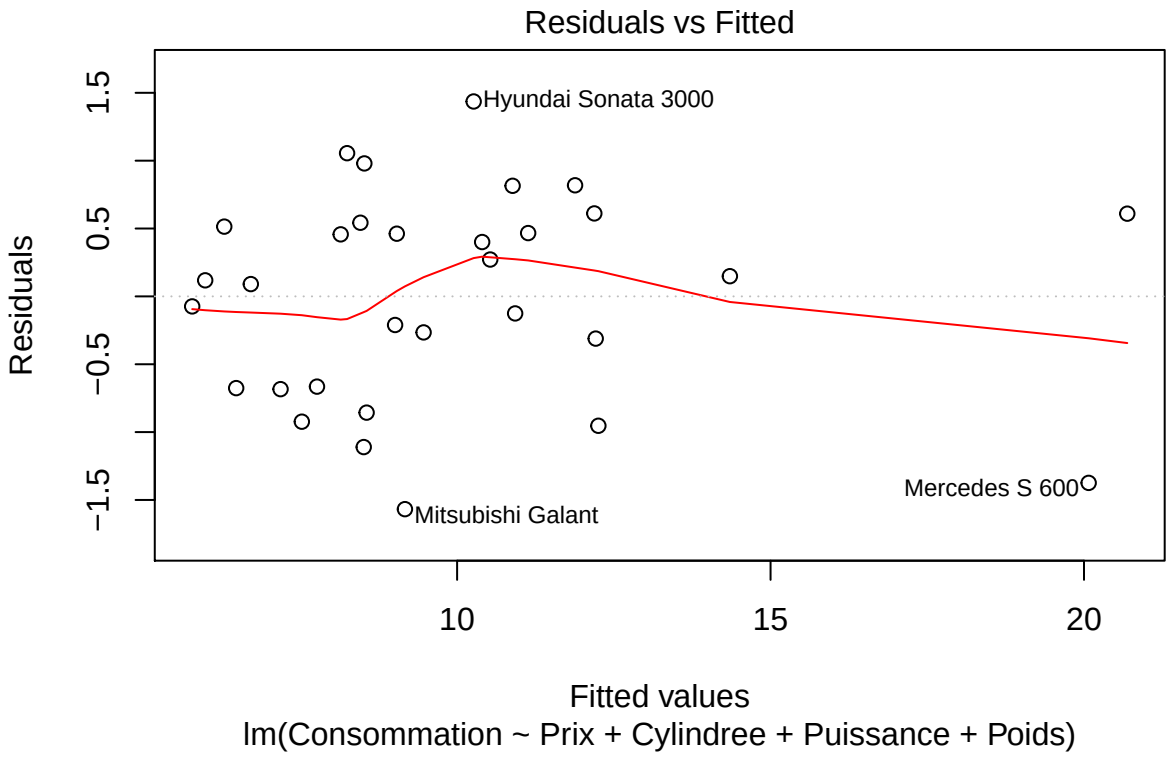
The rank hypothesis is easily verifiable with a simple calculation. For the other assumptions **[Pi]**, this requires an analysis of the residus. First upload the needed **R** library.

```
library(MASS)
library(carData)
library(car)
```

Validation of the postulates [P1]: Errors are centered

The centered postulat (the linearity assumption in practice) can be assesed by inspecting the *Residuals vs Fitted*-plot (or the *Studentized residuals*-plots). The command for the *Residuals vsFitted*-plot is `plot( ,which=1)`.

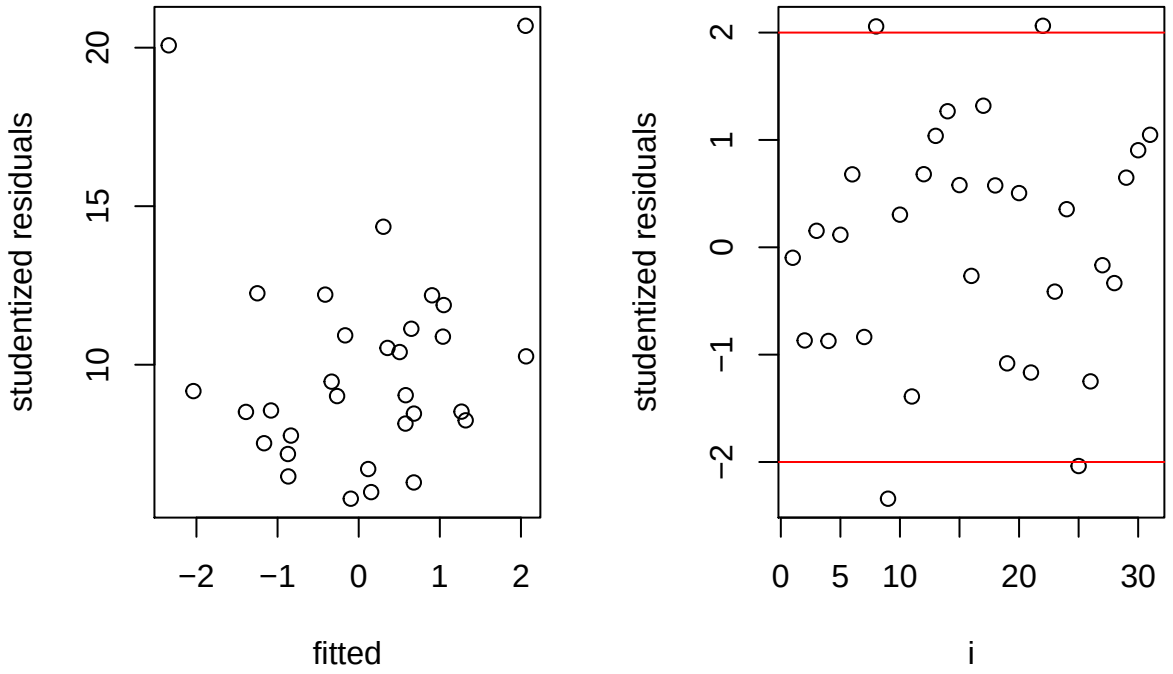
```
plot(reg,which=1)
```



The plot shows that when the responses predicted by the model (fitted values) increase, the residuals remain globally uniformly distributed on both sides of 0. The red line is approximately horizontal at zero.

The command `stdres()` displays the *Studentized residuals* $(t_i^*)_i$, needed to draw the *Studentized residuals*-plot.

```
par(mfrow=c(1,2))
SR=stdres(reg)
plot(SR,fitted(reg),xlab="fitted",ylab="studentized residuals")
plot(SR,xlab="i",ylab="studentized residuals")
abline(h=2,col="red")
abline(h=-2,col="red")
```

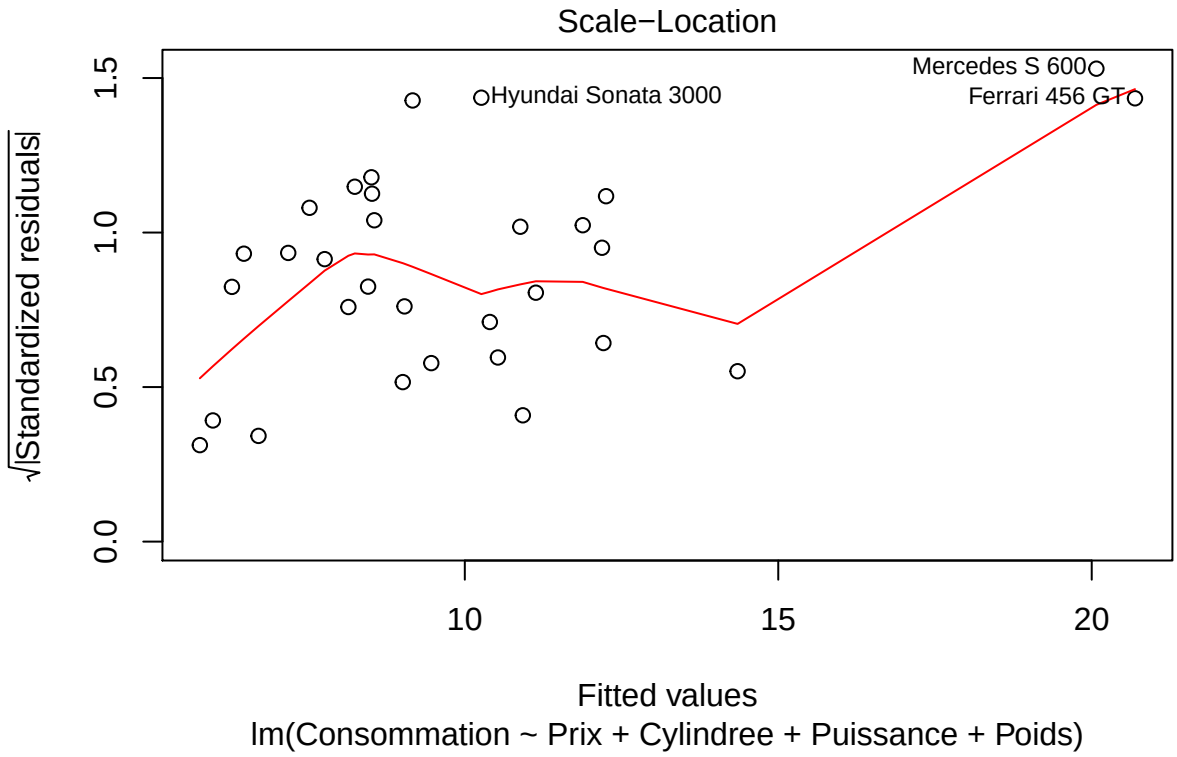


The plots show no fitted pattern, the residuals remain globally (reasonably) uniformly distributed. therefore that the assumption of linearity is acceptable. Thus, we validate the postulate.

Validation of the postulates [P2]: Errors have homoscedastic variance

The homoscedastic assumption can be checked by examining the *Scale-location*-plot (the command `plot( ,which=3)`). The postulate is validated if we see a horizontal line with equally spread points. In our example, it seems difficult to validate the postulat. So, let us make a Breush-Pagan test (H_O : homoscedasticity) to assess it.

```
plot(reg,which=3)
```



The command for the Breush-Pagan test is `ncvTest`. The homoscedasticity is rejected if the *p-value* is less than 0.05. Here, *p-value*= 0.38455 > 0.05, the posutlate is validated.

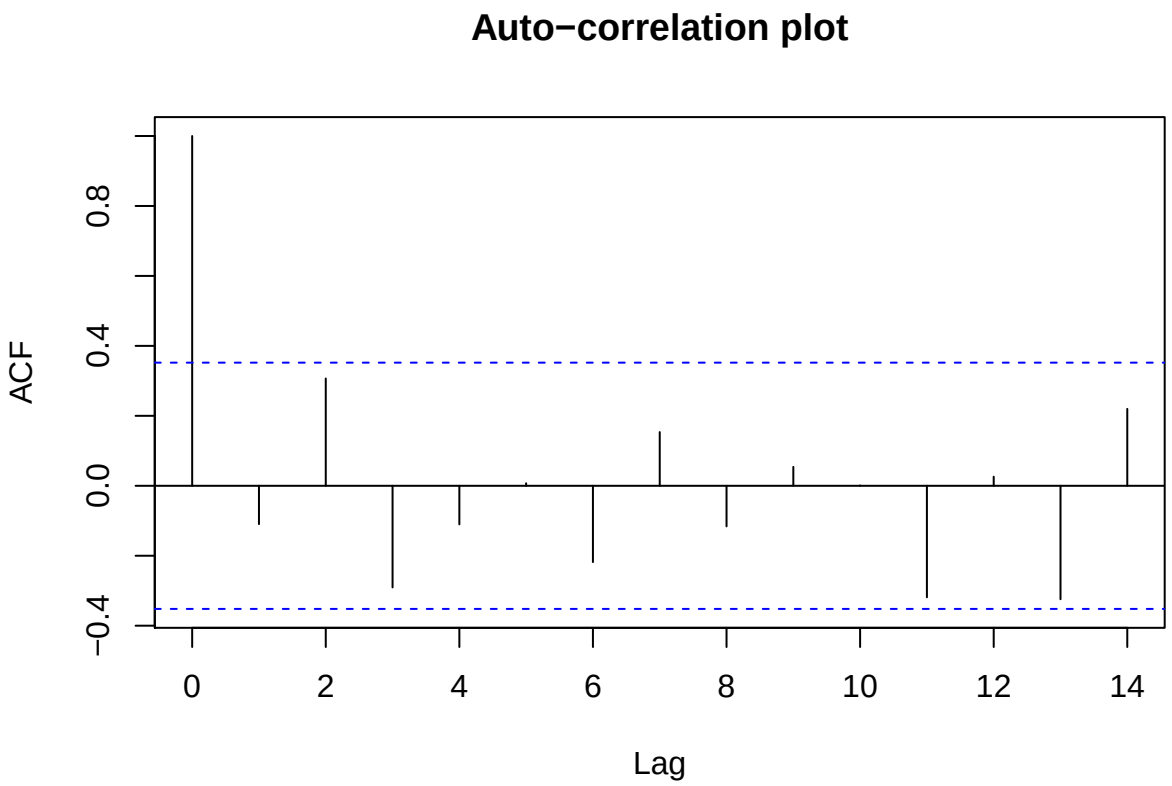
```
ncvTest(reg)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.7560996, Df = 1, p = 0.38455
```

Validation of the postulate [P3]: Errors are uncorrelated

Under **R**, we can represent the auto-correlation of the residuals using the command `acf()`. In our example, exept the first one, none exceeds dashed thresholds thus uncorrelation is satisfied.

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



The Durbin-Watson test can be also used to validate this assumption. The command is `durbinWatsonTest`. Under the null hypothesis the residual are considered auto-uncorrelated.

```
durbinWatsonTest(reg)
```

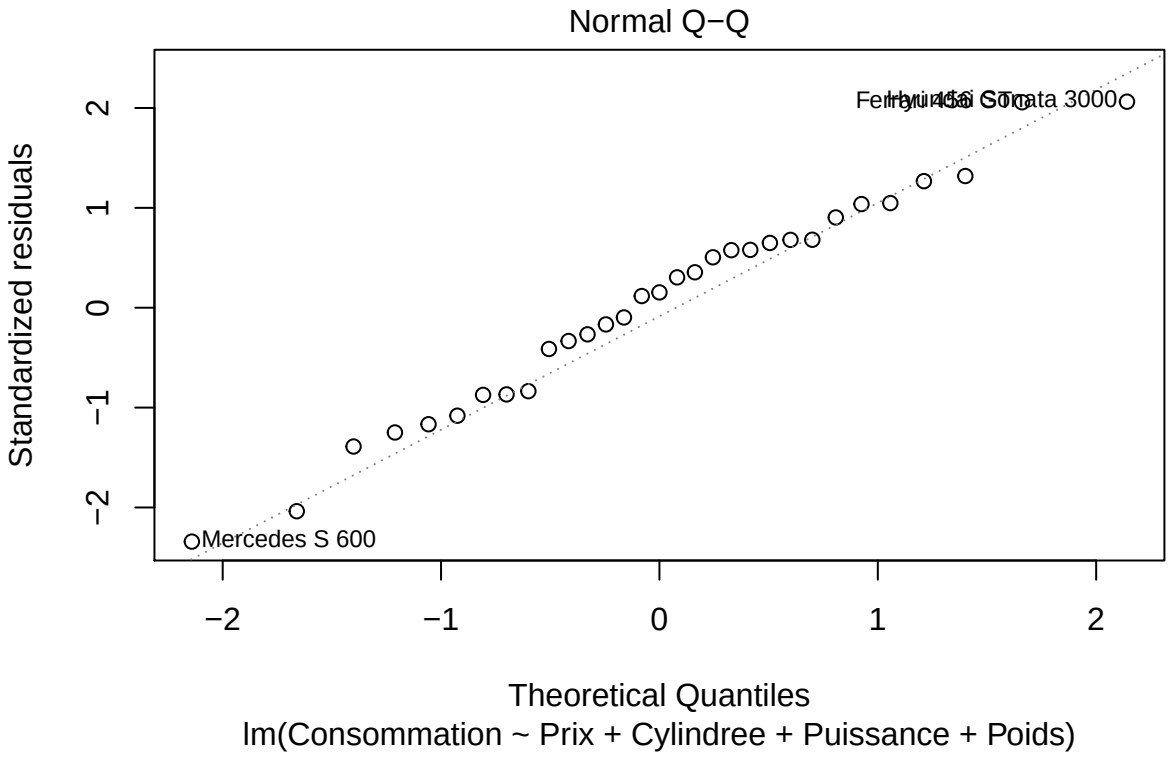
```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.1096954      2.180495   0.764
##  Alternative hypothesis: rho != 0
```

Here, the *p-value*= 0.782 > 0.05 thus we can't reject H_0, the posutlate is validated.

Validation of the postulate [P4]: Errors are gaussian

To analyze the normality, we use the Q-Q plot with the command `plot( ,which=2)`. The points appear reasonably aligned along the reference line even the sample size $n = 31$ is small, then the postulate is validated.

```
plot(reg,which=2)
```



The Shapiro-Wilk test can also be used to assess the normality of residuals. The normality assumption is rejected if the *p-value* is less than 0.05. Here, *p-value*> 0.05, the posutlate is unvalidated. But as the sample size is small, it was expected. However, we assume that the postulate is verified.

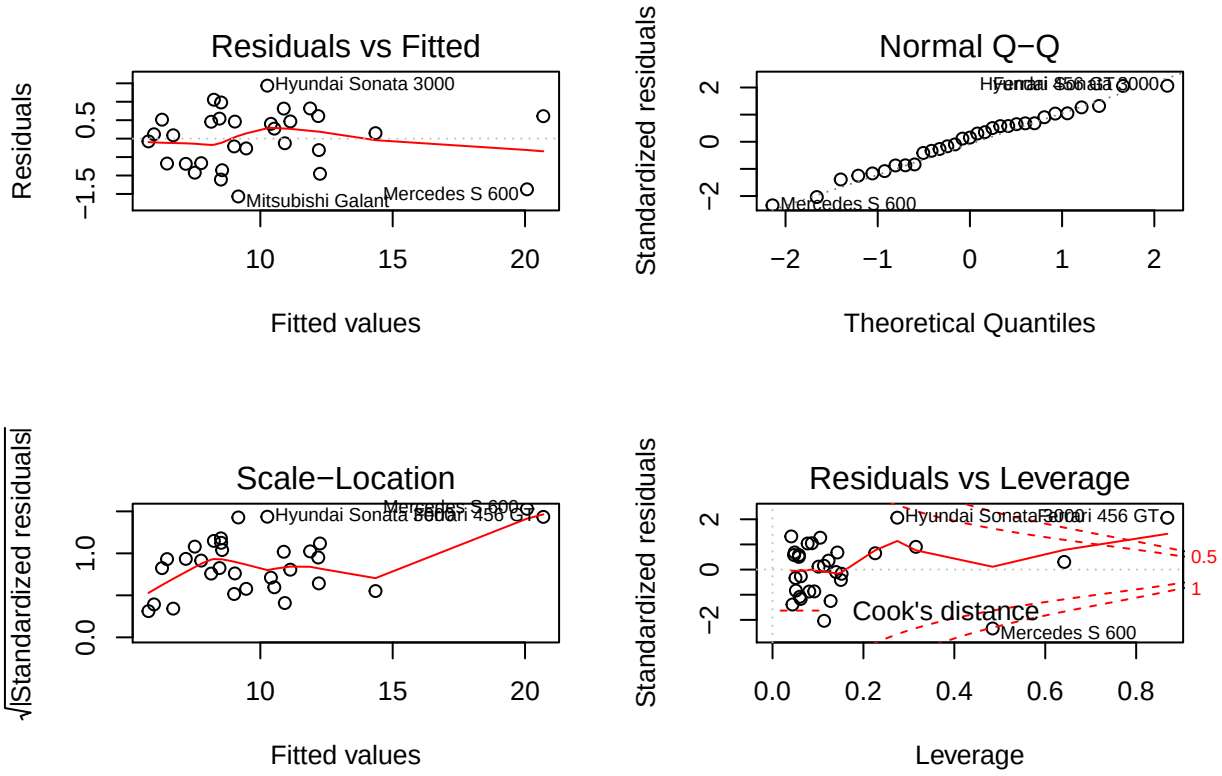
```
shapiro.test(residuals((reg)))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals((reg))
## W = 0.9709, p-value = 0.5442
```

Command `plot`: to resume

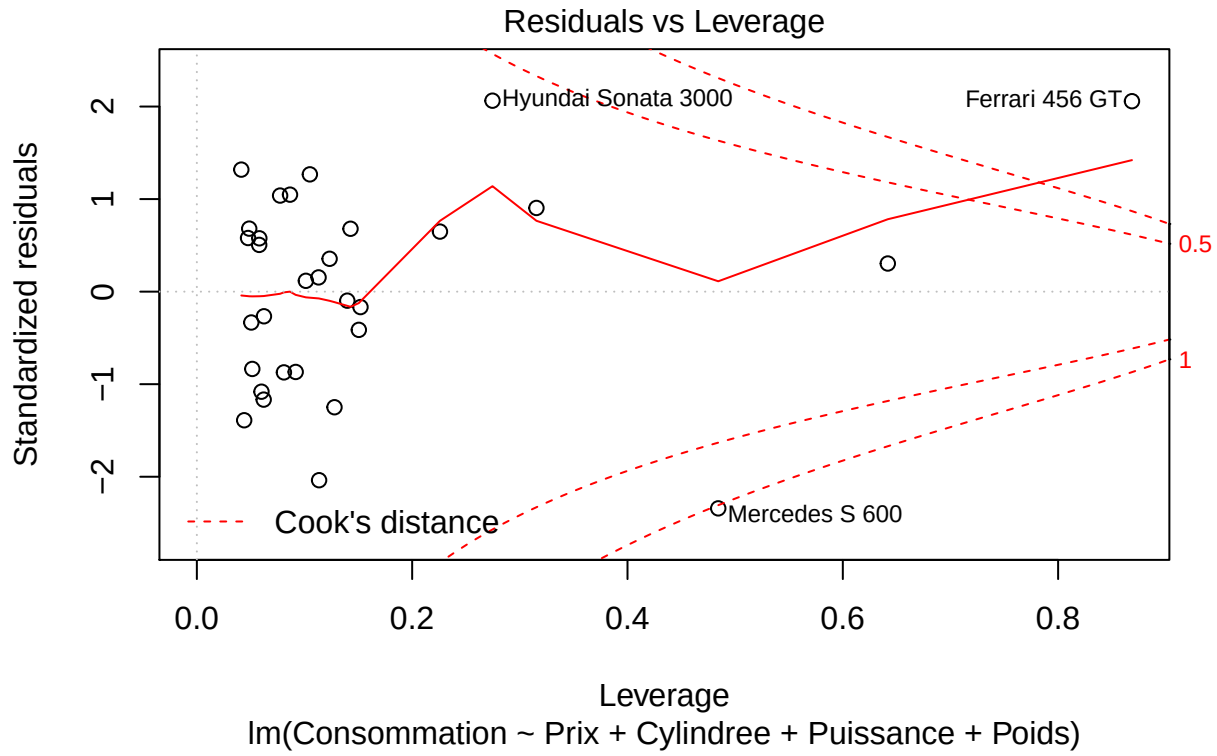
Note that by default, the command `plot)` gives 4 different plots in the linear regression setting. Three of them have been seen and used to validate the postulat. The last one is the `Residuals vs Leverage`-plot.

```
par(mfrow=c(2,2))
plot(reg)
```



The `Residuals vs Leverage`-plot can be called alone as follows :

```
plot(reg,which=5)
```

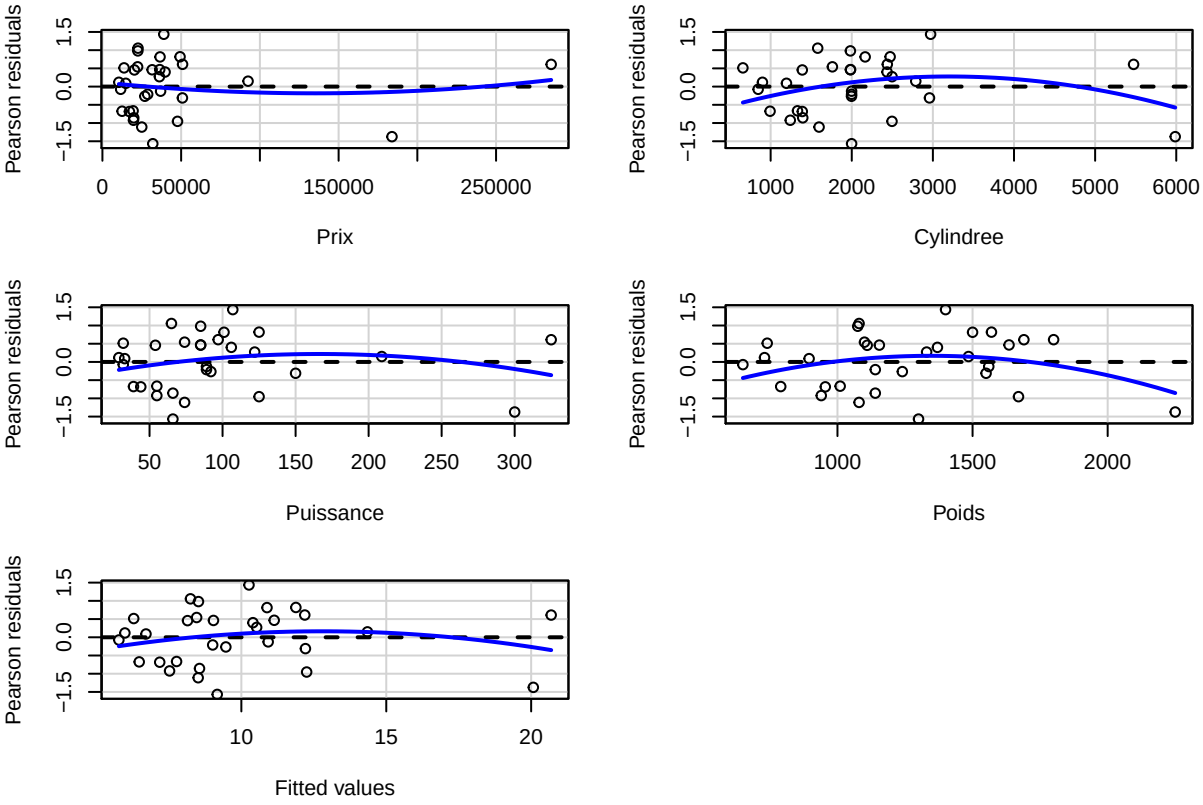


It appears in this plot, that 2 observations have a Cook's distance larger than 1. They are *ouliers* : *regression ouliers* or *leverage points* or both. A sudy of ouliers is done in a further section.

For going further...

For going further, to see the contribution of each variable, we can use the `residualPlots` function from the `cars` library. (it will be discussed in classroom)?

```
residualPlots(reg)
```




```
##             Test stat Pr(>|Test stat|)
## Prix           1.1523          0.26009
## Cylindree     -2.2748          0.03176 *
## Puissance     -2.4246          0.02289 *
## Poids         -1.6631          0.10878
## Tukey test    -2.1976          0.02798 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.5 Confidence Interval

The `confint` command easily displays confidence intervals for the parameters β_j of the model. They are based on a Student's law, if the postulate **[P4]** is satisfied. If not, these intervals are biased.

```
cbind(confint(reg),coef(reg))
```

```
##                      2.5 %       97.5 %
## (Intercept)  1.167851e+00 3.744737e+00  2.456294e+00
## Prix         2.474392e-06 3.836669e-05  2.042054e-05
## Cylindree   -1.682157e-03 6.809703e-04 -5.005933e-04
## Puissance    4.455929e-03 4.553302e-02  2.499448e-02
## Poids        2.354210e-03 5.966955e-03  4.160583e-03
```

The `predict( ,interval = "confidence")` command displays confidence interval for $x_i^T\beta$ (say, for *estimation*); while the `predict( ,interval = "prediction")` command displays confidence interval for Y_i (say, for *prediction*). Note that $x_i^T\beta$ and Y_i are both estimate by $x_i^T\widehat{\beta}$, but the bound of the confidence interval are differents.

```
ICconf = predict(reg, interval = "confidence", level = 0.95)
head(ICconf)
```

```
##                           fit      lwr      upr
## Daihatsu Cuore       5.773872 5.145857 6.401888
## Suzuki Swift 1.0 GLS 6.475902 5.966890 6.984914
## Fiat Panda Mambo L   5.981720 5.416875 6.546566
## VW Polo 1.4 60       7.183591 6.705853 7.661329
## Opel Corsa 1.2i Eco  6.709359 6.174759 7.243959
## Subaru Vivio 4WD     6.285932 5.651261 6.920603
```

```
ICpred= predict(reg, interval = "prediction", level = 0.95)
```

```
## Warning in predict.lm(reg, interval = "prediction", level = 0.95): predictions on c
```

```
head(ICpred)
```

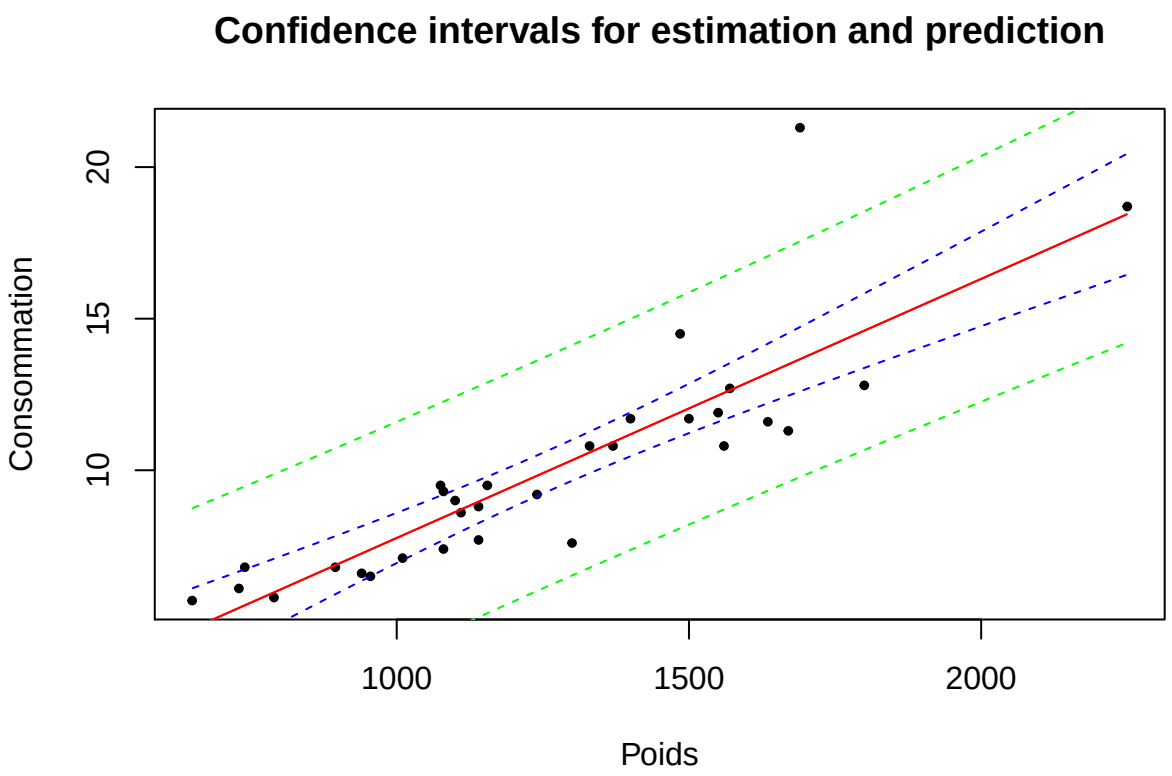
```
##                           fit      lwr      upr
## Daihatsu Cuore       5.773872 3.980461 7.567284
## Suzuki Swift 1.0 GLS 6.475902 4.720620 8.231184
## Fiat Panda Mambo L   5.981720 4.209442 7.753999
## VW Polo 1.4 60       7.183591 5.437121 8.930060
## Opel Corsa 1.2i Eco  6.709359 4.946486 8.472231
## Subaru Vivio 4WD     6.285932 4.490179 8.081685
```

Let us consider the simple regression of the variable `Consommation` on the predictor `Poids`, and consider new data `Poids` (called `newgrille` in our program). The command `predict.lm( ,interval = 'confidence')` displays prediction of the new $x_{i_{new}}^T\beta$ and the command `predict.lm( ,interval = 'prediction')` displays prediction of the new $Y_{i_{new}}$.

```
regP=lm(Consommation~Poids,data=conso_voit)
newgrille=data.frame(Poids=seq(min(conso_voit$Poids)+1,max(conso_voit$Poids)-1),2)
predicgrille=predict.lm(regP,newgrille, interval='confidence',level=0.95)
head(predicgrille)
```

```
##        fit      lwr      upr
## 1 4.783165 3.455416 6.110914
## 2 4.791711 3.465595 6.117828
## 3 4.800257 3.475773 6.124742
## 4 4.808804 3.485950 6.131658
## 5 4.817350 3.496126 6.138574
## 6 4.825897 3.506302 6.145491
```

```
plot(conso_voit$Poids,conso_voit$Consommation,ylab="Consommation",xlab="Poids",pch=20,
#abline(regP,col='red')
matlines(newgrille$Poids,predicgrille,lty=c(1,2,2),col=c("red","blue","blue"))
predicgrille=predict.lm(regP,newgrille, interval='prediction',level=0.95)
matlines(newgrille$Poids,predicgrille,lty=c(1,2,2),col=c("red","green","green"))
```

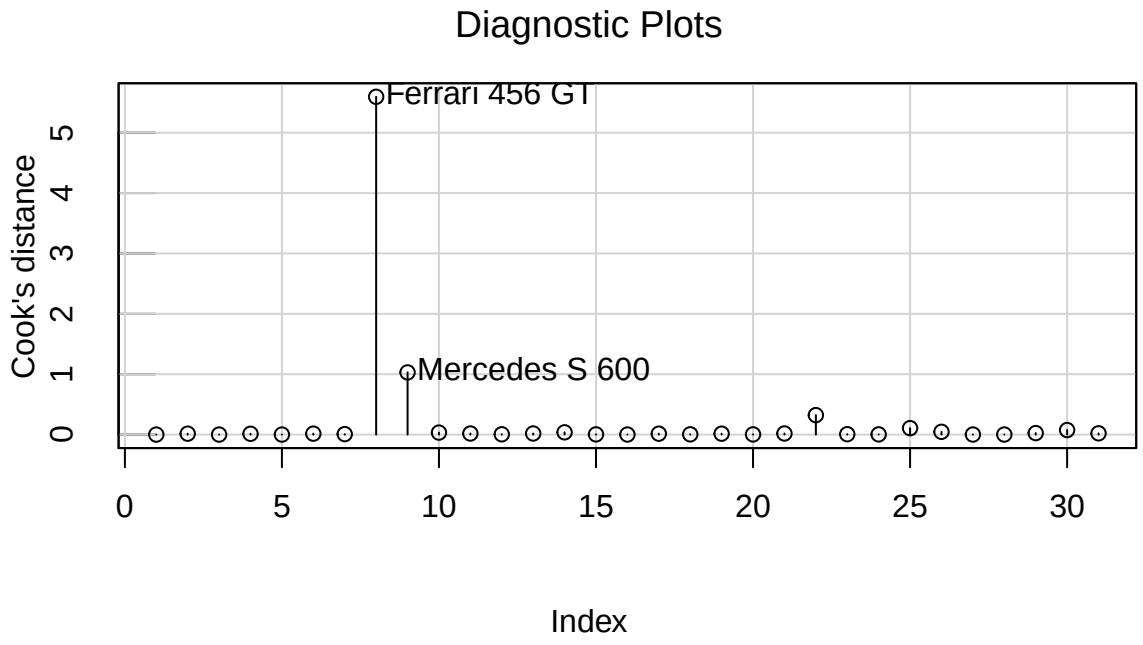


4.6. Outliers and leverage points

The library `car` offers an easy way to detect graphically atypical observations and to assess about their nature by tests. The command `influenceIndexPlot` is an important one.

Cook's distance plot

```
influenceIndexPlot(reg,vars="Cook")
```

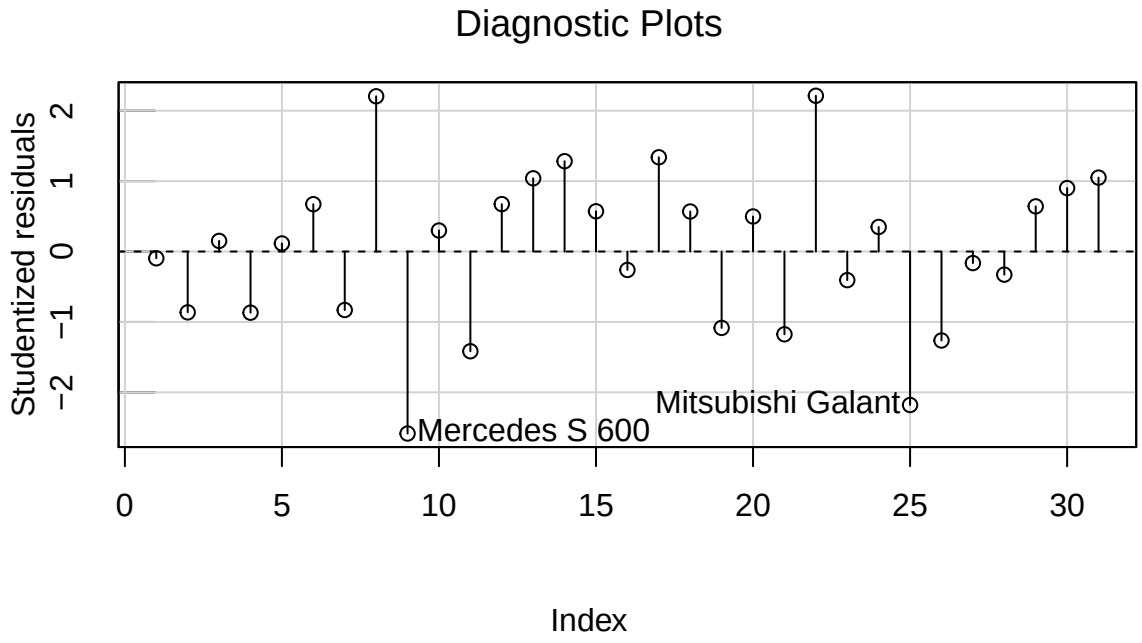


The Cook's distance plot highlight the influence of each observation on the estimation of the model (on β). As seen on the previous chapter, we compare the Cook's distance with 1. Here, two observations have a Cook's distance larger than 1 :

`Ferrari 456 GT` and `Mercedes S 600`

Studentized plot

```
influenceIndexPlot(reg,vars="Studentized")
```

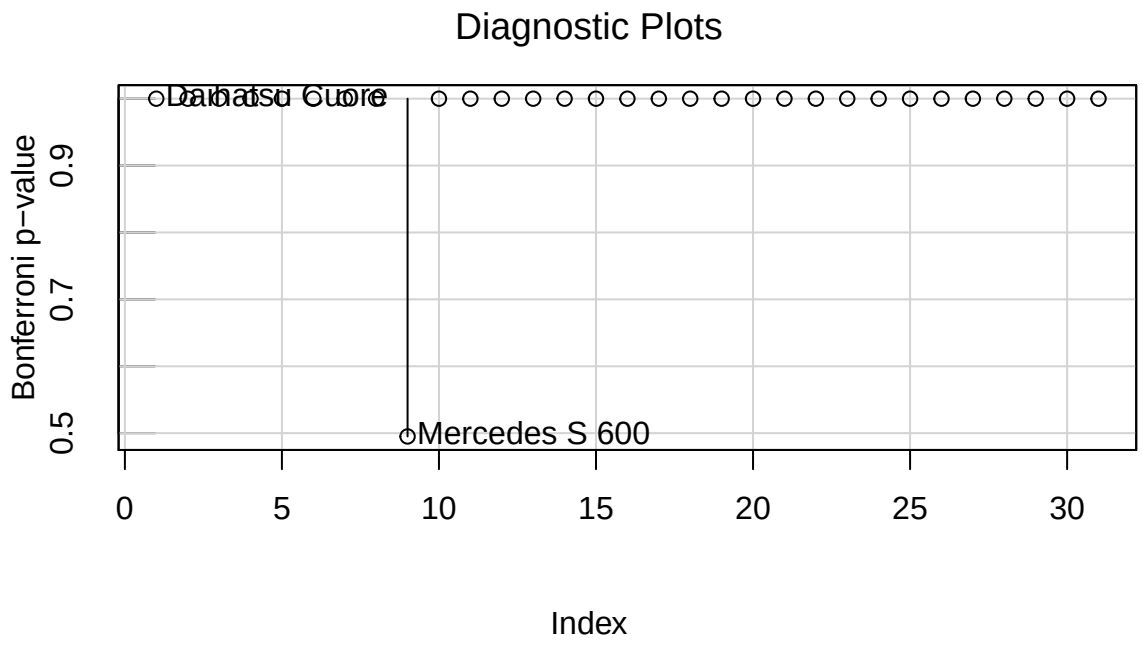


This plot, that of studentized residuals also makes it possible to highlight `outliers`. Here, two observations seems doubtful :

`Mitsubishi Galant` and `Mercedes S 600`

Bonferroni plot

```
influenceIndexPlot(reg,vars="Bonf")
```

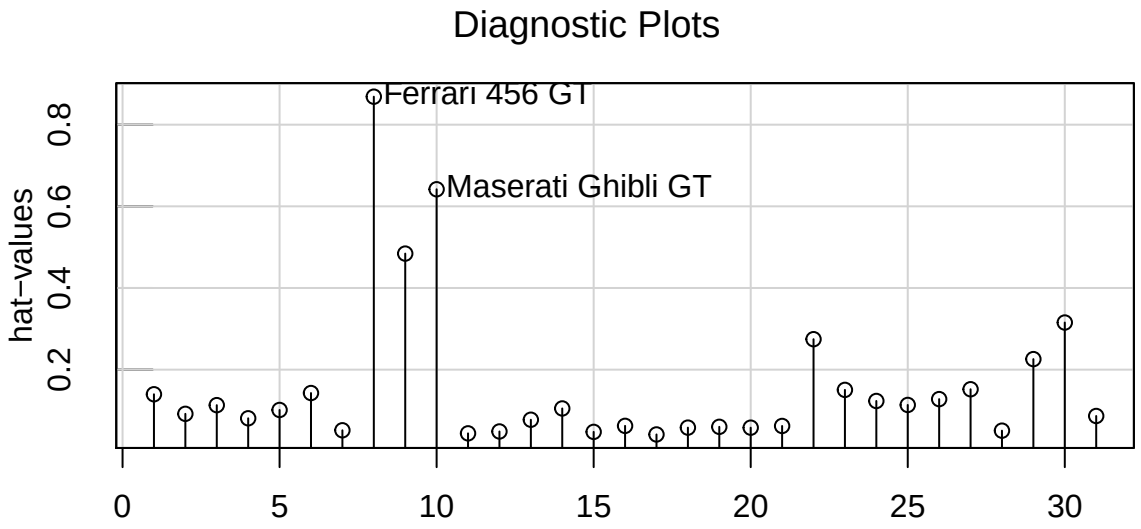


This is the plot of the *p-value* Bonferroni. Is considered as an outlier is a observation with a p-value less than 0.05. Here, the plot detetects one observation :

```
Mercedes S 600
```

Hat plot

```
influenceIndexPlot(reg,vars="hat")
```



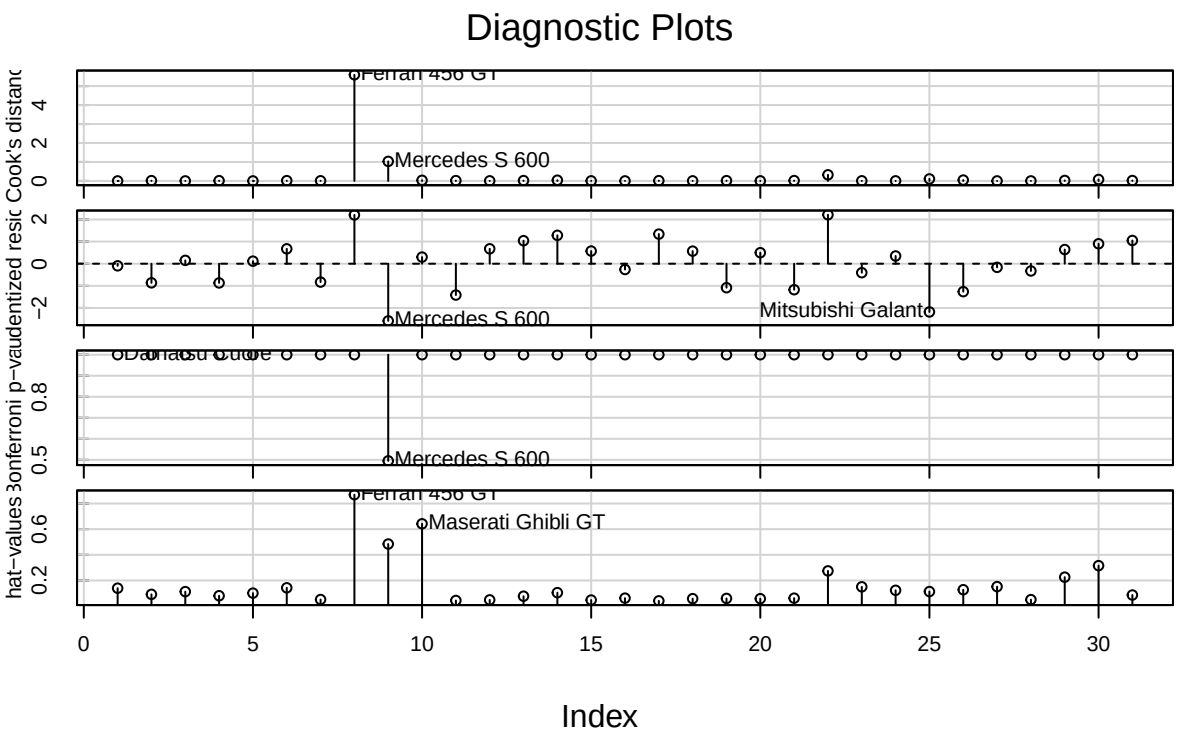
This plot called the *hat value*-plot, reflects the leverage (h_{ii}) of each observation on its own estimate. An observation is considered to be a *leverage point* when this value is less than 0.05. Here, the plot detetects two observations :

`Ferrari 456 GT` and `Maserati Ghibli GT`

The all plots

It's better to display these four graphs in parallel to have a better vision of the atypical points found in the various plot. It can be done as follows :

```
influenceIndexPlot(reg)
```



Here, the doubtful observations are :

`Mercedes S 600` and `Ferrari 456 GT`

We can access to Bonferroni's with the `outlierTest` command.

```
outlierTest(reg)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##                 rstudent unadjusted p-value Bonferroni p
## Mercedes S 600 -2.584781           0.01597      0.49506
```

The adjusted *p-value* by the Bonferonni method is equal to 0.49506 and is very far from the threshold of 0.05. The `Mercedes S 600` observation can not be considered as `outlier`.

To assess if the observations `Ferrari 456 GT` and `Mercedes S 600` really affect the estimation ouf our model (say β), we can compare the results of the estimation of β with and without these observation. We can done this with the command `comparCoefs`.

```
regbis = lm(Consommation~Prix + Cylindree + Puissance + Poids, data=
               conso_voit[-c(which(conso_voit_complet$Type=="Ferrari 456 GT"),
                             which(conso_voit_complet$Type=="Mercedes S 600")),])
compareCoefs(reg ,regbis)
```

```
## Calls:
## 1: lm(formula = Consommation ~ Prix + Cylindree + Puissance + Poids,
##   data = conso_voit)
## 2: lm(formula = Consommation ~ Prix + Cylindree + Puissance + Poids,
##   data = conso_voit[-c(which(conso_voit_complet$Type ==
##   "Ferrari 456 GT"), which(conso_voit_complet$Type == "Mercedes S 600")),
##    ])
##
##                Model 1   Model 2
## (Intercept)      2.456     2.071
## SE               0.627     0.664
##
## Prix          2.04e-05  2.56e-05
## SE            8.73e-06  3.03e-05
##
## Cylindree    -0.000501  0.000327
## SE            0.000575  0.000700
##
## Puissance      0.02499   0.01777
## SE             0.00999   0.01554
##
## Poids         0.004161  0.003598
## SE            0.000879  0.001054
##
```

It comes out that, that the observations `Ferrari 456 GT` and `Mercedes S 600` have little influence on the coefficients of the model parameters, as well as on their standard error, since the values do not vary much, even if they have a Cook's distance larger than 1.

```
#summary(regbis)
#plot(regbis)
```

Introduction to Regression - Chapter 5

MAP 535

Contents

Chapter 5 : Model selection

5.1. Introduction

The purpose of the regression is twofold: Explain and predict using estimation tools. In previous chapters, it has been assumed that the model

$$Y = X\beta + \varepsilon$$

is the "good" where $X = (X_1, \ldots, X_p)$.. In practice, nothing assures us that we have not forgotten variables. It is also possible that too many variables are used. If the goal is to explain, it seems justified to take the model having the largest R^2. If the goal is to estimate or predict, we will see that this is not necessarily the case. To do this, we use the mean squared error (MSE).

Definition 1 *Let $\theta \in \mathbb{R}^k$ be the parameter to be estimated and $\hat{\theta}$ an estimator of θ.* ***The mean squared error (MSE)*** *of $\hat{\theta}$ is given by:*

$$\mathbb{E}[||\hat{\theta} - \theta||^2] = \sum_{j=1}^{k} \mathbb{E}[(\hat{\theta}_j - \theta_j)^2].$$

Comment:

☛ The use of $||\cdot||^2$ is consistent with the idea of ordinary least squares estimation.

Proposition 1 *For all $\theta \in \mathbb{R}^p$:*

$$\mathbb{E}[||\hat{\theta} - \theta||^2] = \sum_{j=1}^{k} (\mathbb{V}ar(\hat{\theta}_j) + (\mathbb{E}[\hat{\theta}_j] - \theta_j)^2).$$

Proof : Obvious. □

To illustrate the purpose of this chapter, let's do some calculations for the following example. We assume the model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon = X\beta + \varepsilon, \tag{1}$$

where $X = [X_1 \ X_2]$ is a $n \times 2$ matrix of rank 2. Let $\beta = (\beta_1, \beta_2)^T \in \mathbb{R}^2$ and such that $\beta_2 \neq 0$. One may wonder if the X_2 varible is useful, and study the case where we would consider $\beta_2 = 0$ even if it is false, and look for when to omit an explanatory variable can be advantageous in terms of risk .

Let define the following model

$$Y = X_1\beta_1 + \varepsilon, \tag{2}$$

on which the OLSE is determined for the estimate of β_1, which is

$$\widetilde{\beta}_1 = (X_1^T X_1)^{-1} X_1^T Y,$$

where Y is defined by the model~(1), thus $\widetilde{\beta}_1$ is biased. Denote by $\hat{\beta}$, the OLSE of the estimation of β calculate from the model~(1). Thus, we have 2 estimators, one biased and the other one unbiased

$$\widetilde{\beta} = (\widetilde{\beta}_1, 0)^T \text{ and } \hat{\beta} = (X^T X)^{-1} X^T Y.$$

Proposition 2 *In the previous context,* $\forall \beta \in \mathbb{R}^p$

$$\mathbb{E}[||\hat{\beta} - \beta||^2] - \mathbb{E}[||\widetilde{\beta} - \beta||^2] \geq \sigma^2 \frac{||X_1||^2}{D} - \beta_2^2 \left(1 + \frac{(X_1^T X_2)^2}{||X_1||^4}\right),$$

where D denote the determinant of the matrix $(X^T X)^{-1}$.

Comment:

☛ This result does not contradict the Gauss-Markov theorem, because $\widetilde{\beta}$ is biased. By introducing (for $\beta_2 \neq 0$ and small enough) a slightly biased estimator with a lower variance, the quadratic risk is improved. For the estimation (and therefore the prediction), we must be wary of too rich models.

Proof : We easily prove that

$$(X^T X)^{-1} = \frac{1}{D} \begin{pmatrix} ||X_2||^2 & -X_1^T X_2 \\ -X_1^T X_2 & ||X_1||^2 \end{pmatrix},$$

where $D := ||X_1||^2 ||X_2||^2 - (X_1^T X_2)^2 > 0$.

• Moreover, the estimator $\hat{\beta}$ is unbiased, it comes

$$\mathbb{E}[(\hat{\beta} - \beta)^2] = \sum_{j=1}^{2} \mathbb{V}\text{ar}(\hat{\beta}_j) = \sigma^2 \text{Tr}((X^T X)^{-1})) = \frac{\sigma^2}{D} \left(||X_2||^2 + ||X_1||^2\right).$$

• For the estimator $\widetilde{\beta} = (\widetilde{\beta}_1, 0)^T$, we have

$$\begin{aligned}
\mathbb{E}[||\widetilde{\beta} - \beta||^2] &= \sum_{j=1}^{2} \mathbb{E}[(\widetilde{\beta}_j - \beta_j)^2] = \mathbb{E}[(\widetilde{\beta}_1 - \beta_1)^2] + \beta_2^2 = \mathbb{E}[((X_1^T X_1)^{-1} X_1^T Y - \beta_1)^2] + \beta_2^2 \\
&= \mathbb{E}[((X_1^T X_1)^{-1} X_1^T (\beta_1 X_1 + \beta_2 X_2 + \varepsilon)\beta_1)^2] + \beta_2^2 = \left((X_1^T X_1)^{-1} X_1^T X_2\right)^2 \beta_2^2 + \sigma^2 (X_1^T X_1)^{-1} + \beta_2^2 \\
&= \frac{\sigma^2}{||X_1||^2} + \beta_2^2 \left(1 + \frac{(X_1^T X_2)^2}{||X_1||^4}\right).
\end{aligned}$$

For $D > 0$, it comes that $D < \|X_1\|^2\|X_2\|^2$. Therefore, we get

$$\begin{aligned}\mathbb{E}[(\hat{\beta}-\beta)^2] - \mathbb{E}[\|\widetilde{\beta}-\beta\|^2] &= \frac{\sigma^2}{D}\left(\|X_2\|^2 + \|X_1\|^2\right) - \frac{\sigma^2}{\|X_1\|^2} - \beta_2^2\left(1 + \frac{(X_1^T X_2)^2}{\|X_1\|^4}\right) \\ &> \frac{\sigma^2\|X_1\|^2}{D} - \beta_2^2\left(1 + \frac{(X_1^T X_2)^2}{\|X_1\|^4}\right).\end{aligned}$$

□

In the following, we will be interested in methods to choose a set of variables (which we will call **model**). While it may be easy to decide between two models, the question of model choice is more delicat. Indeed,

- There is no natural order between the variables.
- There are many possible models. For example, if there are 8 possible variables in addition to the vector $\mathbb{1}_n$ (always take the intercept), then we have $\sum_{j=0}^{8} C_j^8 = 2^8 = 256$ possible models to compare.

More specifically, we will focus on methods that rely on the following tools/criterions:

- Tests between nested models
- R^2
- R_a^2 adjusted
- C_p of Mallows
- AIC- criterion
- BIC- criterion

5.2. Notations and illustrative example

We note $p = q + 1$ the number of explanatory variables (the intercept $\mathbb{1}_n$ included). We define

$$X = (\mathbb{1}_n, X_1, \ldots, X_q).$$

We denote by $[m]$ any model of size m, *i.e.* $m := \text{card}([m])$. We consider the framework of linear regression models.

$$Y = X\beta + \varepsilon,$$

where $\text{rang}(X) = p, \quad \mathbb{E}[\varepsilon] = 0, \quad \mathbb{V}\text{ar}(\varepsilon) = \sigma^2 I.$

We define for all model $[m]$

$$RSS(m) = \|Y - P_m Y\|^2,$$

where P_m is the matrix of orthogonal projection into the space generated by thethe variables of $[m]$.

For $1 \leq m_0 \leq p-1$, we define $[m_0]$ a model composed by m_0 variables (the intercept $\mathbb{1}_n$ is considered to be in the model). Let $[m_1]$ be a model with $m_1 = m_0 + 1$ variables such that

$$[m_1] = [m_0] \cup \{\text{one more variable} \notin [m_0]\}.$$

Now, let's describe various criterions for choosing betxeen these two nested models $[m_0]$ and $[m_1]$ in view of the data.

5.3. criterions

5.3.1. Fisher test for nested models.

As part of this approach, two different test statistics are given:

$$\begin{aligned} F &= \frac{RSS(m_0) - RSS(m_1)}{RSS(m_1)} \times (n - m_0 - 1), \\ \widetilde{F} &= \frac{RSS(m_0) - RSS(m_1)}{\hat{\sigma}^2} = \frac{RSS(m_0) - RSS(m_1)}{RSS} \times (n - p) \end{aligned}$$

Theorem 1 *We assume Y to be gaussian. Let $\alpha \in]0, 1[$. The statistics F and $\widetilde{F}$ allow us to test*

$$H_0 : \text{"the model is } [m_0]\text{"} \quad cvs \quad H_1 : \text{"the model is} [m_1]\text{"}$$

Indeed, :

- *If $F > f_{1,n-m_0-1,1-\alpha}$, then the model $[m_1]$ must be chosen at a level of risk α.*
- *If $\widetilde{F} > f_{1,n-p,1-\alpha}$, then the model $[m_1]$ must be chosen at a level of risk α.*

Comment:

☛ Note that it is difficult to compare these two results. The previous theorem is only valid under the condition $[m_0] \subset [m_1]$.

Proof : We recall that

$$RSS = \|Y - P_X Y\|^2 = (n - p)\hat{\sigma}^2.$$

The Pythagore theorem gives by projecting $Y - P_{m_0} Y$ into $[m_1]$:

$$\|Y - P_{m_0} Y\|^2 = \|P_{m_1} Y - P_{m_0} Y\|^2 + \|Y - P_{m_1} Y\|^2,$$

which is equivalent to :

$$RSS(m_0) = \|P_{m_1} Y - P_{m_0} Y\|^2 + RSS(m_1).$$

Then, we deduce that

$$\begin{aligned} F &= \frac{\|P_{m_1}Y - P_{m_0}Y\|^2/1}{\|Y - P_{m_1}Y\|^2/(n - m_0 - 1)}, \\ \widetilde{F} &= \frac{\|P_{m_1}Y - P_{m_0}Y\|^2/1}{\|Y - P_X Y\|^2/(n-p)}. \end{aligned}$$

Using the quantiles of Fisher's law, Theorem~3 and the theorem~6, we obtain the result. □

5.3.2. The determination coefficient R^2

It is recalled that, in general, the definition of the coefficient of determination R^2 is

$$R^2 = \frac{\|\widehat{Y} - \overline{Y}\mathbb{1}_n\|^2}{\|Y - \overline{Y}\mathbb{1}_n\|^2},$$

with $\mathbb{1}_n = (1, 1, \ldots, 1)^T$, $\overline{Y} = n^{-1}\sum_{i=1}^n Y_i$ and $\widehat{Y} = X\widehat{\beta}$. Also, for any model $[m]$, we define

$$R^2(m) = \frac{\|P_m Y - \overline{Y}\mathbb{1}_n\|^2}{\|Y - \overline{Y}\mathbb{1}_n\|^2} = 1 - \frac{\|Y - P_m Y\|^2}{\|Y - \overline{Y}\mathbb{1}_n\|^2}.$$

Denoting

$$TSS = \|Y - \overline{Y}\mathbb{1}_n\|^2, \quad RSS(m) = \|Y - P_m Y\|^2,$$

we have :

$$R^2(m) = 1 - \frac{RSS(m)}{TSS}.$$

We deduce the following property:

Proposition 3 *We have :*

$$R^2(m_1) - R^2(m_0) = \frac{RSS(m_0) - RSS(m_1)}{TSS} \geq 0.$$

Proof : Equality is obvious, inequality follows from Pythagore's theorem. □

Comment:

- ☛ In general, we do not use the R^2 as a selection criterion because it will always increase with the number of variables. But it is an indicative criterion when it remains constant and the number of variables is increased. It is also used to compare two models with the same number of variables

5.3.3. The adjusted determination coefficient R_a^2

It is recalled that, in general, the definition of the adjusted R^2 coefficient of determination is

$$R_a^2 = 1 - \frac{(n-1)\|Y - \widehat{Y}\|^2}{(n-p)\|Y - \overline{Y}\mathbb{1}_n\|^2} = 1 - \frac{(n-1)(1-R^2)}{n-p}.$$

We set, for all model $[m]$,

$$R_a^2(m) = 1 - \frac{(n-1)(1-R^2(m))}{n-m} = 1 - \frac{RSS(m)}{n-m} \times \frac{(n-1)}{TSS}.$$

The $R_a^2(m)$ is therefore a function of the sum of residual squares divided by the number of degrees of freedom. Note that if m increases then $RSS(m)$ decreases and $n-m$ decreases. This helps to correct the disadvantages of the R^2 coefficient.

5.3.4. The C_p of Mallows

For all model $[m]$, we denote $\widehat{Y}_m = P_m Y$. It is recalled that $RSS(m) = \|P_m Y - Y\|^2$ and

$$RSS(m) = \|P_m Y - Y\|^2 \neq \|P_m Y - X\beta\|^2.$$

Definition 2 *Let* $[m]$ *be any model. The Mallows criterion associated with* $[m]$ *is defined by:*

$$C_p(m) = \frac{RSS(m)}{\widehat{\sigma}^2} - n + 2m.$$

We can show that

(a) $\mathbb{E}[RSS(m)] = \mathbb{E}[\|\widehat{Y}_m - Y\|^2] = \|(\mathbb{I} - P_m)X\beta\|^2 + (n-m)\sigma^2$.

(b) $\mathbb{E}[\|\widehat{Y}_m - X\beta\|^2] = \|(\mathbb{I} - P_m)X\beta\|^2 + m\sigma^2$.

(c) $\mathbb{E}[C_p(m)\widehat{\sigma}^2] = \mathbb{E}[\|\widehat{Y}_m - X\beta\|^2]$.

Proof : Will be proved in lecture class.

Comments:

- **Unbiased estimator of the mean quadratic error:** We deduce from (c) that $C_p(m)\widehat{\sigma}^2$ is an unbiased estimator of the unknown mean quadratic prediction error $\mathbb{E}[\|\widehat{Y}_m - X\beta\|^2]$.

☛ **Minimisation of the criterion:** For any model $[m]$, the mean squared error of $\widehat{Y}_m$ is $\mathbb{E}[\|\widehat{Y}_m - X\beta\|^2]$. Ideally, it is a good criterion for estimating the estimator $\widehat{Y}_m$. Selecting a good $[m]$ model is like minimizing

$$m \longmapsto \mathbb{E}[\|\widehat{Y}_m - X\beta\|^2].$$

Unfortunately, this quantity depends on the unknown parameter β. We have at our disposal an unbiased estimator of this quantity. We could then minimize

$$m \longmapsto C_p(m)\widehat{\sigma}^2.$$

Since $\widehat{\sigma}^2$ does not depend on the model, it is natural, especially when trying to estimate $X\beta$ to minimize

$$m \longmapsto C_p(m).$$

Discussion around the criterion:

➷ **A penalized criterion:** We defined the C_p of Mallows criterion as follows

$$C_p(m)\widehat{\sigma}^2 = RSS(m) + 2m\widehat{\sigma}^2 - n\widehat{\sigma}^2 := RSS(m) + \text{pen}(m).$$

When studying the classic R^2, it appeared that the more variables were added, the more the RSS decreased:

$$m \text{ increases} \Rightarrow RSS(m) \text{ decreases.}$$

Adding a penalty $\text{pen}(m) := 2m\widehat{\sigma}^2$ to the $RSS(m)$ in the criterion is an alternative way to the adjusted R^2 to counterbalance this effect

$$m \text{ increases} \Rightarrow \text{pen}(m) \text{ increases.}$$

We say that we **penalize the big models**.

➷ **Adding useless variables to the real model** Set that the "real" model denoted by $[m^*]$ is included in the model $[m_0]$, then

$$X\beta = P_{m_0}X\beta \Leftrightarrow X\beta - P_{m_0}X\beta = 0_n.$$

The equation (a) then becomes : $\mathbb{E}[RSS(m_0)] = \mathbb{E}[\|\widehat{Y}_{m_0} - Y\|^2] = (n - m_0)\sigma^2$ and we have

$$RSS(m_0) \approx (n - m_0)\widehat{\sigma}^2$$

Equations (b) and (c) give then: $\mathbb{E}[C_p(m_0)\widehat{\sigma}^2] = \mathbb{E}[\|\widehat{Y}_{m_0} - X\beta\|^2] = m_0\widehat{\sigma}^2$. Therefore,

$$C_p(m_0) \approx m_0.$$

Thus, if we add useless variables (increases m_0) to the true model (included in $[m_0]$), then $RSS(m_0) \approx (n - m_0)\sigma^2$ will not significantly decrease compared to the $C_p(m_0) \approx m_0$ which will increase more significantly.

➴ **Forgetting important variables to the real model**

If the "real" model $[m^*]$ is not fully included in $[m_0]$ then

$$X\beta \neq P_{m_0}X\beta \Leftrightarrow X\beta - P_{m_0}X\beta = C.$$

So with the same reasoning as before we have:

(a) $\mathbb{E}[RSS(m_0)] = \mathbb{E}[\|\widehat{Y}_{m_0} - Y\|^2] = C + (n - m_0)\sigma^2$.

(b) $\mathbb{E}[\|\widehat{Y}_{m_0} - X\beta\|^2] = C + m_0\sigma^2$.

(c) $\mathbb{E}[C_p(m_0)\widehat{\sigma}^2] = \mathbb{E}[\|\widehat{Y}_{m_0} - X\beta\|^2]$.

We have then

$$RSS(m_0) \approx (n - m_0)\widehat{\sigma}^2 + C \quad \text{et } C_p(m_0) \approx m_0 + C$$

where $C > 0$. In this case, $C_p(m_0) > m_0$.

➴ **To resume**

- If we add useless variables to the "real" model, then $C_p(m_0) \approx m_0$.
- If we forget important variables to the "real" model, then $C_p(m_0) \approx m_0 + C$. where $C > 0$.

So if beyond the problem of estimating $X\beta$, we are interested, by the detection of the good variables, we will be interested in models $[m_0]$ such that $C_p(m_0) \leq m_0$.

Important Comment:

☛ It should be noted that the previous interpretations are only true if the choice of the model (selection of the optimal $[m]$) is independent of the data (computation of $\widehat{Y}_m = P_m Y$), so we must cut the sample in 2:

- A sample for the learning to compute $\widehat{Y}_m$ for all $[m]$.
- Another sample for validation to select $[m_{optimal}]$.

5.3.5. AIC/BIC criterion

Consider a linear regression model

$$Y = X\beta + \varepsilon,$$

where

$$\text{rank}(X) = p, \quad \mathbb{E}[\varepsilon] = 0, \quad \mathbb{V}\text{ar}(\varepsilon) = \sigma^2 I \quad \text{and } \varepsilon \sim \mathcal{N}(0, \sigma^2 I).$$

The likelihood of the model is :

$$L(Y,\beta) = -\frac{1}{2\sigma^2}\|Y - X\beta\|^2 - \frac{n}{2}\log(2\pi) - \frac{}{2}\log(\sigma^2)$$

Let $\widehat{\beta} = (X^TX)^{-1}X^TY$ be the OLSE which is in our gaussian setting the ordinary maximum likelihood estimator (OMLE). Then, by définition the OMLE $\widehat{\beta}$ maximize the likelihood. In others words, the maximized likelihood (ML) is

$$L(Y,\widehat{\beta}).$$

Proposition 4 *The model* $[m]$ *that maximizes the maximized likelihood (ML) on m is the model that minimizes*

$$m \longmapsto RSS(m).$$

Proof : Will be proved in Lecture class. □

Comments:

- ☛ We have seen that minimizing the RSS is not necessarily the best thing to do because it amounts to taking the largest model ($p = n = m$).
- ☛ As for the C_p of Mallows, we want to add a (positive) penalty to penalize the big models.

For the AIC criterion, the penalty is simply $\text{pen}(m) = m$ and for the BIC criterion, the penalty is $\text{pen}(m) = \frac{\log n}{2} \times m$.

Definition 3

• *The AIC of a model* $[m]$ *is defined by*

$$AIC(m) = \frac{n}{2}\log(RSS(m)) + m.$$

• *The BIC of a model* $[m]$ *is defined by*

$$BIC(m) = n\log(RSS(m)) + \log(n) \times m.$$

Comments:

- ☛ We choose the model $[m]$ that minimizes

$$m \longmapsto AIC(m) \text{ or } m \longmapsto BIC(m)$$

- ☛ If $n > 7$ ($\Rightarrow \log(n) > 2$) then the BIC will tend to select models smaller than those selected by AIC.

5.4. Comparaison of criterions

The purpose of this section is to compare the criterions in the previous section. For that we will consider two nested models $[m_0] \subset [m_1]$ such that $m_1 = m_0 + 1$. We set

$$H_0 : \text{ the model is } [m_0] \quad vs \quad H_1 : \text{ the model is } [m_1]$$

We study the cases where $[m_0]$ is chosen at the expense of $[m_1]$, *i.e* we look for a test statistic

$$T \leq q$$

where $q = C_\alpha > 0$ is a constant which depends of the level $\alpha \in (0, 1)$ of the test. In this section, we consider $\alpha = 0.05$ and $n - m_0 - 1 \geq 16$.

Fisher Test :

$$T := F = \frac{RSS(m_0) - RSS(m_1)}{RSS(m_1)} \times (n - m_0 - 1) \leq 4$$

The R^2 coefficient :

In our setting, $[m_1]$ is always chosen at the expense of $[m_0]$.

The adjusted R^2 coefficient :

$$\begin{aligned} R_a^2(m_0) \geq R_a^2(m_1) \quad &\Longleftrightarrow \quad \frac{RSS(m_0)}{n - m_0} \leq \frac{RSS(m_1)}{n - m_0 - 1} \\ &\Longleftrightarrow \quad T := \frac{RSS(m_0) - RSS(m_1)}{RSS(m_1)} \times (n - m_0 - 1) \leq 1. \end{aligned}$$

The C_p of Mallows :

$$\begin{aligned} C_p(m_0) \leq C_p(m_1) \quad &\Longleftrightarrow \quad \frac{RSS(m_0)}{\widehat{\sigma}^2} \leq \frac{RSS(m_1)}{\widehat{\sigma}^2} + 2 \\ &\Longleftrightarrow \quad \frac{RSS(m_0) - RSS(m_1)}{\widehat{\sigma}^2} \leq 2. \end{aligned}$$

If $\widehat{\sigma}^2$ is replaced by $RSS(m_1)/(n - m_0 - 1)$, then the following condition appears

$$T := \frac{RSS(m_0) - RSS(m_1)}{RSS(m_1)} \times (n - m_0 - 1) \leq 2.$$

AIC and BIC criterions:

In this case, we minimize the function

$$C : m \longmapsto \log(RSS(m)) + f(n)m$$

where for AIC criterion $f(n) = 2/n$ and for BIC criterion $f(n) = \log(n)/n$. Then, we have

$$\begin{aligned} C(m_0) \leq C(m_1) &\iff \log(RSS(m_0)) - \log(RSS(m_1)) \leq f(n) \\ &\iff \frac{RSS(m_0) - RSS(m_1)}{RSS(m_1)} \times (n - m_0 - 1) \leq (e^{f(n)} - 1) \times (n - m_0 - 1). \end{aligned}$$

Asymptotically, when $n \to +\infty$,

- **For AIC**

$$C(m_0) \leq C(m_1) \iff T := \frac{RSS(m_0) - RSS(m_1)}{RSS(m_1)} \times (n - m_0 - 1) \leq \frac{2}{n} \times (n - m_0 - 1).$$

- **For BIC**

$$C(m_0) \leq C(m_1) \iff T := \frac{RSS(m_0) - RSS(m_1)}{RSS(m_1)} \times (n - m_0 - 1) \leq \frac{\log n}{n} \times (n - m_0 - 1).$$

To resume

For each of the 6 criteria, we have roughly reduced ourselves to the study of

$$T := \frac{RSS(m_0) - RSS(m_1)}{RSS(m_1)} \times (n - m_0 - 1) \leq q$$

avec

- $q = 4$ for the Fisher test.
- $q = -\infty$ for the R^2 coefficient.
- $q = 1$ for the adjusted R^2 coefficient.
- $q = 2$ for the C_p of Mallows.
- $q = \frac{2}{n} \times (n - m_0 - 1)$ for the AIC.
- $q = \frac{\log n}{n} \times (n - m_0 - 1)$ for the BIC.

Comments:

☛ This roughly orders each of the criteria: the most favorable to $[m_0]$ is the BIC criterion, the most favorable to $[m_1]$ is the R^2 coefficient. We must be wary of these comparisons because they still depend on the value of n, of $\widehat{\sigma}^2$,...

☛ It should be remembered that for the Fisher test, the criterion for two models can only be compared if one model contains the other (nested models).

5.5. Step-by-step method

We have seen in the introduction that the minimization of criteria can be a delicate task when the number of explanatory variables is high. Indeed, if we have p variables (whose constant vector $\mathbb{1}_n$, the intercept), we have 2^{p-1} different models (all containing $\mathbb{1}_n$). When exhaustive search is not possible (either because we want to use Fisher's test, or because p is too big), we can use a step-by-step method combined with one of the 6 criterions previously studied. The disadvantage is that it does not test all possible combinations. We are therefore not sure of obtaining a global minimum. The three famous step-by-step methods are the following:

- **Forward selection:** We start with the model resume to the intercept $\mathbb{1}_n$. At each step, a regressor/variable is added to the model, the one with the best contribution (*i.e.* the ones which improves the chosen criterion). We stop when the criterion can not be improved by adding a new regressor/variable.
- **Backard selection :** We stat the "biggest" model whose intercept. At each step, a regressor/variable is removed to the model, the one which improves the chosen criterion. We stop when the criterion can not be improved by removing a new regressor/variable.
- **Stepwise selection/both selction :** This is the same method as the Forward selection method, except that at each step, a regressor/variable present in the model can be challenged (removed or added).

Comments:

☛ Remember that the $\mathbb{1}_n$ intercept/variable is always present for all models.

☛

5.6. Illustrative example under R

Comme back to the example of the Chapter 4.

- Y =`Consommation` = Fuel consumption in liters per 100 km.
- X_1 =`Prix` = Vehicle price in Swiss francs.
- X_2 =`Cylindree` = Cylinder capacity in cm3.
- X_3 =`Puissance` = Power in kW.
- X_4 =`Poids` = Weight in kg.

Consider the following linear model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i, \ \forall i = 1, \cdots, n. \tag{3}$$

We recall that we assume the model~(3), under the Rank assumption and under **[P1]–[P4]** where

- Errors are centered (linearity of the model in pratice) : $\forall i = 1, \cdots, n \quad \mathbb{E}_\beta[\varepsilon_i] = 0$.
- Errors have homosedasctic variance : $\forall i = 1, \cdots, n \quad \mathbb{V}\text{ar}_\beta[\varepsilon_i] = \sigma^2 > 0$.
- Errors are uncorralated: $\forall i \neq j \quad \mathbb{C}\text{ov}(\varepsilon_i, \varepsilon_j)$.
- Errors are gaussian : $\forall i = 1, \cdots, n \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

First download the dataset.

```
conso_voit = read.table("conso.txt", header=TRUE, sep="\t", dec=",",row.names=1)
```

The linear regression of the `Consommation` variable on the other variables is done using the `lm` function.

```
reg = lm(Consommation~Prix+Cylindree+Puissance+Poids,data=conso_voit)
```

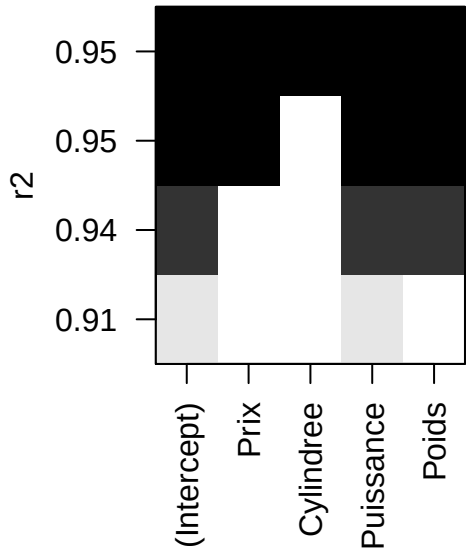
The question is : Are all the predictors relevant? To answer this question, if p is small we can procede as follows.

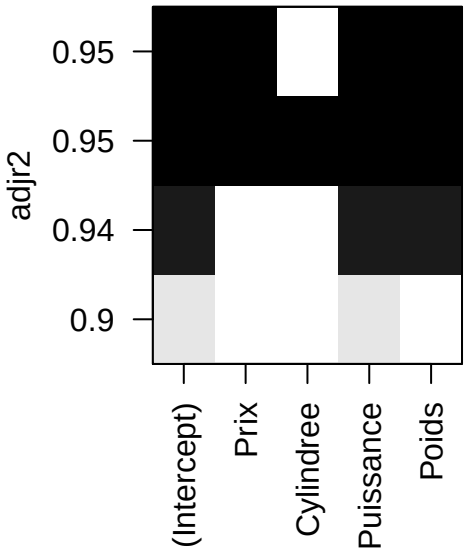
```
library(leaps)
# int=T (to include the intercept)
# nbest=1 (to select one model by dimension)
# nvmax=4 (the maximum number of regressors
   #(except the intercept) to include in a model)
choosen_model=regsubsets(Consommation~Prix+Cylindree+Puissance+Poids,int=T,
                         nbest=1,nvmax=4,method="exhaustive",data=conso_voit)

summary(choosen_model)
```

```
## Subset selection object
## Call: regsubsets.formula(Consommation ~ Prix + Cylindree + Puissance +
##     Poids, int = T, nbest = 1, nvmax = 4, method = "exhaustive",
##     data = conso_voit)
## 4 Variables  (and intercept)
##           Forced in Forced out
## Prix          FALSE      FALSE
## Cylindree     FALSE      FALSE
## Puissance     FALSE      FALSE
## Poids         FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: exhaustive
##          Prix Cylindree Puissance Poids
## 1  ( 1 ) " "  " "       "*"       " "
## 2  ( 1 ) " "  " "       "*"       "*"
## 3  ( 1 ) "*"  " "       "*"       "*"
## 4  ( 1 ) "*"  "*"       "*"       "*"
```

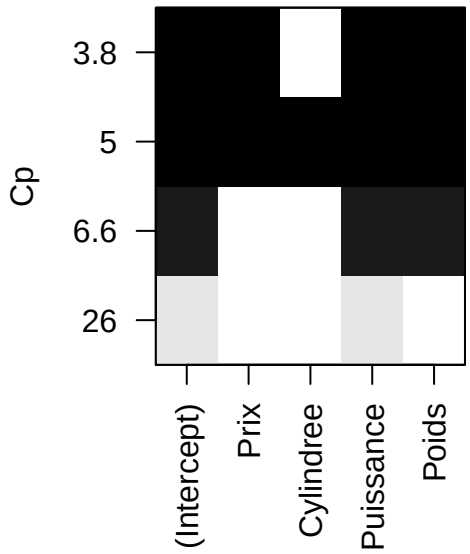
Plot make it easier to understand.

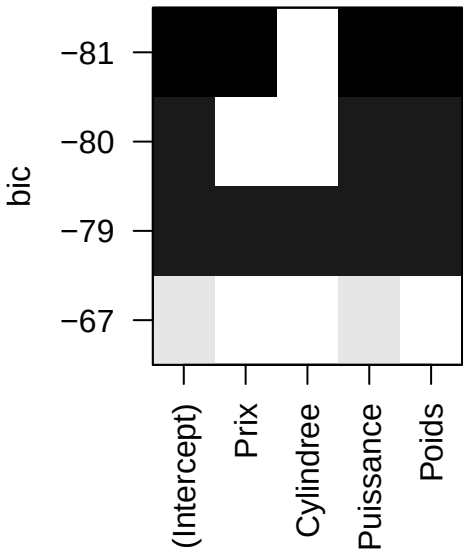
```
par(mfrow=c(1,2))
plot(choosen_model,scale="r2")
plot(choosen_model,scale="adjr2")
```






```
plot(choosen_model,scale="Cp")
plot(choosen_model,scale="bic")
```





5.6.1. Step by step methodes AIC

```
library(MASS)
```

5.6.1.1. Forward method

```
reg0=lm(Consommation~1,data=conso_voit)
stepAIC(reg0, Consommation~Prix+Cylindree+Puissance+Poids,data=conso_voit,
        trace=T,direction=c('forward'))
```

```
## Start:  AIC=79.87
## Consommation ~ 1
##
##              Df Sum of Sq    RSS    AIC
## + Puissance   1    346.79  35.35  8.071
## + Cylindree   1    338.37  43.77 14.692
## + Prix        1    303.45  78.69 32.878
## + Poids       1    285.17  96.96 39.351
## <none>                     382.14 79.866
##
## Step:  AIC=8.07
## Consommation ~ Puissance
##
##              Df Sum of Sq    RSS     AIC
## + Poids       1   14.2733 21.077 -5.9605
## + Cylindree   1    3.0114 32.339  7.3104
## <none>                     35.350  8.0706
## + Prix        1    0.0002 35.350 10.0704
##
## Step:  AIC=-5.96
## Consommation ~ Puissance + Poids
##
##              Df Sum of Sq    RSS     AIC
## + Prix        1    3.2053 17.871 -9.0744
## <none>                     21.077 -5.9605
## + Cylindree   1    0.0580 21.019 -4.0460
##
## Step:  AIC=-9.07
## Consommation ~ Puissance + Poids + Prix
##
##              Df Sum of Sq    RSS     AIC
## <none>                     17.871 -9.0744
## + Cylindree   1   0.50652 17.365 -7.9657
```

```
##
## Call:
## lm(formula = Consommation ~ Puissance + Poids + Prix, data = conso_voit)
##
## Coefficients:
## (Intercept)    Puissance        Poids         Prix
```

```
##   2.499e+00    2.013e-02    3.735e-03    1.852e-05
```

5.6.1.2. Backward method

```
stepAIC(reg,~.,trace=TRUE,direction=c("backward"))
```

```
## Start:  AIC=-7.97
## Consommation ~ Prix + Cylindree + Puissance + Poids
##
##             Df Sum of Sq    RSS     AIC
## - Cylindree  1    0.5065 17.871 -9.0744
## <none>                   17.365 -7.9657
## - Prix       1    3.6537 21.019 -4.0460
## - Puissance  1    4.1792 21.544 -3.2805
## - Poids      1   14.9706 32.335  9.3075
##
## Step:  AIC=-9.07
## Consommation ~ Prix + Puissance + Poids
##
##             Df Sum of Sq    RSS     AIC
## <none>                   17.871 -9.0744
## - Prix       1    3.2053 21.077 -5.9605
## - Puissance  1    3.9434 21.815 -4.8934
## - Poids      1   17.4783 35.350 10.0704
```

```
##
## Call:
## lm(formula = Consommation ~ Prix + Puissance + Poids, data = conso_voit)
##
## Coefficients:
## (Intercept)         Prix    Puissance        Poids
##   2.499e+00    1.852e-05    2.013e-02    3.735e-03
```

5.6.1.3. Both method

```
stepAIC(reg0,Consommation~Prix+Cylindree+Puissance+Poids,data=conso_voit,
        trace=TRUE,direction=c("both"))
```

```
## Start:  AIC=79.87
## Consommation ~ 1
##
##             Df Sum of Sq    RSS    AIC
## + Puissance  1    346.79  35.35  8.071
## + Cylindree  1    338.37  43.77 14.692
## + Prix       1    303.45  78.69 32.878
```

```
## + Poids      1    285.17  96.96 39.351
## <none>                   382.14 79.866
##
## Step:  AIC=8.07
## Consommation ~ Puissance
##
##             Df Sum of Sq    RSS    AIC
## + Poids      1     14.27  21.08 -5.961
## + Cylindree  1      3.01  32.34  7.310
## <none>                    35.35  8.071
## + Prix       1      0.00  35.35 10.070
## - Puissance  1    346.79 382.14 79.866
##
## Step:  AIC=-5.96
## Consommation ~ Puissance + Poids
##
##             Df Sum of Sq    RSS    AIC
## + Prix       1     3.205 17.871 -9.074
## <none>                   21.077 -5.961
## + Cylindree  1     0.058 21.019 -4.046
## - Poids      1    14.273 35.350  8.071
## - Puissance  1    75.888 96.964 39.351
##
## Step:  AIC=-9.07
## Consommation ~ Puissance + Poids + Prix
##
##             Df Sum of Sq    RSS     AIC
## <none>                   17.871 -9.0744
## + Cylindree  1    0.5065 17.365 -7.9657
## - Prix       1    3.2053 21.077 -5.9605
## - Puissance  1    3.9434 21.815 -4.8934
## - Poids      1   17.4783 35.350 10.0704
##
## Call:
## lm(formula = Consommation ~ Puissance + Poids + Prix, data = conso_voit)
##
## Coefficients:
## (Intercept)    Puissance        Poids         Prix
##   2.499e+00    2.013e-02    3.735e-03    1.852e-05
```

5.6.2. Step by step methodes BIC

Recall that the size of the sample is $n = 31$. Here note that `trace=F`, then the details won't appear. The command `k=log(n)` has to be added if we want to use BIC criterion (AIC is by default).

```
dim(conso_voit)
```

```
## [1] 31  5
```

```
n=31
```

```
## Forward method
reg0=lm(Consommation~1,data=conso_voit)
stepAIC(reg0, Consommation~Prix+Cylindree+Puissance+Poids,data=conso_voit,
        trace=F,direction=c('forward'),k=log(n))
```

```
##
## Call:
## lm(formula = Consommation ~ Puissance + Poids + Prix, data = conso_voit)
##
## Coefficients:
## (Intercept)    Puissance        Poids         Prix
##   2.499e+00    2.013e-02    3.735e-03    1.852e-05
```

```
## Backward method
stepAIC(reg,~.,trace=F,direction=c("backward"),k=log(n))
```

```
##
## Call:
## lm(formula = Consommation ~ Prix + Puissance + Poids, data = conso_voit)
##
## Coefficients:
## (Intercept)         Prix    Puissance        Poids
##   2.499e+00    1.852e-05    2.013e-02    3.735e-03
```

```
## Both  method
stepAIC(reg0,Consommation~Prix+Cylindree+Puissance+Poids,data=conso_voit,
        trace=F,direction=c("both"),k=log(n))
```

```
##
## Call:
## lm(formula = Consommation ~ Puissance + Poids + Prix, data = conso_voit)
##
## Coefficients:
## (Intercept)    Puissance        Poids         Prix
##   2.499e+00    2.013e-02    3.735e-03    1.852e-05
```

5.6.3 To conclude

Note that in our example the given result of the 3 methods is the same even for 2 differents criterions. It is not always the case.

Introduction to Regression - Chapter 6

MAP 535

Contents

Chapter 6 : ANOVA

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

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*

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

➡ *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 ...
```

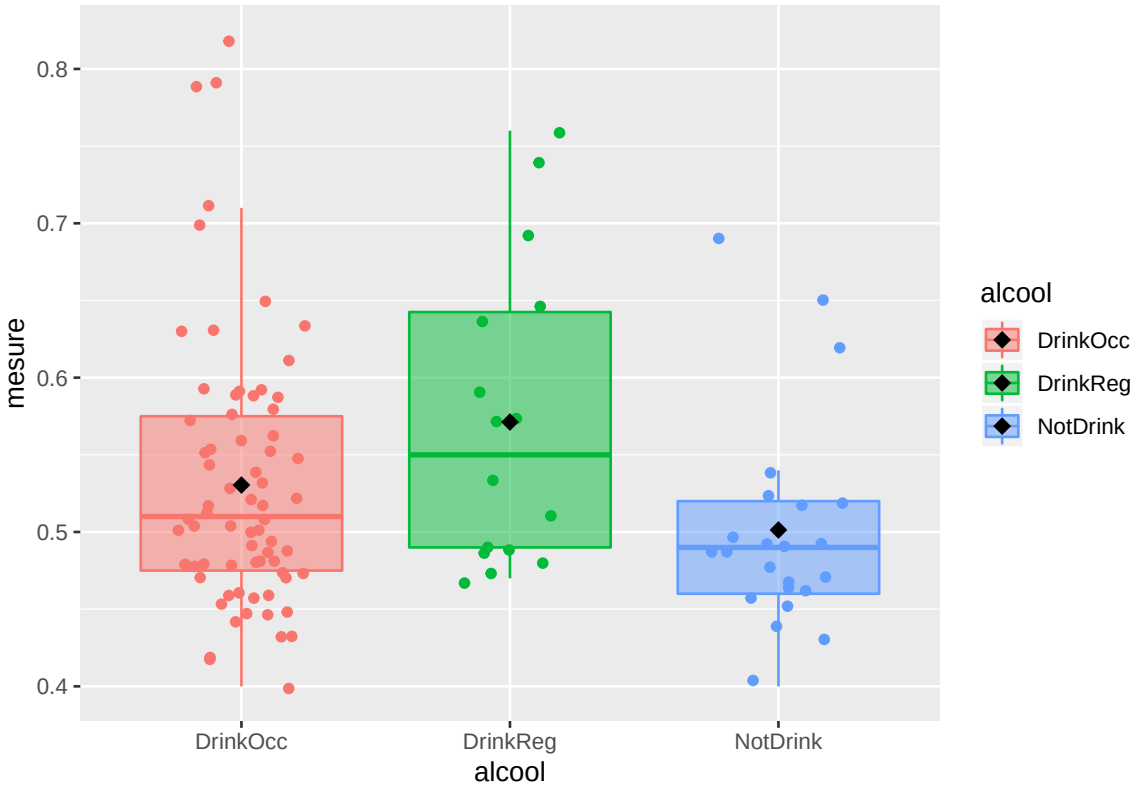
Plot of the dataset

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)
```



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,

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

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,

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

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.

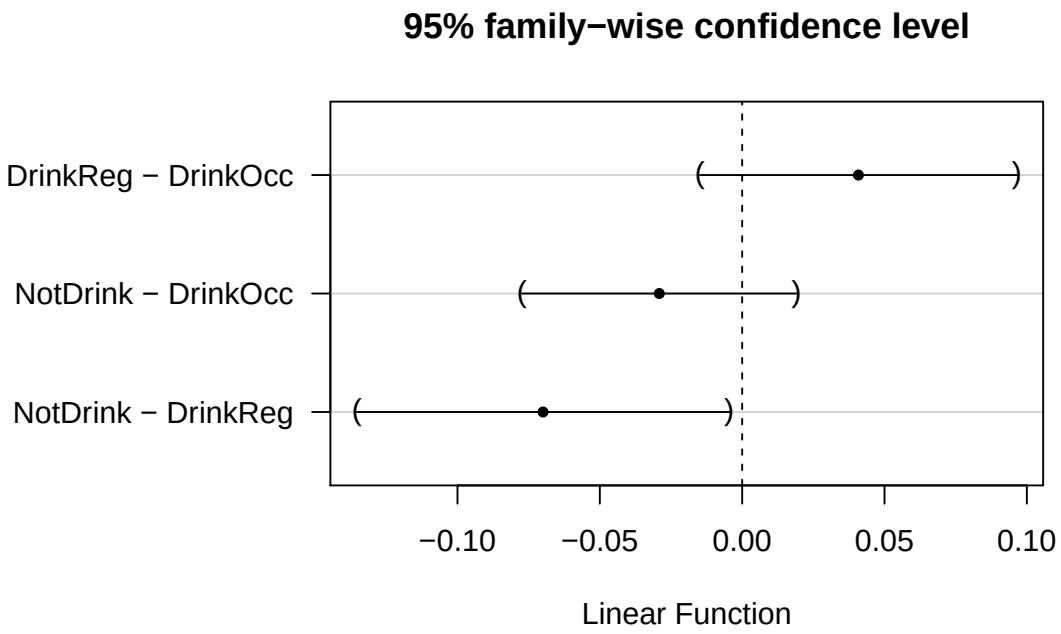
```
library(multcomp)
mc_tukey =glht(mod1, linfct=mcp(alcool="Tukey"))
summary(mc_tukey)
```

```
##
##   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   0.1924
## NotDrink - DrinkOcc == 0 -0.02912    0.02038  -1.429   0.3247
## 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

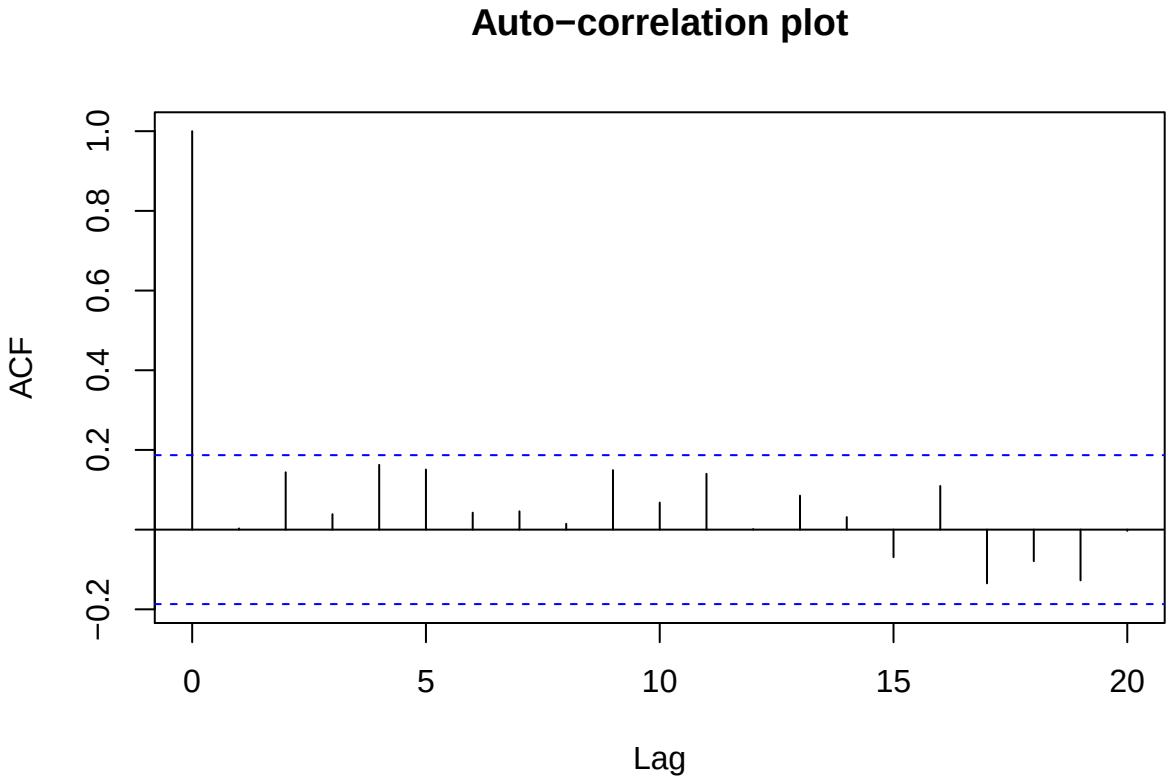
The Durbin Watson test tests the auto correlation. It is therefore concluded that there is no autocorrelation as the test *p-value* is heregreater than 5%.

```
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.

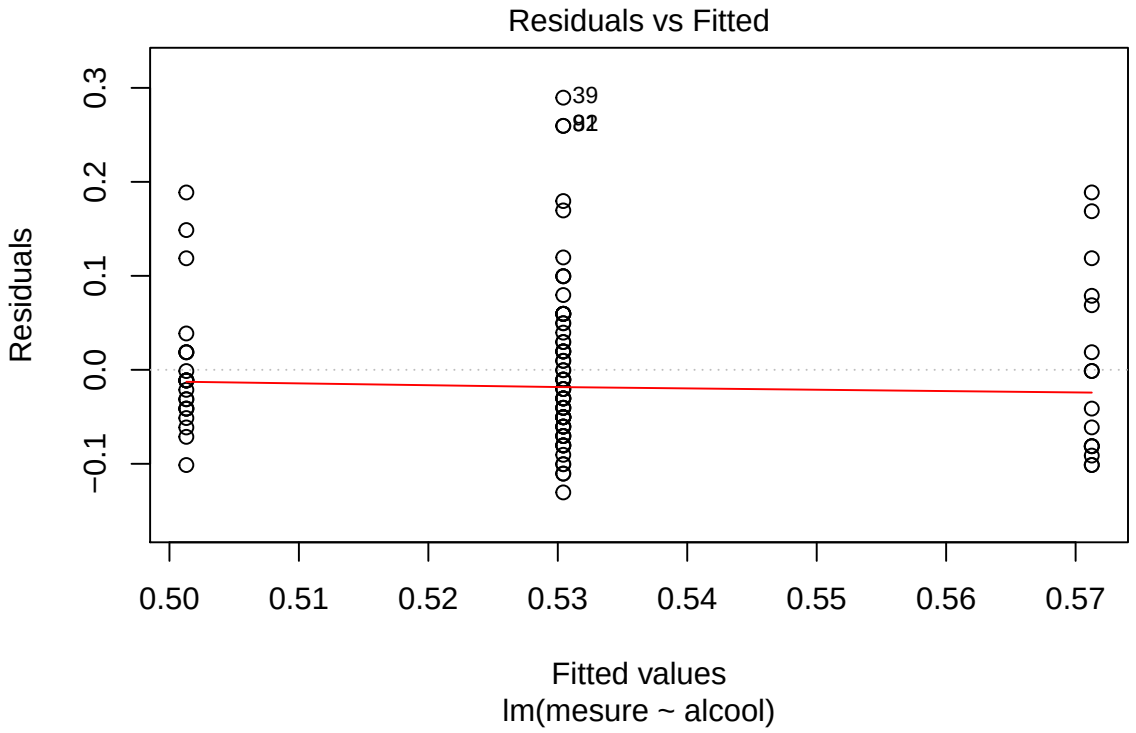
```
acf(residuals(mod1),main="Auto-correlation plot")
```



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

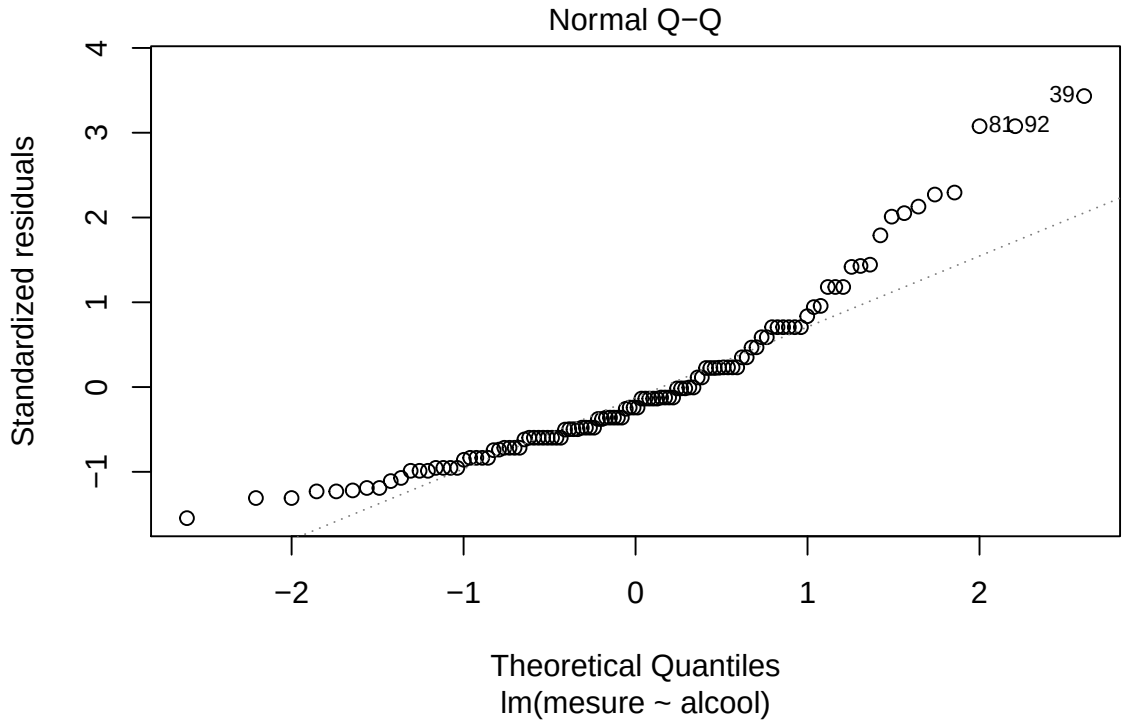
The *p-value* of the Shapiro test is very small, so we reject the postulat on the normality of the residues.

```
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

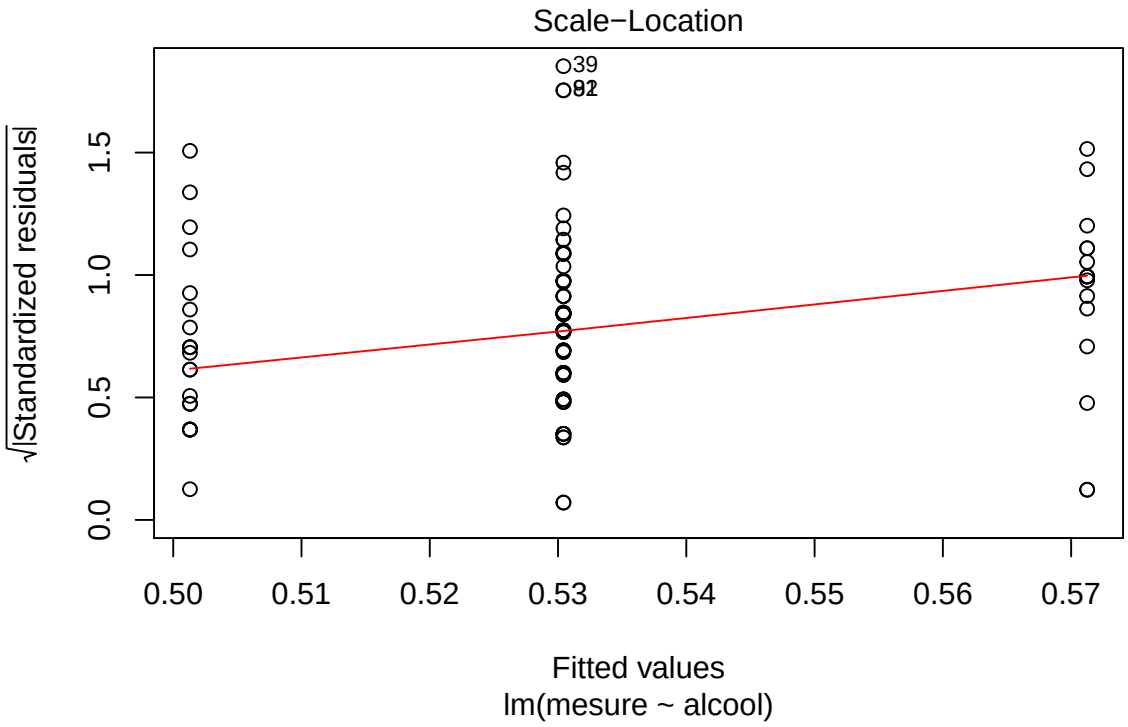
We can use the Bartlett test, (H_0: the variances of the different groups are globally identical and $H_1 = \overline{H_0}$). In our setting, the p-value is larger than 5%, we can't reject H_0

```
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.

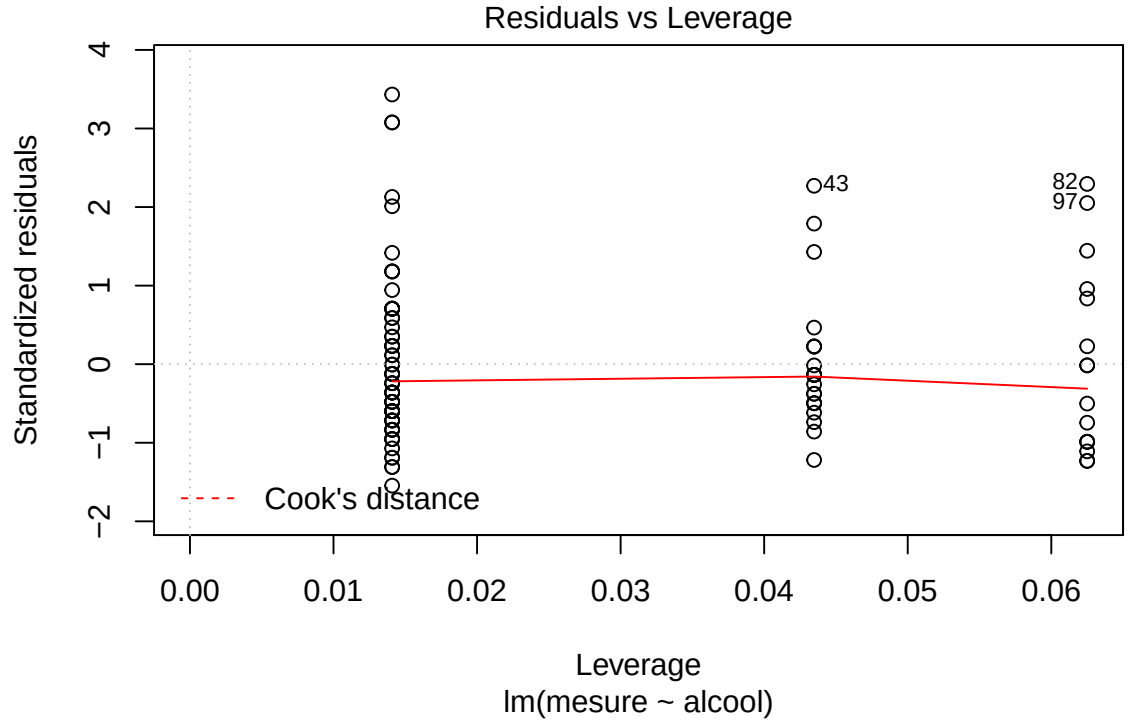
```
plot(mod1,3)
```



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 :

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

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.

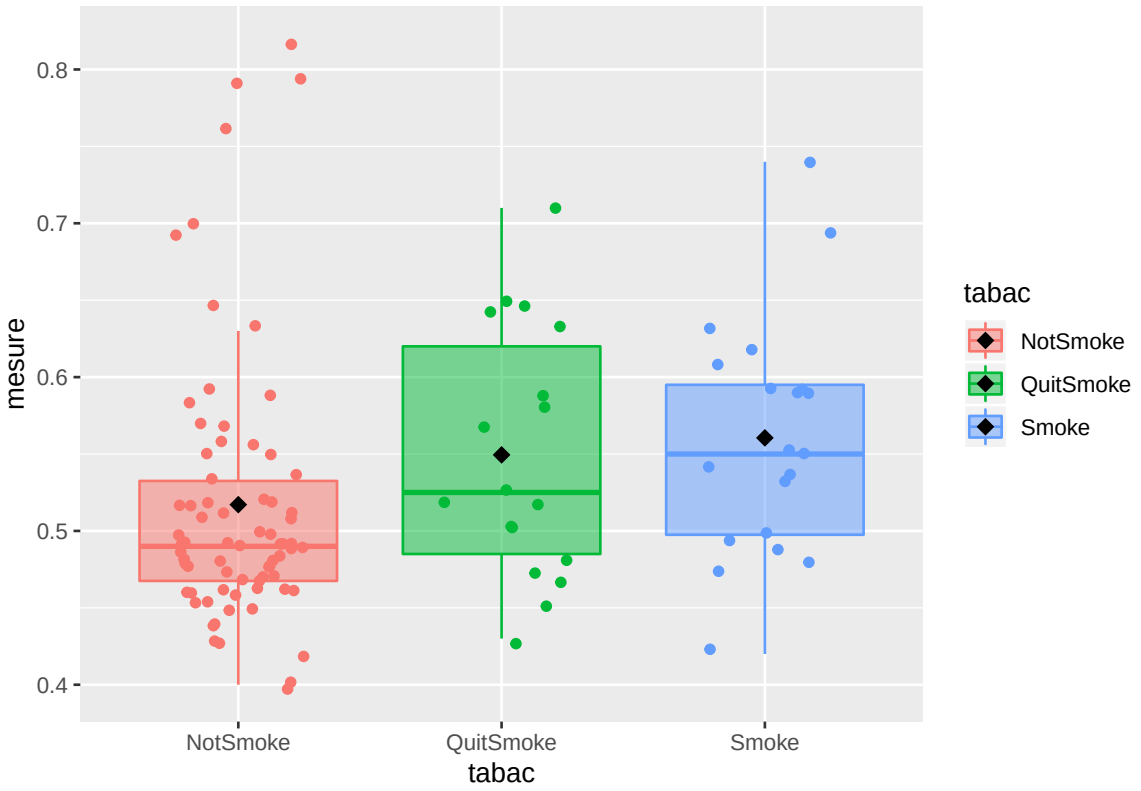
```
marqueur$tabac=replace(marqueur$tabac,marqueur$tabac==0,"NotSmoke")
marqueur$tabac=replace(marqueur$tabac,marqueur$tabac==1,"QuitSmoke")
marqueur$tabac=replace(marqueur$tabac,marqueur$tabac==2,"Smoke")
marqueur$tabac=as.factor(marqueur$tabac)
```

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)
```



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	$n_{\cdot k}$
$k=1$: NotSmoke	18	9	45	**72**
$k=2$: QuitSmoke	3	1	14	**18**
$k=3$: Smoke	2	6	12	**20**
$n_{j\cdot}$	**23**	**16**	**71**	$n=110$

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

➤➤ Note that an easy way to display $\overline{Y}_{\cdot jk}$, 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 $\overline{Y}_{\cdot jk}=\frac{1}{n_{jk}}\sum_{i=1}^{n_{jk}}Y_{ijk}$

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

```
mean(Tp)
```

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

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

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

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

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

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

➤➤ 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"
##  [6] "AcTot"     "Tartrique" "Malique"   "Citrique"  "Acetique"
## [11] "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.

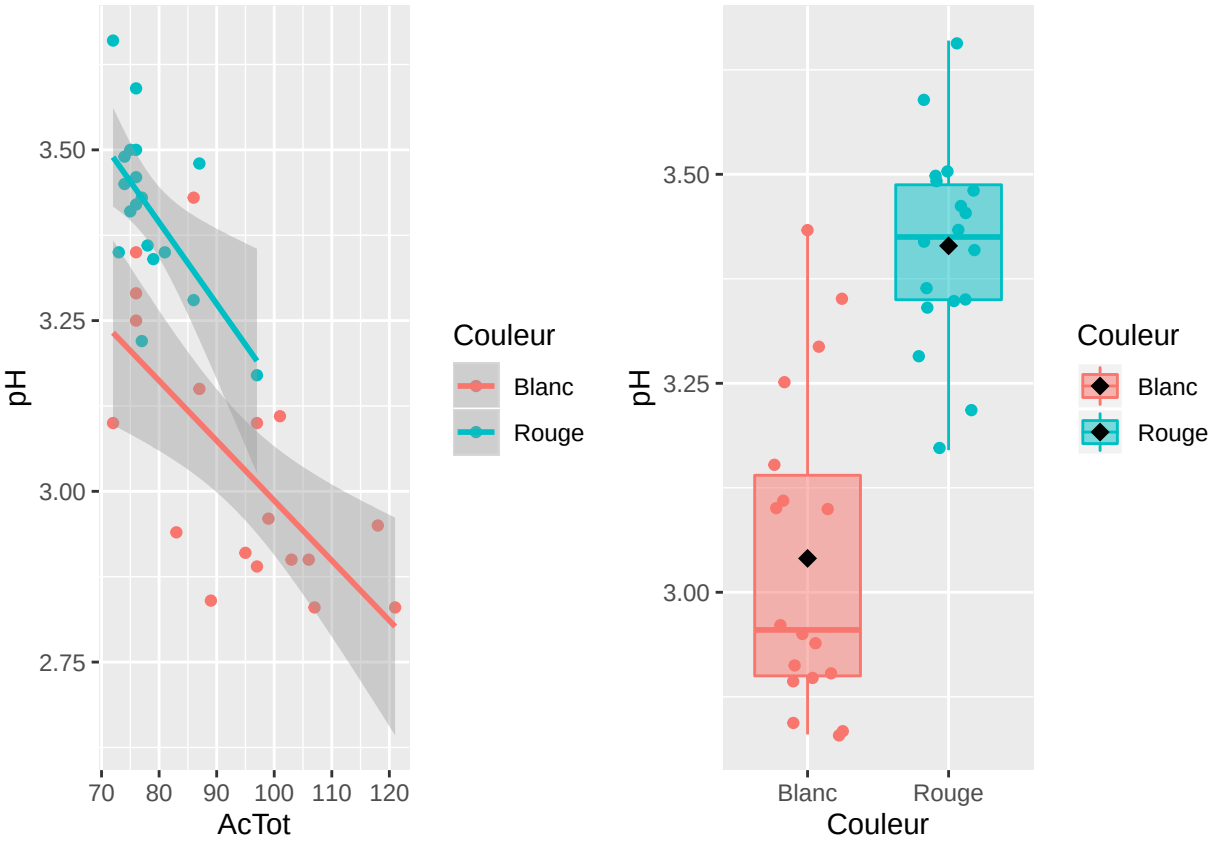
```
c(mean(Tmean),mean(Cepages$pH))
```

```
## [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)
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

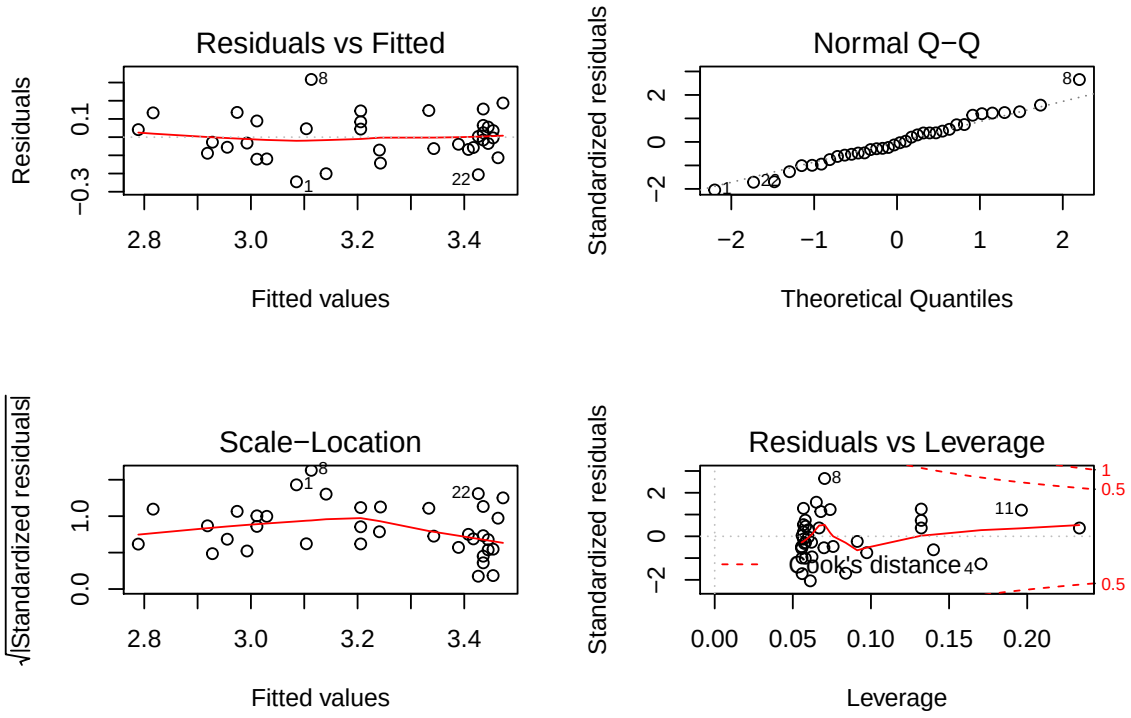
$$Y = \mu \mathbb{1}_n + A\alpha + bx + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(O_n, \sigma^2 \mathbb{I}_n),$$

```r
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
```

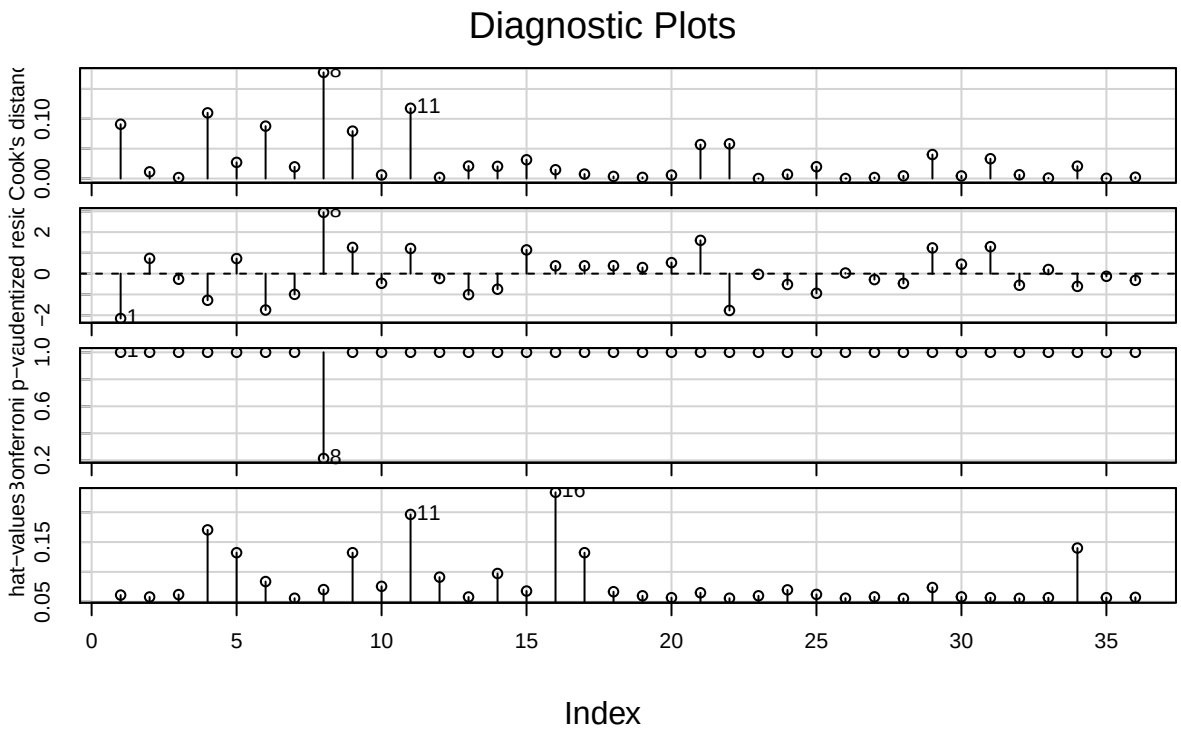
➤➤ We have to validate the model. Graphically, it can be done as follows

```r
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.