INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

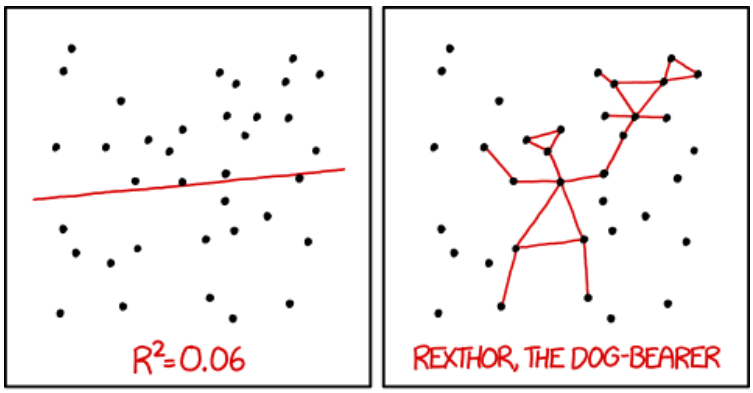
GMM Department
4th year MA

Academic year 2022–2023

Elements of Statistical Modeling

Statistical Tests
Linear Model & Generalized Linear Model

Juliette Chevallier *



* Office GMM 109
juliette.chevallier@insa-toulouse.fr

Acknowledgements
This handbook is based on those previously produced by GMM colleagues, for which I would like to thank them.

All models are wrong, but some are useful.

– George Box*

* George Box (1919 – 2013), British statistician.

Contents

Introduction to Statistical Modeling

1

A large part of applied mathematics consists, in a certain way, in modeling, that is to say, in designing one (or several) model(s) of a mathematical nature, allowing to explain, in a sufficiently general way, a given phenomenon, whether it is physical, biological, economic or other. Schematically, we can distinguish between deterministic modeling and stochastic modeling. In a deterministic model, we do not consider random variations; on the contrary, stochastic modeling considers these random variations by associating them with a probability law.

The classical tools of deterministic modeling are the ordinary differential equations (ODE) and the partial differential equations (PDE), which consider the variations of a phenomenon according to factors such as time, temperature... These equations rarely have explicit solutions, and their resolution often requires the implementation of numerical algorithms to obtain a solution, possibly approximate.

The main objective of stochastic modeling is to specify probability laws that take into account the random variations of certain phenomena, variations due to unknown or unmeasurable causes (for example, because they are to come). Within stochastic modeling, probabilistic modeling aims to provide a formal framework for describing the random variations mentioned above and studying the general properties of the phenomena that govern them. In a more applied sense, statistical modeling essentially consists in defining appropriate tools to model the observed data, taking into account their random nature.

Note that the term statistical modeling is very general. Hence, ultimately, any statistical approach falls under it. However, what we will deal with in this course is relatively precise and constitutes a specific part of statistical modeling. As a consequence, there are many statistical modeling methods. Here, we will study only a small part of them. At the same time, the considerable increase in the amount of data (internet, high-speed biology, marketing, *etc.*), the need to exploit these data statistically, and modern computing tools have given birth to numerous methods in the last few years.[1] However, these methods are not only more sophisticated, they are also more and more "greedy" in terms of computation time.

1: Let's say since the beginning of the XXIst century

There is almost always a privileged variable in the methods we will study, generally called the variable to be explained or the response variable, and denoted Y (of course, a random variable). The objective is then to build a model that explains "as well as possible" this variable Y as a function of explanatory variables observed on the same sample.

1.1 Illustrative Dataset

To illustrate the statistical approach and the problems that linear and generalized linear models can address, we present here a statistical analysis on a simple example.

In a population-based study, a hospital was interested in the snoring propensity of 100 patients. The variables considered are:

- *age*: in years,
- *weight* : in kg,
- *height*: in cm,
- *alcohol*: number of glasses drunk per day (in red wine equivalent),
- *sex*: sex of the person (F=female, M=male),
- *snoring*: diagnosis of snoring (Y=snoring, N=not snoring),
- *smoking*: smoking behavior (Y=smoker, N=non-smoker).

An extract of the data is presented below:

	age	weight	height	alcohol	sex	snoring	smoking
1	47	71	158	0	M	N	Y
2	56	58	164	7	M	Y	N
3	46	116	208	3	M	N	Y
4	70	96	186	3	M	N	Y
5	51	91	195	2	M	Y	Y
6	46	98	188	0	F	N	N

The dataset associated with this chapter is available on the moodle page of the course: `snore.txt`.

Listing 1.1: Exploratory statistics of quantitative variables.

```
> summary(snore[c("age",
    "weight","height")])

      age
 Min.   :23.00
 1st Qu.:43.00
 Median :52.00
 Mean   :52.27
 3rd Qu.:62.25
 Max.   :74.00

     weight
 Min.   : 42.00
 1st Qu.: 77.00
 Median : 95.00
 Mean   : 90.41
 3rd Qu.:107.00
 Max.   :120.00

     height
 Min.   :158.0
 1st Qu.:166.0
 Median :186.0
 Mean   :181.1
 3rd Qu.:194.0
 Max.   :208.0
```

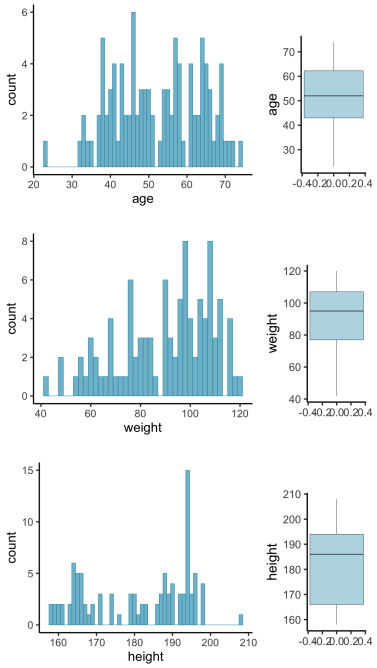


Figure 1.1: Graphical representations of the distribution of quantitative variables: age, weight and height.

1.1.1 Variables of Different Nature

Variables are analyzed differently depending on their nature: quantitative or qualitative.

A quantitative variable is a variable that can be represented by numbers on which the basic arithmetic operations have a meaning. They are usually summarized in the form of an indicator: mean, standard deviation, *etc*. as in Listing 1.1. Graphically, we generally opt for a histogram or a box plot for continuous quantitative variables, and for a bar chart for discrete quantitative variables (*cf*. Figure 1.1).

On the other hand, qualitative variables characterize an individual's membership in a group (or category). A qualitative variable is therefore coded with mutually exclusive classes (each individual can only belong to one category). Categorical variables are therefore described in terms of counts (absolute frequency of each modality) and percentages (relative frequencies), as in Table 1.1. They are graphically displayed in bar charts (*cf*. Figure 1.2).

1.2 Modeling Quantitative Variables

In this part, we try to evaluate the possible effect of the individuals' characteristics on their weight (quantitative variable). Depending on the nature of the variables, the methods of analysis are different.

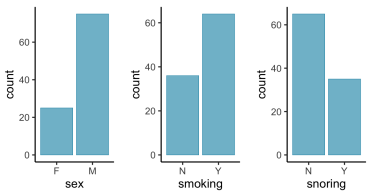


Figure 1.2: Bar graphs representing the distribution of categorical variables: gender, smoking and snoring.

1.2.1 Comparative Study of Two Populations

Before proposing a statistical model, we would like to know if gender impacts people's weight.

Visually, in Figure 1.3, we represent the empirical distribution of weights, on the one hand, for women and, on the other hand, for men. Assuming that gender does not affect weight, we should observe similar, or at least comparable, distributions. This does not seem to be the case here. Nevertheless, the question remains whether this difference is statistically significant or not.

Table 1.1: Frequency table by gender, smoking and snoring.

	Modality	Freq.(%)
Gender	Female	25
	Male	75
Smoking	Yes	64
	No	36
Snoring	Yes	35
	No	65

In 3rd year, you studied a test to test the equality of two Gaussian variables: the Student t-test. To determine whether we can use such a test, we shall first test the normality of our samples. To this end, we can perform several statistical test procedures: Q-Q plot, Kolmogorov test, or Shapiro-Wilk test for instance.

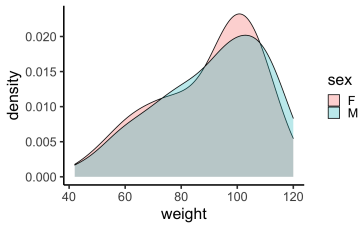


Figure 1.3: Weight distribution of individuals according to their gender.

Listing 1.2 shows the result of a Shapiro-Wilk procedure. The p-values associated with each test are less than 0.05. So, we reject the normality hypothesis with a 5% risk regardless of the gender of the individuals. Thus, we cannot compare these two populations using a Student t-test. We have to use more generic tests: namely, non-parametric tests, such as the Wilcoxon-Mann-Whitney or the Kolmogorov-Smirnov one.

We will detail all these tests in Chapter 3.

Listing 1.2: Normality test for weights.

```
> weightF = snore$weight[
    snore$sex=="F"]
> weightM = snore$weight[
    snore$sex=="M"]

> shapiro.test(weightF)
> shapiro.test(weightM)

Shapiro-Wilk normality test
data:  weightF
W = 0.9146, p-value = 0.03865

Shapiro-Wilk normality test
data:  weightM
W = 0.95406, p-value =
    0.008379
```

1.2.2 Linear Regression

To study the relationship between two quantitative variables (for example, between weight and height, or weight and age), one can plot a scatterplot (Figure 1.4) and calculate the linear correlation coefficient between these two variables.

```
> cor.test(snore$height, snore$weight, method = "pearson", conf.
    level = 0.95)

    Pearson s product-moment correlation

data:  snore$height and snore$weight
t = 24.463, df = 98, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8931821 0.9503567
sample estimates:
      cor
0.9269744
```

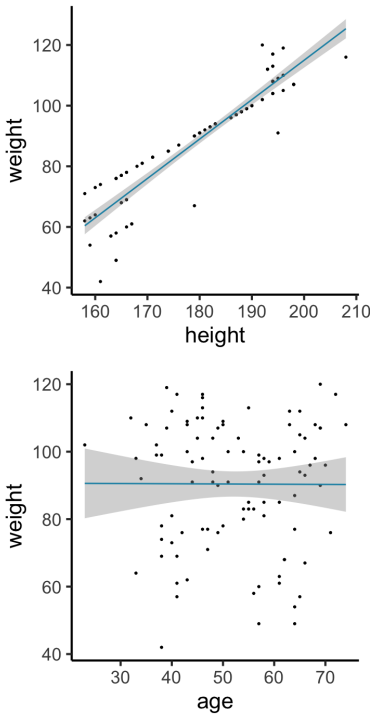


Figure 1.4: Scatterplot representing the relationship between weight and height (top), weight and age (bottom).

Table 1.2: Pearson's linear correlation coefficient and test of nullity of this coefficient.

	height	age
weight	0.92	-0.004
p-value	$< 2.2\,10^{-16}$	0.9687

In other words, the correlation coefficient here is 0.92 with a p-value less than $2.2\,10^{-16}$. For the pair (weight, age), we find that the correlation coefficient is -0.004 with a p-value equal to 0.9687. See Table 1.2. Hence, we find that the linear correlation coefficient is significantly different from 0 in the case of the regression of weight against height. This is not the case for the regression of weight against age.

1.2.2.1 Simple Linear Regression

The scatterplot can be summarized by a line that we will call the *simple linear regression* line. This is the simplest case of a linear model, which allows us to explain a quantitative variable in terms of another quantitative variable. For example, the linear regression line summarizing the relationship between weight and height has the equation:

$$\forall i \in [\![1, 100]\!], \qquad weight_i = a + b \times height_i + \varepsilon_i, \tag{1.1}$$

where ε_i is the error associated with each observation. Generally, these errors are assumed to be independent Gaussian variables with constant variance σ^2 to be estimated.

The statistical model underlying Equation (1.1) can also be presented in a matrix form.

Exercise 1.1 *Let the following vectors:* $\theta = {}^t(a, b)$

$$weight = {}^t(weight_1, \ldots, weight_{100}) \quad \text{and} \quad \varepsilon = {}^t(\varepsilon_1, \ldots, \varepsilon_{100}).$$

Show that the model can be written as

$$weight = X\theta + \varepsilon, \tag{1.2}$$

where X is a matrix to be specified.

In the model (1.2), $\theta = {}^t(a, b)$ and σ^2 are unknown. In order to estimate the parameters a and b, we use the *least squares method*. We thus choose the pair $(\hat{a}, \hat{b})$ verifying :

$$\begin{aligned}(\hat{a}, \hat{b}) &= \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{100} (weight_i - \alpha - \beta\, height_i)^2 \\ &= \underset{\alpha,\beta}{\operatorname{argmin}} \|weight - \alpha\, \mathbb{1}_{100} - \beta\, height\|^2 .\end{aligned}$$

In the chapter dedicated to linear regression (Part II), we will determine the explicit expression of these estimators and study their properties.

Using the `lm` function in R, we can easily fit this linear regression model on the data:

```
> reg <- lm(weight~height,data=snore)
> summary(reg)

Call:
lm(formula = weight ~ height, data = snore)

Residuals:
     Min       1Q   Median       3Q      Max
-22.2927  -1.9744   0.6785   5.7136  15.4269

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -144.90532    9.64523  -15.02   <2e-16 ***
height         1.29937    0.05312   24.46   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.064 on 98 degrees of freedom
Multiple R-squared:  0.8593,    Adjusted R-squared:  0.8578
F-statistic: 598.4 on 1 and 98 DF,  p-value: < 2.2e-16
```

In practice we obtain the following estimates:

- $(\hat{b})^{\text{obs}} = 1.299$: Estimate of the slope of the regression line, *i.e.* estimate of the average variation of weight with respect to height,
- $(\hat{a})^{\text{obs}} = -144.905$: Estimate of the intercept of the regression line,
- $(\sigma^2)^{\text{obs}} = 7.064^2$

The slope estimate equal to 1.299 is significantly different from 0, showing that weight and height are significantly related. These preliminary results only approximate the underlying linear-under model. In many situations, an in-depth study remains to be carried out to first "validate" the model and then exploit it: construction of tests, confidence intervals, *etc*. We will discuss these notions in more detail in the following chapters.

1.2.2.2 Multiple Linear Regression

It can also be interesting to model a variable as a function of several other quantitative variables, using a *multiple linear regression* model. For example, we can model weight as a function of height and age, which gives the following equation:

$$\forall i \in [\![1, 100]\!], \qquad weight_i = a + b \times height_i + c \times age_i + \varepsilon_i,$$

where $(\varepsilon_i)_{i \in [\![1,100]\!]}$ denote independent Gaussian variables with constant variance σ^2.

Exercise 1.2 *Considering the vectors* $\theta = {}^t(a, b, c)$

$$weight = {}^t(weight_1, \ldots, weight_{100}) \quad \text{and} \quad \varepsilon = {}^t(\varepsilon_1, \ldots, \varepsilon_{100}).$$

Show that the model can be written as

$$weight = X\theta + \varepsilon\,, \tag{1.3}$$

where X is a matrix to be specified.

In particular, one can notice by looking at Equation (1.2) and Equation (1.3) that the two linear regression models seen previously are written in a "same" matrix form.

1.2.3 Analysis of Variance (ANOVA)

It is possible to study the relationship between a quantitative variable and a qualitative variable, for example between weight and sex or between weight and smoking. This relationship is represented graphically by parallel boxplots (*cf.* Figure 1.5).

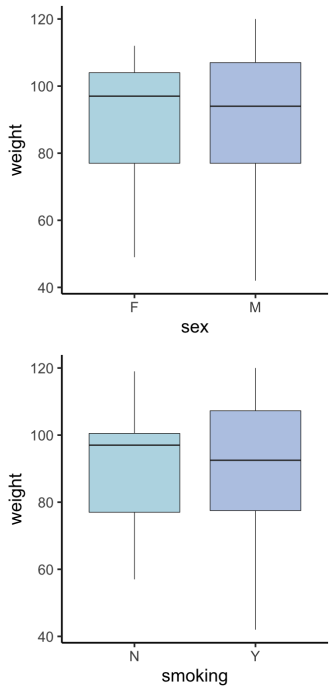


Figure 1.5: Parallel boxplots representing the relationship between weight and sex (top); between weight and smoking (bottom).

1.2.3.1 One-Way Analysis of Variance

Intuitively, to compare the weight of men and women, we will calculate the average weight for each group. Statistically, we model the weight as a function of gender by implementing a one-way analysis of variance model that is written as :

$$\forall i \in [\![1, 100]\!]\,, \qquad weight_i = a \cdot \mathbb{1}_{sex=F} + b \cdot \mathbb{1}_{sex=M} + \varepsilon_i\,,$$

where $(\varepsilon_i)_{i\in[\![1,100]\!]}$ denote independent Gaussian variables with constant variance σ^2. In this case, by reordering the observations according to the factor gender, the model can be written in the following matrix form:

$$\underbrace{\begin{pmatrix} weight_{F,1} \\ \vdots \\ weight_{F,n_F} \\ weight_{M,1} \\ \vdots \\ weight_{M,n_M} \end{pmatrix}}_{weight} = \underbrace{\begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} a \\ b \end{pmatrix}}_{\theta} + \underbrace{\begin{pmatrix} \varepsilon_{F,1} \\ \vdots \\ \varepsilon_{F,n_F} \\ \varepsilon_{M,1} \\ \vdots \\ \varepsilon_{M,n_M} \end{pmatrix}}_{\varepsilon},$$

where $weight_{F,i}$ denotes the weight of the i-th woman, $i \in [\![1, n_F]\!]$. The same for $weight_{M,i}$, $i \in [\![1, n_M]\!]$, for men. In practice, the least squares method is used to estimate the unknown parameters. Still using the `lm` function of R, we obtain the following results:

```
> anova <- lm(weight~sex-1,data=snore)
> summary(anova)

Call:
lm(formula = weight ~ sex - 1, data = snore)

Residuals:
    Min      1Q  Median      3Q     Max
-48.773 -12.773   4.227  16.227  29.227

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
sexF   89.320      3.764   23.73   <2e-16 ***
sexM   90.773      2.173   41.77   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.82 on 98 degrees of freedom
Multiple R-squared:  0.9593,    Adjusted R-squared:  0.9584
F-statistic:  1154 on 2 and 98 DF,  p-value: < 2.2e-16
```

The average weight of women and men is therefore $(\hat{a})^{\text{obs}} = 89$ and $(\hat{b})^{\text{obs}} = 91$ respectively.

1.2.3.2 Two-Way Analysis of Variance

Studying the combined effect of gender and smoking on weight is also possible. Intuitively, we can study class averages by crossing the two categorical variables. We implement a two-way analysis of variance model to explore this combined effect on weight. This model is written as follows: For all $i \in \{F, M\}$, $j \in \{Y, N\}$ and $k \in [\![1, n_{ij}]\!]$,

$$weight_{ijk} = a_i + b_j + c_{ij} + \varepsilon_{ijk},$$

where $weight_{ijk}$ denotes the weight of the $k \in [\![1, n_{ij}]\!]$ individual of sex $i \in \{F, M\}$ and smoking status $j \in \{Y, N\}$. The (ε_{ijk}) denote independent Gaussian variables with constant variance σ^2. We can also write this model in matrix form:

$$weight = X\theta + \varepsilon.$$

This model will allow us to study the effect of each factor (gender and smoking) on weight and detect combinations between gender and smoking that would give a remarkably different weight from other classes.

1.2.4 Analysis of Covariance (ANCOVA)

In our example, we can attempt to explain weight by height (quantitative variable) and gender (qualitative variable). In this case, we can draw two scatterplots between weight and height, one for women and one for men, as shown in Figure 1.6

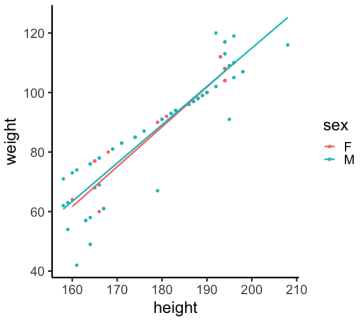


Figure 1.6: Scatterplots representing the relationship between weight and height according to gender.

The analysis of covariance model writes as follows:

$$\forall i \in \{F, M\}, \quad \forall j \in [\![1, n_i]\!], \quad weight_{i,j} = a_i + b_i\, height_{i,j} + \varepsilon_{i,j},$$

where $weight_{i,j}$ denotes the weight of individual j of sex i, and where the errors $\varepsilon_{i,j}$ are assumed to be independent centered Gaussian of variance σ^2.

We can therefore compare the effect of height on weight according to gender by implementing an *analysis of covariance* model. In practice, this corresponds to estimate a regression line of weight versus height for each modality of the sex variable.

In conclusion, in the different problems mentioned above, namely linear regression, analysis of variance, and analysis of covariance, we have used :

- the same type of matrix modeling,
- the same type of assumptions on the errors,
- the least-squares estimator.

In fact, these different problems are not as far apart as they seem a priori because the models used belong to the same family of models: *the linear model*.

1.3 Modeling Qualitative Variables

We can also be interested in studying a quantitative variable: for instance, individuals' smoking or non-smoking character.

1.3.1 Comparative Study of Two Populations

We would like to conclude about the dependence of two qualitative variables. For example, we would like to know if men generally smoke more than women. Table 1.3 gives the numbers of each of the gender/smoker cross-tabulations.

Table 1.3: *Tobacco use by gender.* Amout of smoking male (resp. female).

	F	M	Tot.
Smoker	10	54	64
Non-smoker	15	21	36
Total	25	75	100

To do this, we can, for example, set up a chi-squared test of independence (See Listing 1.3). We conclude negatively about the independence of the variables. In Chapter 4, we will study the so-called chi-squared test.

Listing 1.3: Chi-squared test.

```
> chisq.test(df_smoke)

Pearson's Chi-squared test
    with Yates' continuity
    correction

data:  df_smoke
X-squared = 7.0023, df = 1,
    p-value = 0.00814
```

Note: At this point, we cannot conclude that men smoke more than women, only that there is a dependence between gender and smoking. We need to go further in our statistical study to answer the stated question.

1.3.2 Logistic Regression

Last, let us now consider the case where the response variable Y is qualitative, and we wish to explain this variable Y according to some regressors $z^{(1)}, \dots, z^{(m)}$. Here are some illustrative examples:

Example 1.1 An insurance company seeks to detect fraudulent files. To do so, it has a panel of n files. To each of these files is associated the value 0 (for fraudulent file) or 1. After selecting the most interesting characteristics (household debt, social origin, place of residence, *etc.*), it determines to what extent these variables influence the probability of fraud. In this way, it hopes to be able to detect possible "sensitive" files in the future. We are in the case of a binary response variable Y.

Example 1.2 We seek to explain the number of plant species growing in different locations as a function of the biomass of those locations and the soil pH. The response variable Y here takes its values in $\mathbb{N}$.

In the case of a binary response variable, we observe the vector $Y = (Y_1, \dots, Y_n)$, where $Y_i \sim \mathcal{B}(\pi_i)$ for all $i \in [\![1, n]\!]$. Given the m regressors $z^{(1)}, \dots, z^{(m)}$, it seems quite natural to use the following model:

$$\forall i \in [\![1, n]\!], \quad \mathbb{E}[Y_i] = \pi_i = \sum_{j=1}^{m} a_j z_i^{(j)}.$$

However, as we are trying to model and predict probabilities, this approach does not seem very recommended as some predicted values might not belong to the interval $[0, 1]$. So instead, we will try to model a function of π_i by a linear combination of the explanatory variables $(z_i^{(j)})_j$. For example, in the context of *logistic regression*, we consider the *link function* $g :]0, 1[\to \mathbb{R}$ defined by

$$\forall t \in]0, 1[, \quad g(t) = \ln\left(\frac{t}{1-t}\right),$$

and we model

$$\forall i \in [\![1, n]\!], \quad g(\pi_i) = \sum_{j=1}^{m} a_j z_i^{(j)}.$$

More generally, it is possible to consider other distributions for the variable Y and other link functions. For example, the regression model discussed at the beginning of this chapter corresponds to a Gaussian distribution and a canonical (identity) link function. We will see that it is possible to study all these models by the same path: the *generalized linear model*.

1.4 Data Visualization

We have provided some descriptive statistics in the previous paragraphs to better understand our data. Note that it is always a good idea to actually visualize the data. Indeed, identical descriptive statistics can

hide very different realities. For example, the 13 datasets (the Datasaurus, plus 12 others) shown in Figure Figure 1.7 all have the same summary statistics (x/y mean, x/y standard deviation, and Pearson's correlation) within two decimal places while being drastically different in appearance.

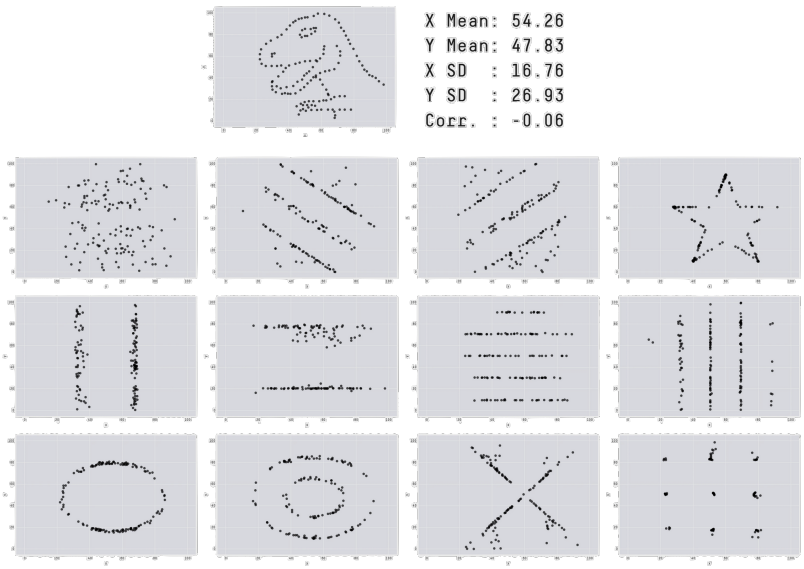


Figure 1.7: *The Datasaurus Dozen*: Alberto Cairo created the Datasaurus dataset which urges people to "never trust summary statistics alone; always visualize your data", since, while the data exhibits normal seeming statistics, plotting the data reveals a picture of a dinosaur

For more information, please visit: autodesk.com/research/publications/same-stats-different-graphs.

Statistical Tests

Reminders on Tests 2

In this first chapter, we propose to start by recalling the basic vocabulary of test theory. However, for the sake of brevity, all the parametric tests seen in 3rd year will not be recalled here.

A hypothesis test is a procedure for inferring (accepting or rejecting) the validity of hypotheses about one or more populations based on the study of one or more random samples. Statistical inference methods allow us to determine, with a given probability, whether the differences found in the samples can be attributed to chance or whether they are large enough to mean that the samples are probably from different populations.

There are several types of statistical tests:

- The *conformity test* compares a parameter calculated on the sample with a pre-established value. In other words, we assume a theoretical law, generally the normal distribution, and we want to check if our sample conforms to this law. The best known are the tests on mean, variance, or proportions.
  For example, we know that the 3rd face of a non-piped die has a chance of $1/6$ to occur. We ask a player to throw, without any particular precaution, a die 100 times. We then test if the frequency of appearance of 3 is compatible with the $1/6$ probability. If not, we can question the integrity of the die.
- The *goodness-of-fit* test checks the compatibility of the data with a distribution chosen a priori. The most commonly used test in this context is the Gaussian distribution test, which allows a parametric test to be applied.
- The *homogeneity* or comparison test tests that $k \geqslant 2$ samples are from the same population. Alternatively, it amounts to testing that the distribution of the variable of interest is the same in the k samples.
  For example: Is there a difference between the mean glucose level measured for two samples of individuals who received different treatments?
- The test of *independence* tests the existence of a link between two variables. The techniques used differ depending on whether the variables are nominal, ordinal, or quantitative.
  Example: Is the distribution of eye color observed in the French population independent of the sex of the individuals?

2.1 General Reminders on Statistical Tests

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and X a random variable from the set of possible outcomes $(\Omega, \mathcal{A})$ to a measurable space $(E, \mathcal{E})$. Let

consider a statistical model, *i.e.* a set of probability distributions on $(E, \mathcal{E})$: $\mathcal{P} = \{\mathbb{P}_\theta,\ \theta \in \Theta\}$. The set Θ defines the parameters of the model. Last, let consider a n-sample $\mathcal{X} = (X_1, \ldots, X_n)$ whose law is assumed to belong to $\mathcal{P}$.

A statistical hypothesis test is a method of statistical inference. The principle of hypothesis testing is to pose a working hypothesis and to predict the consequences of this assumption for the population. These predictions are compared with the observations, and the conclusion is reached by accepting or rejecting the hypothesis based on objective decision rules.

2.1.1 Null Hypothesis & Alternative Hypothesis

The first step is to define two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively called *null hypothesis* and *alternative hypothesis*. We then consider two disjoint subsets of Θ, Θ_0 and Θ_1, and we say that we test

$$\mathcal{H}_0 : \theta \in \Theta_0 \quad vs. \quad \mathcal{H}_1 : \theta \in \Theta_1 \,.$$

From the sample $\mathcal{X}$, we then want to construct a decision rule (rejection region) to discriminate between these two hypotheses.

Recall that the $\mathcal{H}_0$ and $\mathcal{H}_1$ hypotheses do not play a symmetrical role. The null hypothesis is the preferred hypothesis that we wish to control: it consists of saying that there is no difference between the compared parameters or that the observed differences are not significant and due to sampling fluctuations. This hypothesis is formulated with the aim of being rejected. The alternative hypothesis, $\mathcal{H}_1$, is the "negation" of $\mathcal{H}_0$ and is equivalent to saying "$\mathcal{H}_0$ is false". The decision to reject $\mathcal{H}_0$ therefore means that $\mathcal{H}_1$ is true. We refer to a *simple hypothesis* when the associated subset is a singleton and to a *composite hypothesis* in the opposite case.

Remark 2.1 There is an important asymmetry in the tests' conclusions. Indeed, the decision to accept $\mathcal{H}_0$ is not equivalent to "$\mathcal{H}_0$ is true and $\mathcal{H}_1$ is false". It only reflects that there is no clear evidence that $\mathcal{H}_0$ is false. A test leads to rejecting or non-rejecting null hypothesis, never to its straightforward acceptance.

The nature of $\mathcal{H}_0$ determines the way $\mathcal{H}_1$ is formulated and, consequently, the one-sided or two-sided nature of the test. We speak of a *two-sided* test when the alternative hypothesis is "decomposed into two parts"; conversely, we refer to a *one-sided* test when the alternative hypothesis is "composed of a single part".

Example 2.1 Denote p the frequency of smokers among students and p_0 the frequency of smokers in the general population.

- To test whether the student population has a frequency of smokers p representative of the one in general population p_0, we pose $\mathcal{H}_0$: "$p = p_0''$ and $\mathcal{H}_1$: "$p \neq p_0''$. The test considered is

then *two-sided* because the frequency p can be higher or lower than the frequency p_0;
- If we now assume that the frequency of smokers is higher in the student population than in the overall population p_0, we pose $\mathcal{H}_0$: "$p = p_0''$ and $\mathcal{H}_1$: "$p > p_0''$. The test is then *one-sided* because the frequency p can only be higher than p_0.

It would also have been possible to have $\mathcal{H}_0$: "$p = p_0''$ and $\mathcal{H}_1$: "$p < p_0''$. Refer to Table 2.2 for some examples.

Statistical Tests

Definition 2.1 (Statistical test) *A statistical test consists of a partition of Ω into two sets: the set $\mathcal{R}$ of possible values of the sample that lead to the rejection of the null hypothesis $\mathcal{H}_0$ in favor of the alternative $\mathcal{H}_1$, and its complement.*

We call $\mathcal{R}$ the *rejection region*, or critical region, of the test and $\mathcal{R}^{\complement}$ the *region of acceptance*. The threshold value delimiting the regions of acceptance and rejection is called *critical value*.

Definition 2.2 (Statistical test function) *We call test function of rejection region $\mathcal{R}$ the statistic $\phi(x) = \mathbb{1}_{x \in \mathcal{R}}$.*

In other words, if $\phi(x) = 1$, we reject $\mathcal{H}_0$, and if $\phi(x) = 0$, we do not reject $\mathcal{H}_0$ (and so accept $\mathcal{H}_1$).

2.1.2 Type I Error and p-value

The first kind of error is the rejection of a true null hypothesis as the result of a test procedure. We refer to this error as a type I error (*false positive*) and, less frequently,[1] error of the first kind.

1: At least in English, since in French we speak of "erreur de première espèce".

This occurs if the value of the test statistic falls into the rejection region while the $\mathcal{H}_0$ hypothesis is true. The probability of this event is the *significance level* α. The significance level is also said to be the probability of rejecting the null hypothesis incorrectly.

Let a test of rejection region $\mathcal{R}$ to test $\mathcal{H}_0$ against $\mathcal{H}_1$.

Definition 2.3 (Type I error) *For all $\theta_0 \in \Theta_0$, we define the* type I error function *as*

$$\alpha(\theta_0) = \mathbb{P}_{\theta_0}(\mathcal{X} \in \mathcal{R}).$$

The size of the test *corresponds to the maximum type I error:*

$$\alpha^* = \sup_{\theta_0 \in \Theta_0} \mathbb{P}_{\theta_0}(\mathcal{X} \in \mathcal{R}).$$

Definition 2.4 (Alpha-level) *Let $\alpha \in [0,1]$. We say that this test is:*

- *of* α level *if its size is at most* α, $\alpha^* \leqslant \alpha$;
- *of* exactly α *level if it is of* α *size;*
- asymptotically *of level* α *if* $\limsup_{m,n\to+\infty} \alpha^* \leqslant \alpha$;
- asymptotically *of size* α *if* $\lim_{m,n\to+\infty} \alpha^* = \alpha$

We also refer to alpha-levels as *risk* of the test. We usually set these levels to 0.05, 0.01 or 0.001.

Remark 2.2 The value of the risk α should be set *a priori* by the experimenter and never based on the data. It is a compromise between the risk of concluding wrongly and the ability to conclude. The critical region decreases as α decreases, and thus $\mathcal{H}_0$ is rejected less frequently. If we want to make fewer errors, we conclude less frequently.

Example 2.2 (Coin) We want to determine if a coin is rigged or not. Let X be the number of faces obtained by tossing the coin 100 times. We put into equation the hypothesis $\mathcal{H}_0$ "the coin is not rigged" as follows: $\mathcal{H}_0$: "$X \in [40,60]$". In particular, this is a two-sided test since $\mathcal{H}_0$ is rejected if $X < 40$ or $X > 60$.
The type I risk of this test is $\alpha = \mathbb{P}(\mathcal{B}(100, 1/2) \in [40,60])$, where $\mathcal{B}(n,p)$ is the binomial distribution with number of trials $n \in \mathbb{N}$, and success probability $p \in [0,1]$.

Suppose we have constructed for all $\alpha \in]0,1[$ a test of level α and rejection region $\mathcal{R}_\alpha$, allowing to test $\mathcal{H}_0$ against $\mathcal{H}_1$.

Definition 2.5 (p-value) *We call* p-value *of the tests' family the smallest level at which we reject* $\mathcal{H}_0$ *from the observed sample* $\mathcal{X}^{obs}$*:*

$$p(\mathcal{X}^{obs}) = \inf\left\{ \alpha \in]0;1[\;\middle|\; \mathcal{X}^{obs} \in \mathcal{R}_\alpha \right\}.$$

Intuitively, the p-value is the probability of obtaining test results at least as extreme as the observed results, assuming that the null hypothesis is correct. In other words, a small p-value means that such an extreme observed outcome would be very unlikely under the null hypothesis. The smaller the p-value, the stronger the evidence in favor of the alternative hypothesis: the p-value provides the lowest significance level at which the null hypothesis would be *rejected*. In particular, the p-value is *not* the probability that the test hypothesis is true. The p-value "only" indicates how well the data conform to the test hypothesis $\mathcal{H}_0$ and the assumptions made about it, *i.e.* the underlying statistical model.

Remark 2.3 (Misuse of p-values) The p-values are often used or interpreted incorrectly. For a more detailed explanation, you can refer to the following sourced Wikipedia article: `https://en.wikipedia.org/wiki/Misuse_of_p-values`. Let us mention here the points that are generally misunderstood about p-values:

1. The p-value is *not* the probability that the null hypothesis is true, or the probability that the alternative hypothesis is false;
2. The p-value is *not* the probability that the observed effects were produced by random chance alone;
3. The 0.05 significance level is merely a convention;
4. The p-value does not indicate the size or importance of the observed effect.

2.1.3 Type II Error and Power

Alternatively, the second type of error is the non-rejection of a false null hypothesis (*false negative*). This is known as type II error or error of the second kind.

This occurs if the value of the test statistic does not fall into the rejection region while hypothesis $\mathcal{H}_1$ is true. Table 2.1 presents the different possible error scenarios.

Let a test of rejection region $\mathcal{R}$ to test $\mathcal{H}_0$ against $\mathcal{H}_1$.

Definition 2.6 (Type II error) *For all $\theta_1 \in \Theta_1$, we define the* type II error function *as*

$$\beta(\theta_1) = \mathbb{P}_{\theta_1}(\mathcal{X} \notin \mathcal{R})$$

and the maxima type II error is

$$\beta^* = \sup_{\theta_1 \in \Theta_0} \mathbb{P}_{\theta_1}(\mathcal{X} \notin \mathcal{R}) .$$

Remark 2.4 To quantify the risk β, we need to know the probability distribution of the statistic under assumption $\mathcal{H}_1$

Example 2.3 (Coin) Go back to the previous example with the coin. We suppose the probability of getting a face is 0.6 for a rigged coin. By adopting the same decision rule for $\mathcal{H}_0$, the type II risk is $\beta = \mathbb{P}(\mathcal{B}(100, 0.6) \in [40, 60])$

Definition 2.7 (Power function) *We call* power function *of the test of rejection region $\mathcal{R}$ the application defined by:*

$$\pi : \ \theta_1 \in \Theta_1 \longmapsto \mathbb{P}_{\theta_1}(\mathcal{X} \in \mathcal{R}) = 1 - \beta(\theta_1) \in [0, 1] .$$

Obviously, as the power increases, the probability of a type II error decreases. In particular, among tests of the same level, the most powerful is always preferred.

Remark 2.5 The power of a test depends on the nature of $\mathcal{H}_1$: A one-sided test is more powerful than a two-sided test. The power $1 - \beta$ increases with the size of the sample studied, and it decreases when the significance α decreases. Therefore, a trade-off between power and significance is necessary when conducting a statistical test procedure.

Table 2.1: *Error types* according to the truthfulness of the null hypothesis and the outcome of the test.

$\mathcal{H}_0$	*True*	*False*
Accept	Correct $p = 1 - \alpha$	**Type II** $p = \beta$
Reject	**Type I** $p = \alpha$	Correct $p = 1 - \beta$

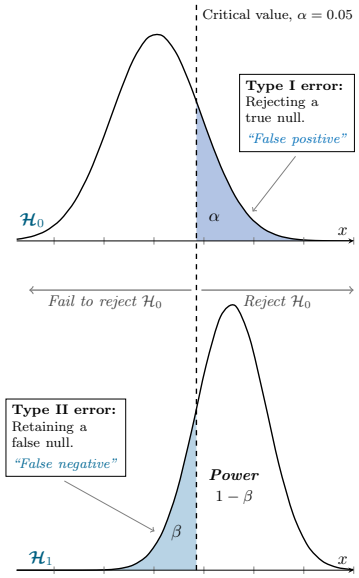


Figure 2.1: Statistical power of a test.

2: In French, we speak of "test UPP" for "Test uniformément plus puissant".

Table 2.2: *Alternative hypotheses* and *rejection region* associated with the null assumption $\mathcal{H}_0$: "$t = t_0''$" depending on whether the test is one- or two-sided.

	Alternative hypothesis	Rejection region
one-	$\mathcal{H}_1$: "$t > t_0''$"	$\mathcal{R} = \{S > S^\star\}$
	$\mathcal{H}_1$: "$t < t_0''$"	$\mathcal{R} = \{S < S^\star\}$
two-	$\mathcal{H}_1$: "$t \neq t_0''$"	$\mathcal{R} = \{\|S\| > S^\star\}$

Figure 2.1 summarises this situation.

A test based on the rejection region $\mathcal{R}$ is said to be *better* than one based on the rejection region $\mathcal{R}'$ if they are both of α level and

$$\forall \theta \in \Theta_1\,, \quad \mathbb{P}_\theta(\mathcal{X} \in \mathcal{R}) \geqslant \mathbb{P}_\theta(\mathcal{X} \in \mathcal{R}')\,.$$

Definition 2.8 (Uniformly most powerful) *A test based on the rejection region $\mathcal{R}_\alpha$ is said to be* uniformly more powerful *(UMP) at level α if*

1. $\sup_{\theta\in\Theta_0} \mathbb{P}_\theta(\mathcal{X} \in \mathcal{R}_\alpha) \leqslant \alpha$;
2. *For all rejection region $\mathcal{R}'_\alpha$ such that* $\sup_{\theta\in\Theta_0} \mathbb{P}_\theta(\mathcal{X} \in \mathcal{R}'_\alpha) \leqslant \alpha$,

$$\forall \theta \in \Theta_1, \quad \mathbb{P}_\theta(\mathcal{X} \in \mathcal{R}_\alpha) \geqslant \mathbb{P}_\theta(\mathcal{X} \in \mathcal{R}'_\alpha)$$

In other words, a uniformly most powerful test has the greatest power $1-\beta$ among all possible tests of a given size α. Please note that there is not always a UMP[2] test.

2.1.4 Methodological Considerations

A statistic (Definition 2.2) is a function of the random variables representative of the sample. The choice of the statistic depends on the nature of the data, the type of hypothesis that we wish to control, the assumptions that we can make about the populations studied... The numerical value of the statistic obtained for the considered sample allows us to judge the veracity of $\mathcal{H}_0$.

Assume that the probability distribution followed by the S-statistic under $\mathcal{H}_0$ is known. At a given probability α, it is then possible to establish a threshold value $S^\star$ of the statistic. Hence, by choosing the value of α as the significance level (Definition 2.4), the critical region $\mathcal{R}(S^\star)$ corresponds to the set of values such that $\mathbb{P}(S \in \mathcal{R}) = \alpha$. The definition of the critical region varies depending on whether the test is one- or two-sided (See Table 2.2).

There are two strategies for reaching a decision regarding the test of interest: the first strategy sets the value of the significance level α *a priori*, and the second sets the value of the critical probability α^{obs} *a posteriori*.

Decision Rule #1: Under the assumption "$\mathcal{H}_0$ is true" and for a fixed significance level α,

- If the value of the computed statistic S^{obs} belongs to the critical region $\mathcal{R}$, then assumption $\mathcal{H}_0$ is rejected at the risk of error α and hypothesis $\mathcal{H}_1$ is accepted;
- If the value of the statistic S^{obs} does not belong to the critical region, then hypothesis $\mathcal{H}_0$ cannot be rejected.

Decision Rule #2: We evaluate the critical probability α^{obs} such that $\mathbb{P}(S \in \mathcal{R}^{\text{obs}}) = \alpha^{\text{obs}}$, where $\mathcal{R}^{\text{obs}}$ is the observed counterpart of the rejection region $\mathcal{R}$, for instance $\mathcal{R}^{\text{obs}} = \{|S| \geqslant S^{\text{obs}}\}$ for a two-sided test. For a fixed significance level α,

- If $\alpha^{\text{obs}} \geqslant \alpha$, we accept assumption $\mathcal{H}_0$ since the risk of rejecting $\mathcal{H}_0$ while it is true is too large;
- If $\alpha^{\text{obs}} < \alpha$, we reject the $\mathcal{H}_0$ hypothesis because the risk of rejecting $\mathcal{H}_0$ while it is true is very low.

2.2 Parametric Tests (MIC 3)

Different statistical tests have been studied in the MIC3 Statistics UF. To construct all these tests, we assume that the law of the samples belongs to a parametric model, *i.e.* to a given family of laws described by a finite number of parameters. We then speak of *parametric tests*. Since these tests depend crucially on the nature of the statistical distribution of the observations, certain validity conditions must be satisfied to ensure their reliability. For example, the Student's t-test for independent samples is only reliable if the data associated with each sample follow a normal distribution and if the variances of the samples are homogeneous.

In practice, assumptions can be difficult to test. *Non-parametric tests* remove this limitation by offering a family of tests that are not based on statistical distributions. Therefore, they can be used whatever the distribution of the samples and even if the validity conditions of parametric tests are not verified.

Non-parametric tests are more robust than parametric tests. In other words, they can be used in a broader range of situations. However, parametric tests are, in general, more powerful than their non-parametric counterparts: A parametric test will be more likely to lead to a rejection of $\mathcal{H}_0$, if this rejection is justified. Therefore, when parametric tests are valid, they should be favored over their non-parametric counterparts. Non-parametric tests are used when the conditions for the application of other methods are not met, even after a possible transformation of the variables. They can be used even for very small sample sizes.

Tests Based on the Empirical Distribution Function

3

In this chapter, our aim is twofold: to estimate the distribution of a random variable and to address the testing problems that arise from it. To tackle these issues, we will seek to estimate the distribution function of this variable. We are thus facing a non-parametric statistical problem, which we will try to solve using the notion of empirical distribution function.

3.1 Empirical Distribution Function

Recall that the cumulative distribution function (cdf) of a real-valued random variable X is the function given by

$$F\colon x \in \mathbb{R} \longmapsto \mathbb{P}(X \leqslant x)\,.$$

This function is characteristic of the probability distribution of the random variable. We will therefore try to estimate it by introducing the notion of empirical distribution function.

3.1.1 Quantile Function

Let F be a cumulative distribution function of a random variable X.

Definition 3.1 (Quantile function) *For all probability $p \in [0,1]$, we define the* quantile function *of F as its generalized inverse, i.e.:*

$$F^{\leftarrow}(p) := \inf\{x \in \mathbb{R} \,|\, F(x) \geqslant p\}\,.$$

In other words, the quantile function $F^{\leftarrow}$ returns a threshold value x below which random draws from the given cdf would fall p percent of the time. It is also called the percent-point function or inverse cumulative distribution function.[1]

Roughly speaking, a quantile of order p is a value where the graph of the distribution function crosses (or jumps over) p.

Remark 3.1 If F is a invertible, $F^{\leftarrow}$ is the inverse function F^{-1}.

Exercise 3.1 *Compute $F^{\leftarrow}$ for the Bernoulli distribution of parameter θ.*

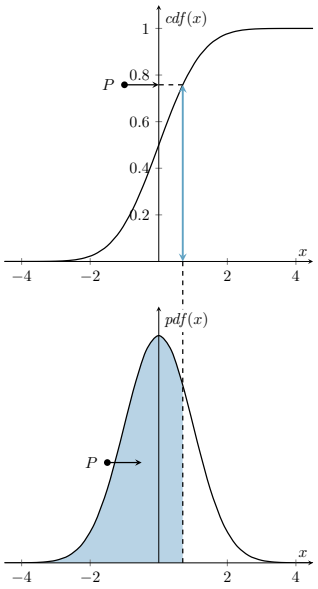


Figure 3.1: Probability density function and cumulative distribution function of the normal distribution

1: Or “inverse généralisé” in French.

The evaluation of quantile functions often involves numerical methods, and there are only a few distributions for which a closed-form expression can be found. When the cdf itself has a closed-form expression, one can use a numerical root-finding algorithm such as the bisection method. Otherwise, several approximation methods have been developed.

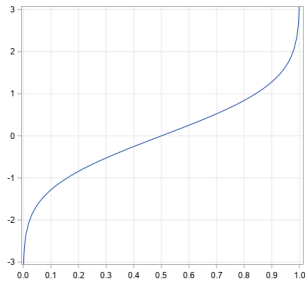


Figure 3.2: Gaussian quantile function

Example 3.1 (Gaussian distribution) The quantile function of the normal distribution itself does not admit a closed-form representation. Indeed, for arbitrary parameters, its quantile function can be derived from a simple transformation of the quantile function of the standard normal distribution, known as the *probit function*.

Proposition 3.2 *Let $x \in \mathbb{R}$, $p \in]0,1[$ and F be a cumulative distribution function.*

1. *F is non-decreasing and right-continuous, which makes it a càdlàg*[2] *function;* $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to+\infty} F(x) = 1$;
2. $\{x \in \mathbb{R} | F(x) \geqslant p\} = [F^{\leftarrow}(p), +\infty[$;
3. *$F^{\leftarrow}$ is non-decreasing;*
4. $F \circ F^{\leftarrow}(p) \geqslant p$, *with equality if* $p \in \mathcal{I}m(F)$;
5. $F(x) \geqslant p$ *if and only if* $x \geqslant F^{\leftarrow}(p)$.

2: *Càdlàg*: French: "continue à droite, limite à gauche".

Remark 3.2 (French vs English) Attention! In English non-decreasing means "croissant" and increasing "strictement croissant". The same goes for non-incresing and decreasing.

Proposition 3.3 *Let X be a random variable with cumulative distribution function F.*

1. *Assume that F is continuous. Then $F(X) \sim \mathcal{U}([0,1])$;*
2. *If $U \sim \mathcal{U}([0,1])$, then $F^{\leftarrow}(U)$ admits F as its cumulative distribution function.*

Proof. Let $U \sim \mathcal{U}([0,1])$. Note that for all $p \in [0,1]$ and $x \in \mathbb{R}$,

$$F^{\leftarrow}(p) \leqslant x \iff p \leqslant F(x).$$

Hence,

$$\mathbb{P}\left[F^{-1}(U) \leqslant x\right] = \mathbb{P}\left[U \leqslant F(x)\right] = F(x),$$

and $X = F^{-1}(U)$ admits F as its cumulative distribution function. □

Proposition 3.3 makes it possible to simulate random variables with a given distribution, as soon as we know how to calculate F^{-1}.

Exercise 3.4 *How to simulate a random variable distributed according to the exponential distribution of parameter λ? A Bernoulli random variable of parameter θ?*

3.1.2 Empirical Distribution Function

Let a n-sample $(X_1, X_2, \dots X_n)$ of real i.i.d. random variables whose cumulative distribution function is given by F.

Definition 3.2 (Empirical distribution function) *We call* empirical distribution function *associated to the n-sample $(X_1, X_2, \dots X_n)$ the function*

$$\begin{aligned} \hat{F}_n : \mathbb{R} &\longrightarrow [0,1] \\ x &\longmapsto \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \leqslant x} \,. \end{aligned}$$

Remark 3.3 (Order statistics) The empirical distribution function of $(X_1, X_2, \dots X_n)$ can also be expressed from the order statistics $(X_{(1)} \leqslant X_{(2)} \leqslant \dots \leqslant X_{(n)})$, where

$$\{X_{(1)}, X_{(2)}, \dots, X_{(n)}\} = \{X_1, X_2, \dots X_n\} \,.$$

We have:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_{(i)} \leqslant x} \,.$$

This makes it easy to plot its graph: The function $\hat{F}_n$ is trivially an non-decreasing step function, discontinuous at the points $(X_{(i)})_{i \in [\![1,n]\!]}$ and constant on $[X_{(i)}, X_{(i+1)}[$ for all $i \in [\![1, n-1]\!]$ (*cf.* Figure 3.3).

Be careful! The variables X_i are random. Therefore, to plot the graph of the empirical distribution function, we first need to observe a realization of these random variables. We refer to it as the *observed* empirical distribution function.

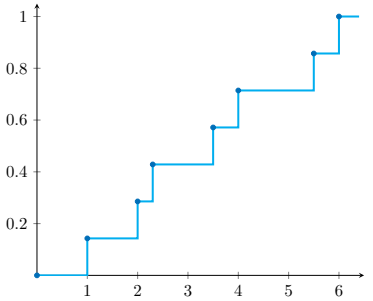


Figure 3.3: Empirical distribution function of the observed sample (2, 3.5, 1, 4, 2.3, 6, 5.5)

Proposition 3.5 *Let $x \in \mathbb{R}$.*

1. *$\hat{F}_n$ is càdlàg, non-decreasing, $\lim_{x \to -\infty} \hat{F}_n(x) = 0$, $\lim_{x \to +\infty} \hat{F}_n(x) = 1$;*
2. *$n\hat{F}_n(x)$ follows a binomial distribution with parameter $(n, F(x))$.*

Link with the cumulative distribution function.

Let $\hat{F}_n$ be the empirical distribution function associated with the n-sample $(X_1, X_2, \dots X_n)$ of cumulative distribution function F. As its name suggests, the function $\hat{F}_n$ is a natural estimator of F; the following proposition makes this idea explicit.

Proposition 3.6 *Let $x \in \mathbb{R}$.*

1. *$\hat{F}_n(x)$ is an unbiased estimator of $F(x)$, $\mathbb{E}\left[\hat{F}_n(x)\right] = F(x)$,*

$$\text{and} \qquad \mathcal{V}ar\left(\hat{F}_n(x)\right) = \frac{F(x)\,(1-F(x))}{n} \underset{n\to+\infty}{\longrightarrow} 0\,;$$

2. $\hat{F}_n(x) \underset{n\to+\infty}{\overset{\mathbb{P}}{\longrightarrow}} F(x)\,;$
3. *For all $x \in \mathbb{R}$ such that $F(x)\,(1-F(x)) \neq 0$,*

$$\sqrt{n}\left(\hat{F}_n(x) - F(x)\right) \underset{n\to+\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}\left(0, F(x)\,(1-F(x))\right)\,.$$

Proof. 1. $\mathcal{V}ar\left(n\hat{F}_n(x)\right) = nF(x)\,(1-F(x))\,;$

2. By Chebichev's inequality,

$$\forall \varepsilon > 0 \quad \mathbb{P}\left(|\hat{F}_n(x) - F(x)| \geqslant \varepsilon\right) \leqslant \frac{1}{\varepsilon^2}\mathcal{V}ar\left(\hat{F}_n(x)\right) \underset{n\to+\infty}{\longrightarrow} 0\,;$$

3. The last statement follows from the central limit theorem.

□

We are now no longer interested in simple pointwise convergence of Fn to F but in uniform convergence.

Theorem 3.7 (Glivenko–Cantelli)

$$\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| \underset{n\to+\infty}{\overset{a.s.}{\longrightarrow}} 0\,.$$

Proof. The strong law of large numbers applied to i.i.d random variables $\mathbb{1}_{X_i \leqslant x}$ (bounded and therefore integrable) such that $\mathbb{E}[\mathbb{1}_{X_i\leqslant x}] = \mathbb{P}(X_i \leqslant x) = F(x)$ leads to the almost sure convergence of $\hat{F}_n$ to F. It remains to prove that the convergence is uniform in x. Let $n \in \mathbb{N}^*$, we pose:

$$D_n = \sup_{x\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i\leqslant x} - F(x) \right|\,.$$

Let $(U_n))_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables of law $\mathcal{U}([0,1])$. We then have the following equations,

in law:

$$D_n = \sup_{x\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{F^{-1}(U_i)\leqslant x\}} - F(x) \right| = \sup_{x\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{U_i\leqslant F(x)\}} - F(x) \right|$$
$$= \sup_{y\in F(\mathbb{R})} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{U_i\leqslant y\}} - y \right| \leqslant \sup_{y\in[0,1]} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{U_i\leqslant y\}} - y \right|.$$

Therefore, it is sufficient to prove that the theorem is true in the special case where the variables $X_i = U_i$ follow uniform laws on $[0,1]$. Thanks to the law of large numbers, we know that for any $y \in \mathbb{R}$, there exists a negligible set $N_y \subset \Omega$ satisfying:

$$\forall \omega \in \Omega \setminus N_y, \quad \sum_{i=1}^{n} \mathbb{1}_{\{U_i\leqslant y\}} \to y\,.$$

A countable union of negligible sets being negligible, we deduce the existence of of a negligible subset $N \subset \Omega$ such that

$$\forall \omega \in \Omega \setminus N, \quad \forall y \in [0,1]\cap\mathbb{Q}, \quad \sum_{i=1}^{n} \mathbb{1}_{\{U_i\leqslant y\}} \to y\,.$$

And using the growth of $y \mapsto \sum_{i=1}^{n} \mathbb{1}_{\{U_i\leqslant y\}}$, it follows that:

$$\forall \omega \in \Omega \setminus N, \quad \forall y \in [0,1], \quad \sum_{i=1}^{n} \mathbb{1}_{\{U_i\leqslant y\}} \to y\,.$$

In other words, for any $\omega \in \Omega \setminus N$, $\sum_{i=1}^{n} \mathbb{1}_{\{U_i\leqslant y\}}$ converges pointwise toward y on $[0,1]$. The Dini theorem then ensures uniform convergence (on the compact [0,1], the functions under consideration being continuous and monotone). □

The Glivenko-Cantelli theorem expresses the extent to which a probability law can be revealed by the knowledge of a large sample of this probability law. In other words, it is a generalization of the strong law of large numbers to the non-parametric case.

3.2 Kolmogorov Adequacy Test

Let X be a random variable whose cumulative distribution function F is assumed to be *continuous*. Let $X_1, \ldots, X_n$ be real i.i.d. random variables with distribution function F. In particular, they have the same distribution as X.

Let a cumulative distribution function F_0 also supposed to be continuous on $\mathbb{R}$ and Y a real random variable of distribution function F_0. We aim to construct a test of $\mathcal{H}_0$: "X and Y have the same distribution: $F = F_0$" against

$\mathcal{H}_1$: "X and Y do not follow the same distribution: $F \neq F_0$";
$\mathcal{H}_1^+$: "X tends to take smaller values than Y: $F \geqslant F_0$";
$\mathcal{H}_1^-$: "X tends to take larger values than Y: $F \leqslant F_0$".

Example 3.2 (Bulb life) We measure the lifetimes of 20 bulbs of the same type. The results, in hours, are: 673, 389, 1832, 570, 522, 2694, 3683, 644, 1531, 2916. Can we affirm, with a 5% risk, that the life of a bulb of this type does not follow the exponential law $\mathcal{E}xp(1/1500)$?

We model the life of the i-th bulb by X_i, F is its unknown distribution function, and F_0 is the distribution function of the law $\mathcal{E}xp(1/1500)$.

Based on the Glivenko-Cantelli theorem, the key idea of the Kolmogorov test is to estimate the unknown distribution function F by the empirical distribution function $\hat{F}_n$ of the sample $(X_1, \ldots, X_n)$ and to compare this empirical distribution function with the given cumulative distribution function F_0.

Definition 3.3 (Kolmogorov test) *The* Kolmogorov test *is defined by the test statistic*

$$D_n = D_n(\hat{F}_n, F_0) = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)| .$$

It consists of rejecting the $\mathcal{H}_0$ hypothesis if $D_n \geqslant d_{n,1-\alpha}$. In other words, its rejection region at level α is of the form $\mathcal{R}_\alpha = \{D_n \geqslant d_{n,1-\alpha}\}$.

Proposition 3.8 *The distribution of D_n under the hypothesis $\mathcal{H}_0$: "$F = F_0$" is independent of F_0.*

Exercise 3.9 *Using the growth of F, prove Proposition 3.8.*

This result is of major practical importance! Indeed, it ensures that for F *continuous*, we can tabulate the distribution of D_n. See the next paragraph.

Theorem 3.10 (admitted) below provides a generalization of the central limit theorem. It allows to check the convergence of the test statistic of the Kolmogorov test.

Theorem 3.10 *For all $\lambda \in \mathbb{R}_+^*$,*

- *Smirnov (1942):* $\mathbb{P}_{\mathcal{H}_0}(\sqrt{n} D_n^+ \leqslant \lambda) \underset{n \to +\infty}{\longrightarrow} \mathrm{e}^{-2\lambda^2}$;
- *Kolmogorov (1933):* $\mathbb{P}_{\mathcal{H}_0}(\sqrt{n} D_n \geqslant \lambda) \underset{n \to +\infty}{\longrightarrow} 2 \sum_{k=1}^{+\infty} (-1)^{k+1} \mathrm{e}^{-2k^2\lambda^2}$;
- *Massart (1990):* $\mathbb{P}_{\mathcal{H}_0}(\sqrt{n} D_n \leqslant \lambda) \leqslant 2\,\mathrm{e}^{-2\lambda^2}$.

In other words, under $\mathcal{H}_0$, D_n will approach 0 at a $\frac{1}{\sqrt{n}}$ rate when $n \to +\infty$.[3] Moreover, the Kolmogorov test is consistent against all

3: For comparison, this is the same rate of convergence as that of the empirical mean to the true mean in the case of an i.i.d. square-integrable sample.

alternatives:

Corollary 3.11 *A test of $\mathcal{H}_0$: "$F = F_0''$, based on Kolmogorv procedure is consistent against all alternatives $\mathcal{H}_1$: "$F \neq F_0''$ as n go to $+\infty$, i.e. $\mathcal{H}_0$ will be (correctly) rejected.*

Proof. ▸ Under $\mathcal{H}_1$: "$F \neq G$", by the Glivenko–Cantelli theorem, the statistic D_n will converge toward $\sup_{x\in\mathbb{R}}|F(x) - G(x)| > 0$, so $\sqrt{n}D_n$ will become infinite. Hence, for all $\lambda \in \mathbb{R}_+^*$,

$$\mathbb{P}_{\mathcal{H}_1}(\sqrt{n}D_n \geqslant \lambda) \underset{n\to+\infty}{\longrightarrow} 1\,.$$

▸ Whereas, under $\mathcal{H}_0$: "$F = G$", by the Kolmogorov theorem (Theorem 3.10), the probability $\mathbb{P}_{\mathcal{H}_1}(\sqrt{n}D_n \geqslant \lambda)$ is small for λ "large".[4] □

4: For $\lambda = 1.36$ the right therm of Kolmogorov formula is equal to 0.05.

Methodological considerations

Remark 3.4 Let $X_{(1)} \leqslant \ldots \leqslant X_{(n)}$ be the ordered sample. We set $X_{(0)} = -\infty$ and $X_{(n+1)} = +\infty$. Since $\hat{F}_n$ is a step function and F_0 is non-decreasing, the maximum gap between $\hat{F}_n$ and F_0 is reached in one of the jumps of $\hat{F}_n$ (See Figure 3.4 for an illustration). Thus,

$$D_n = \max_{i\in[\![0,n]\!]} \left\{ \max\left(\left| \frac{i}{n} - F_0(X_{(i)}) \right| ; \left| \frac{i}{n} - F_0(X_{(i+1)}) \right| \right) \right\},$$

which makes it easy to compute D_n.

The distribution of D_n under $\mathcal{H}_0$ is tabulated. We find in the tables the quantiles $d_{n,1-\alpha}$ such that

$$\mathbb{P}_{\mathcal{H}_0}(D_n \geqslant d_{n,1-\alpha}) \leqslant \alpha\,,$$

(being as close as possible to α). These tables are obtained from simulations of D_n, under the assumption that the X_i are i.i.d. samples from the uniform distribution $\mathcal{U}([0,1])$, *i.e.* $F_0 = \mathbb{1}_{[0,1]}$. The independence of D_n in F_0 (Proposition 3.8) is crucial for the construction of these tables. Indeed, if this were not the case, we would have to construct a table for each possible F_0 distribution.

This test is asymptotically of level α and its power tends to 1 when n tends to $+\infty$.

One-sided test.

In the same way, to test:

▸ $\mathcal{H}_0$: "$F = F_0$" against $\mathcal{H}_1^+$: "$F \geqslant F_0$", we use the test statistic

$$D_n^+ = \sup_{x\in\mathbb{R}} \left(\hat{F}_n(x) - F_0(x) \right),$$

and we reject $\mathcal{H}_0$ if $D_n^+ \geqslant d_{n,1-\alpha}^+$;

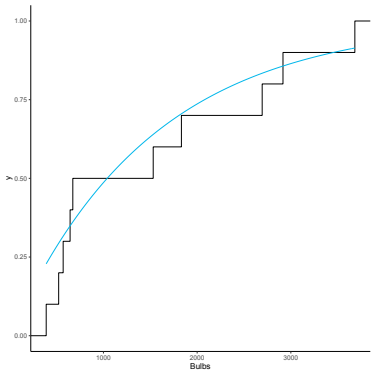


Figure 3.4: Empirical distribution function and cumulative distribution function of the $\mathcal{E}xp(1/1500)$ distribution for the example of light bulbs.

Listing 3.1: Kolmogorov test for the study of bulb life.

```
> Bulbs = c(673, 389, 1832,
    570, 522, 2694, 3683,
    644, 1531, 2916)

> ggplot(data.frame(Bulbs),
    aes(Bulbs))
+ stat_ecdf(geom = "step")
+ stat_function(fun = pexp,
    args = list(rate = 1/
    1500))

> ks.test(Bulbs,pexp,1/1500,
    alternative="two.sided")

    One-sample Kolmogorov-
    Smirnov test

data:  Bulbs
D = 0.22843, p-value = 0.597
alternative hypothesis: two-
    sided
```

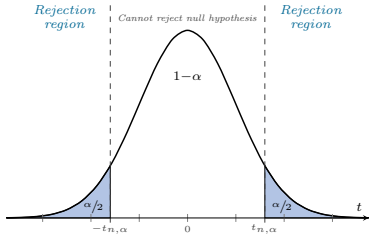


Figure 3.5: Student's t-test.

- $\mathcal{H}_0$: "$F = F_0$" against $\mathcal{H}_1^-$: "$F \leqslant F_0$", we use the test statistic

$$D_n^- = \sup_{x\in\mathbb{R}} \left(F_0(x) - \hat{F}_n(x)\right),$$

  and we reject $\mathcal{H}_0$ if $D_n^- \geqslant d_{n,1-\alpha}^-$.

Likewise, the quantiles are read from the tables.

Example 3.3 (Bulb life, continuation of Example 3.2) The empirical distribution function and the distribution function of the distribution $\mathcal{E}xp(1/1500)$ are shown in Figure 3.4. We establish a Kolmogorov test to test $\mathcal{H}_0$: "$F = F_0$" against $\mathcal{H}_1$: "$F \neq F_0$", where F_0 is the distribution function of the exponential distribution $\mathcal{E}xp(1/1500)$, and using the `ks.test` function (See Listing 3.1).

The p-value being 0.597, we do not reject the null hypothesis at the 5% risk level.

Other tests based on the empirical distribution function.

There are other tests based on the empirical distribution function. The Cramer Von Mises test uses for example the statistic

$$C_n = n\int_{-\infty}^{+\infty} \left(\hat{F}_n(x) - F_0(x)\right)^2 f_0(x)\,\mathrm{d}x,$$

and the Anderson Darling test uses the test statistic

$$A_n = n\int_{-\infty}^{+\infty} \left(\hat{F}_n(x) - F_0(x)\right)^2 \frac{f_0(x)}{F_0(x)(1-F_0(x))}\,\mathrm{d}x.$$

As for the Kolmogorov test, we show that the laws of C_n and A_n are independent of F_0 under $\mathcal{H}_0$, and these laws are therefore tabulated.

3.3 Comparison Tests of Two Samples

In the same spirit, we will construct a homogeneity test. We observe two independent samples of size n and m:

- $X_1, \ldots, X_n$ i.i.d. with cumulative distribution function F;
- $Y_1, \ldots, Y_m$ i.i.d. with cumulative distribution function G.

We want to test if the two samples are from the same distribution. In the case of two Gaussian samples, namely F and G corresponding to the normal distributions $\mathcal{N}(m_X, \sigma_X^2)$ and $\mathcal{N}(m_Y, \sigma_Y^2)$ respectively, we can use a Student's t-test to distinguish between $\mathcal{H}_0 : F = G$ and $\mathcal{H}_1 : F \neq G$ (see the 3rd year course). We do not return to this framework here and place ourselves in a non-parametric setting. In other words, we do not assume anymore that we know the laws of the variables X_i and Y_j.

3.3.1 Kolmogorov-Smirnov Test

In this section, we aim to test $\mathcal{H}_0 : F = G$ against $\mathcal{H}_1 : F \neq G$. We note $\hat{F}_n$ the empirical distribution function of the sample $(X_1, \dots, X_n)$ and $\hat{G}_m$ the one of the sample $(Y_1, \dots, Y_m)$.

Definition 3.4 (Kolmogorov-Smirnov test) *The* Kolmogorov-Smirnov test *is defined by the test statistic*

$$D_{(n,m)} = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - \hat{G}_m(x)| .$$

It consists of rejecting the $\mathcal{H}_0$ hypothesis if $D_{(n,m)} \geqslant d_{n,m,1-\alpha}$. In other words, its rejection region at level α is of the form $\mathcal{R}_\alpha = \{D_{(n,m)} \geqslant d_{n,m,1-\alpha}\}$.

We preserve the existence of tables as for the Kolmogorov test as shown in Proposition 3.12.

Proposition 3.12 . *If F is continuous, the distribution of $D_{(n,m)}$ under the null hypothesis $F = G$ is independent of F.*

This distribution is therefore tabulated.

Proof. Proceed as for Proposition 3.8. □

By the Glivenko–Cantelli theorem, if $\mathcal{H}_0$ holds, then $D_{(n,m)} \to 0$ almost surely as m and n both go to $+\infty$. Therefore, we can hope that test of $\mathcal{H}_0$ based on a suitable multiple of $D_{(n,m)} \to 0$ will have the same, or at least a similar, asymptotic distribution as $\sqrt{n} D_n$. More precisely, the correct convergence rate is $\sqrt{\frac{nm}{n+m}}$, and we have the following uniform convergence result.

Theorem 3.13 *For all $\lambda \in \mathbb{R}_+^*$, under $\mathcal{H}_0$: "$F = G$"*

- *If $F = G$ is continuous,*

$$\lim_{m,n \to +\infty} \mathbb{P}_{\mathcal{H}_0}\left(\sqrt{\frac{nm}{n+m}} D_{(n,m)} \geqslant \lambda\right) = 2 \sum_{k=1}^{+\infty} (-1)^{k+1} \mathrm{e}^{-2k^2\lambda^2} ;$$

- *If $F = G$ is* not *continuous,*

$$\limsup_{m,n \to +\infty} \mathbb{P}_{\mathcal{H}_0}\left(\sqrt{\frac{nm}{n+m}} D_{(n,m)} \geqslant \lambda\right) \leqslant 2 \sum_{k=1}^{+\infty} (-1)^{k+1} \mathrm{e}^{-2k^2\lambda^2} .$$

Corollary 3.14 *A test of $\mathcal{H}_0$: "$F = G$", based on Kolmogorv-Smirnov procedure is consistent against all alternatives $\mathcal{H}_1$: "$F \neq G$" as m and n both go to $+\infty$, i.e. $\mathcal{H}_0$ will be (correctly) rejected.*

Proof. Proceed as for Corollary 3.11. □

One-sided test.

To do a one-sided test ($\mathcal{H}_0 : F = G$ vs. $\mathcal{H}_1 : F \geqslant G$), we use the test statistic

$$D^+_{(n,m)} = \sup_{x \in \mathbb{R}} \left(\hat{F}_n(x) - \hat{G}_m(x) \right),$$

associated with the rejection region $\mathcal{R}_\alpha = \{D^+_{(n,m)} \geqslant d^+_{n,m,1-\alpha}\}$.

Table 3.1: *Pain relief.* Time (in hours) between taking the drug and feeling relief.

Drug A	Drug B
6,8	4,4
3,1	2,5
5,8	2,8
4,5	2,1
3,3	6,6
4,7	1,5
4,2	4,8
4,9	2,3

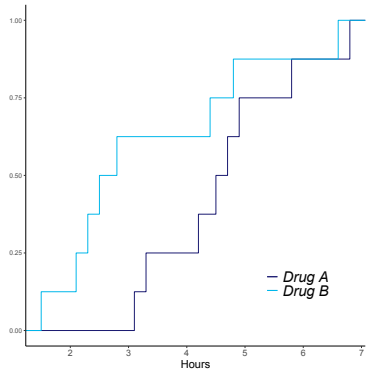


Figure 3.6: Empirical distribution function for drugs A and B.

Example 3.4 (Comparative of analgesics) We would like to compare two drugs for postoperative pain relief. We observed 16 patients, 8 of whom took the usual drug A, and the other 8 an experimental drug B. In Table 3.1, the time (in hours) between the taking of the drug and the feeling of relief is reported. The empirical distribution functions of the two samples are shown in Figure 3.6.

- Is there a difference in efficiency between the two drugs?

  To answer this question, we test $\mathcal{H}_0 : F_A = F_B$ against $\mathcal{H}_1 : F_A \neq F_B$, where F_A and F_B are the cumulative distribution functions associated with samples A and B respectively.

```
> ks.test(dB, dA, alternative="two.sided")

    Two-sample Kolmogorov-Smirnov test

data: dB and dA
D = 0.625, p-value = 0.08702
alternative hypothesis: two-sided
```

- Is drug B more effective than drug A?

  We now test $\mathcal{H}_0 : F_A = F_B$ against $\mathcal{H}_1 : F_B \geqslant F_A$.

```
> ks.test(dB, dA, alternative="greater")

    Two-sample Kolmogorov-Smirnov test

data: dB and dA
D^+ = 0.625, p-value = 0.04394
alternative hypothesis: the CDF of x lies above that of y
```

3.3.2 Wilcoxon-Mann-Whitney Test

In this section, we will focus on the Wilcoxon-Mann-Whitney tests based on ranks. This is a test of whether two samples come from the same distribution, against the alternative that members of one sample tend to be larger than those of the other sample (a location or shift alternative). No parametric form of the distributions is assumed. They can be quite general, as long as the distribution functions are continuous. There are two formulations of the test, one due to Mann and Whitney and the other to Wilcoxon; R uses the Mann–Whitney form.

To simplify the presentation, we will first assume that there are no match in the two samples:

- the X_i are all distinct,

- the Y_j are all distinct,
- the X_i are distinct from the Y_j.

We will come back to the case of ex-aequo in Section 3.3.2.3.

3.3.2.1 Mann-Whitney U Test

The hypothesis to be tested, as in the Kolmogorov–Smirnov test, is $\mathcal{H}_0$: "$F = G$". The principle of the Mann-Whitney test is to determine the number of pairs (X_i, Y_j) for which $Y_j > X_i$. Then $\mathcal{H}_0$ will be rejected if either this count is too large, indicating that the X's tend to be less than the Y's, or if the count is too small, indicating that the Y's tend to be less than the X's.

Suppose we want to test $\mathcal{H}_0$: "$F = G$" against $\mathcal{H}_1^+$: "$F \geqslant G$", and that F and G are continuous. Under $\mathcal{H}_1^+$, for all x,

$$G(x) = \mathbb{P}(Y \leqslant x) \leqslant \mathbb{P}(X \leqslant x) = F(x)\,,$$

with sometimes strict inequality. So, for all x, $\mathbb{P}(Y > x) \geqslant \mathbb{P}(X > x)$, and the number of pairs (X_i, Y_j) for which $Y_j > X_i$ takes larger values under $\mathcal{H}_1^+$ than under $\mathcal{H}_0$.

Definition 3.5 (Mann-Whitney U test) *The* Mann-Whitney test *for $\mathcal{H}_0$: "$F = G$" vs. $\mathcal{H}_1^+$: "$F \geqslant G$" is the test defined from the statistic*

$$U_{(n,m)}^{X<Y} = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{1}_{X_i<Y_j}\,.$$

The test consists of rejecting $\mathcal{H}_0$ if $U_{(n,m)}^{X<Y} \geqslant u_{(n,m),1-\alpha}^{X<Y}$. In other words, its rejection region at level α is of the form $\mathcal{R}_\alpha = \left\{U_{(n,m)}^{X<Y} \geqslant u_{(n,m),1-\alpha}^{X<Y}\right\}$.

The law of $U_{(n,m)}^{X<Y}$ under $\mathcal{H}_0$ can be established by recurrence (cf [2, p. 126]). We note:

$$\forall k \in [\![1, mn]\!], \quad p_{(n,m)}(k) = \mathbb{P}_{\mathcal{H}_0}(U_{(n,m)}^{X<Y} = k)\,,$$

$$p_{(n,0)}(0) = p_{(0,m)}(0) = 1\,, \quad \text{and} \quad \forall k \in \mathbb{N}^*,\ p_{(n,0)}(k) = p_{(0,m)}(k) = 0\,.$$

Then for all k,

$$(n+m)p_{(n,m)}(k) = np_{(n-1,m)}(k) + mp_{(n,m-1)}(k-n)\,.$$

This recurrence formula allows to compute the law of $U_{(n,m)}^{X<Y}$ under the null hypothesis $\mathcal{H}_0$. In other words, for m and n not too large, one can tabulate the distribution.

For large n and m, we can also use the (admitted) following asymptotic result:

Theorem 3.15 (Hajek (1968)) *Under* $\mathcal{H}_0$,

$$\frac{U_{(n,m)}^{X<Y} - \mathbb{E}_{\mathcal{H}_0}\left[U_{(n,m)}^{X<Y}\right]}{\sqrt{\mathcal{V}ar_{\mathcal{H}_0}\left(U_{(n,m)}^{X<Y}\right)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)\,,$$

when

$$n \to +\infty \quad \textit{and} \quad \frac{n}{n+m} \to \lambda \in]0,1[\,.$$

In practice, we use this result if $n, m \geqslant 8$. Under assumption $\mathcal{H}_0$, it comes

$$\mathbb{E}_{\mathcal{H}_0}\left[U_{(n,m)}^{X<Y}\right] = \frac{mn}{2} \quad \text{and} \quad \mathcal{V}ar_{\mathcal{H}_0}\left(U_{(n,m)}^{X<Y}\right) = mn\left(\frac{n+m+1}{12}\right).$$

A similar reasoning can be used to test $\mathcal{H}_0$: "$F = G$" against $\mathcal{H}_1^-$: "$F \leqslant G$". In this case the number of pairs (X_i, Y_j) for which $Y_j < X_i$ takes larger values under $\mathcal{H}_1^-$ than under $\mathcal{H}_0$.

Definition 3.6 (Mann-Whitney U test) *The* Mann-Whitney test *for* $\mathcal{H}_0$*: "*$F = G$*" vs.* $\mathcal{H}_1^-$*: "*$F \leqslant G$*" is the test defined from the statistic*

$$U_{(n,m)}^{X>Y} = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{1}_{X_i > Y_j}\,.$$

Its rejection region at level α *is of the form* $\mathcal{R}_\alpha = \left\{U_{(n,m)}^{X>Y} \geqslant u_{(n,m),1-\alpha}^{X>Y}\right\}$.

The $U_{(n,m)}^{X>Y}$ statistic checks properties similar to those of $U_{(n,m)}^{X<Y}$ seen above. Moreover, since there are mn total pairs, $U_{(n,m)}^{X>Y} + U_{(n,m)}^{X<Y} = mn$.

Finally, in the case of a two-sided test of $\mathcal{H}_0$: "$F = G$" versus $\mathcal{H}_1$: "$F \neq G$", we combine the two previous tests.

Definition 3.7 (Mann–Whitney U test) *The* Mann-Whitney test *for* $\mathcal{H}_0$*: "*$F = G$*" vs.* $\mathcal{H}_1$*: "*$F \neq G$*" is the test defined from the statistic*

$$U_{(n,m)} = \max\left(U_{(n,m)}^{X<Y}, U_{(n,m)}^{X>Y}\right).$$

Its rejection region at level α *is of the form* $\mathcal{R}_\alpha = \left\{U_{(n,m)} \geqslant u_{n,m,1-\alpha}\right\}$.

In particular, $U_{(n,m)} = \min\left(U_{(n,m)}^{X<Y}, mn - U_{(n,m)}^{X<Y}\right)$

3.3.2.2 Wilcoxon Rank-Sum Test

Let's come back to the $\mathcal{H}_0$: "$F = G$" versus $\mathcal{H}_1^+$: "$F \geqslant G$". There is another equivalent form of the Mann-Whitney U test, called the Wilcoxon rank-

sum test.

Let $N = n + m$ and

$$Z = (Z_1, \ldots, Z_n, Z_{n+1}, \ldots, Z_N) = (X_1, \ldots, X_n, Y_1, \ldots, Y_m)$$

the full sample. We define $(R_1, \ldots, R_m)$, where R_j is the rank of Y_j in the ordered full sample

$$R_j \;=\; 1 + \sum_{k=1}^{N} \mathbb{1}_{Z_k < Y_j} \,.$$

Definition 3.8 (Wilcoxon rank-sum test) *The* Wilcoxon *statistic consists in computing the sum of the ranks of the individuals in the second sample*

$$W^Y_{(n,m)} \;=\; \sum_{j=1}^{m} R_j \,.$$

Since we have the relation

$$U^{X<Y}_{(n,m)} \;=\; W^Y_{(n,m)} - \frac{m(m+1)}{2} \,,$$

both statistics lead to the same test.

Similarly, one can construct the Wilcoxon test statistic $W^X_{(n,m)}$, sum of ranks of X_i in Z, related to the test statistic $U^{X>Y}_{(n,m)}$ to test $\mathcal{H}_0$: "$F = G$" versus $\mathcal{H}_1^-$: "$F \leqslant G$".

Finally, note that we have the following relationship:

$$W^X_{(n,m)} + W^Y_{(n,m)} \;=\; \sum_{k=1}^{N} k \;=\; \frac{N(N+1)}{2} \,.$$

Example 3.5 (Comparative of analgesics, continued from Example 3.4) We continue with the study of analgesics. We want to test if drug B is more effective than drug A, *i.e.* $\mathcal{H}_0$: "$F_A = F_B$" versus $\mathcal{H}_1$: "$F_B \geqslant F_A$".

We then observe the complete ordered sample

$$\begin{aligned} z \;=\; & (1.5,\ 2.1,\ 2.3,\ 2.5,\ 2.8,\ 3.1,\ 3.3,\ 4.2,\\ & \qquad 4.4,\ 4.5,\ 4.7,\ 4.8,\ 4.9,\ 5.8,\ 6.6,\ 6.8)\\ \;=\; & (B_6,\ B_4,\ B_8,\ B_2,\ B_3,\ A_2,\ A_5,\ A_7,\\ & \qquad B_1,\ A_4,\ A_6,\ B_7,\ A_8,\ A_3,\ B_5,\ A_1) \end{aligned}$$

The observed ranks for the values of B are therefore

$$R_1 = 9,\ R_2 = 4,\ R_3 = 5,\ R_4 = 2,\ R_5 = 15,\ R_6 = 1,\ R_7 = 12,\ R_8 = 3,$$

which leads to $W^B_{(8,8)} = 51$ and $W^A_{(8,8)} = \frac{16\times17}{2} - 51 = 85$.

Moreover,

$$U^{B<A}_{(8,8)} = \sum_{i=1}^{8}\sum_{j=1}^{8} \mathbb{1}_{B_i<A_j}$$
$$= 5+5+5+6+6+7+7+8 = 49 = W^A_{(8,8)} - \frac{8\times9}{2},$$

and

$$U^{B>A}_{(8,8)} = \sum_{i=1}^{8}\sum_{j=1}^{8} \mathbb{1}_{B_i>A_j}$$
$$= 3+3+3+2+2+1+1+0 = 15 = W^B_{(8,8)} - \frac{8\times9}{2}.$$

An effective way of displaying the data is to use a table as shown in Table 3.2. For example, it is easy to read the rank of the different values, and in the last column that N=16.

Listing 3.2: Wilcoxon rank-sum test

```
> wilcox.test(dB,dA,
    alternative="less")

    Wilcoxon rank sum exact
    test

data:  mB and mA
W = 15, p-value = 0.04149
alternative hypothesis: true
    location shift is less
    than 0
```

Table 3.2: Effective presentation of data in tabular form.

$A_{(.)}$						3.1		3.3		4.2			4.5		4.7			4.9		5.8			6.8
$B_{(.)}$	1.5		2.1		2.3		2.5		2.8					4.4					4.8				6.6	
R_i						6		7		8			10		11			13		14			16
R_j	1		2		3		4		5					9					12				15	

Remark 3.5 (Sample size) The theoretical guarantees available for the Kolmogorov-Smirnov test are asymptotic. Therefore, this test is a priori valid only for "large" sample sizes. Here, we have guarantees even for very small sample sizes. Thus, in the case of small sample sizes, a Wilcoxon-Mann-Whitney test is preferred.

And in the case of a very large sample size? The central limit theorem ensures that the samples are then essentially Gaussian, and we can use the Student's t-test with a controlled margin of error.

3.3.2.3 Treatment of Ex-Aequos

We have assumed that the laws are continuous, so the probability of having a tie is zero. In practice, either because the laws are not continuous, or because we have rounded measures, we can have ex-aequos. In this case, we can "modify" the Mann-Whitney test statistics as follows:

$$\tilde{U}^{X<Y}_{(n,m)} = \sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\mathbb{1}_{X_i<Y_j} + \frac{1}{2}\mathbb{1}_{X_i=Y_j}\right\}$$

and

$$\tilde{U}^{X>Y}_{(n,m)} = \sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\mathbb{1}_{X_i>Y_j} + \frac{1}{2}\mathbb{1}_{X_i=Y_j}\right\}.$$

Note that $\tilde{U}^{X<Y}_{(n,m)} + \tilde{U}^{X>Y}_{(n,m)} = nm$.

For the Wilcoxon test, we use the mean ranks. It consists in assigning to all the elements of a group of ex-aequos the average rank of the elements of the group. We thus correct the R_j defined previously.

Example 3.6 We consider the following observed values for the two samples

$$x = (5,\ 3,\ 6,\ 8,\ 1,\ 6) \quad \text{and} \quad y = (5,\ 7,\ 9,\ 5,\ 2)\,.$$

In particular $n = 5$, $m = 6$, and we obtain the following table of ordered values and ranks:

$x_{(.)}$	1		3		5		6	6		8	
$y_{(.)}$		2		5		5			7		9
$\tilde{R}_i$	1		3		5		7.5	7.5		10	
$\tilde{R}_j$		2		5		5			9		11

Therefore,

$$\tilde{U}^{X<Y}_{(n,m)} = 1 + \left(2 + \frac{1}{2}\right) + \left(2 + \frac{1}{2}\right) + 5 + 6 = 17\,,$$

$$\tilde{W}^{Y}_{(n,m)} = \sum_{j=1}^{m} \tilde{R}_j = 2 + 5 + 5 + 9 + 11 = 32\,,$$

and we still check that $\tilde{U}^{X<Y}_{(n,m)} = \tilde{W}^{Y}_{(n,m)} - \frac{5\times 6}{2}$.

If any of the variables are tied, R gives a warning message saying p-values are not exact. Overlaps within variables are not too "critical" from a statistical point of view. However, ties $X_i = Y_j$ for some i and j are a more severe problem as the value of the statistic becomes uncertain: it can be affected by arbitrarily small changes in X_i or Y_j.

3.3.3 Median test

We want to test $\mathcal{H}_0$: "$F = G$" against $\mathcal{H}_1^+$: "$F \geqslant G$", and we assume that F and G are continuous. The principle of the median test is to determine the number of variables in the second sample that are greater than the median of all observations. We note $N = n + m$.

Definition 3.9 (Median test) *The* median test *is defined from the statistic*

$$M_{(n,m)} = \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}_{R_j > \frac{N+1}{2}}\,.$$

The rejection region at level α is of the form $\mathcal{R}_\alpha = \left\{M_{(n,m)} \geqslant m_{n,m,1-\alpha}\right\}$.

Example 3.7 (Location test) Let $X_1, \dots, X_n$ i.i.d. of distribution function F and $Y_1, \dots, Y_m$ i.i.d. according to the distribution function $G = F(\cdot - \mu)$. For example, we study the blood pressure of patients undergoing treatment for hypertension (Y_j), and compare them with untreated patients (X_i). Suppose that after treatment, the blood pressure law is translated by μ. The treatment is effective if $\mu < 0$, it is ineffective if $\mu = 0$.

Distribution of $M_{(n,m)}$ under $\mathcal{H}_0$:

- If N is even, then

$$\forall k \in \left[\!\left[\max\left(0, m - \frac{N}{2}\right), \min\left(m, \frac{N}{2}\right) \right]\!\right],$$
$$\mathbb{P}_{\mathcal{H}_0}\left(mM_{(n,m)} = k\right) = \frac{\binom{N}{m}\binom{N-m}{N/2-k}}{\binom{N}{N/2}}.$$

  Hence, $M_{(n,m)}$ follows a hypergeometric distribution of parameter $(N, {}^N\!/_2, m)$. We deduce that

$$\mathbb{E}_{\mathcal{H}_0}\left[M_{(n,m)}\right] = \frac{1}{2} \quad \text{and} \quad \mathcal{V}ar\left(M_{(n,m)}\right) = \frac{n}{4m(N-1)}.$$

- If N is odd, then

$$\forall k \in \left[\!\left[\max\left(0, m - \frac{N+1}{2}\right), \min\left(m, \frac{N-1}{2}\right) \right]\!\right],$$
$$\mathbb{P}_{\mathcal{H}_0}\left(mM_{(n,m)} = k\right) = \frac{\binom{N}{m}\binom{N-m}{(N-1)/2-k}}{\binom{N}{(N-1)/2}}.$$

  Hence, $M_{(n,m)}$ follows a hypergeometric distribution of parameter $\left(N, \frac{N-1}{2}, m\right)$. We deduce that

$$\mathbb{E}_{\mathcal{H}_0}\left[M_{(n,m)}\right] = \frac{N-1}{2N} \quad \text{and} \quad \mathcal{V}ar\left(M_{(n,m)}\right) = \frac{n(N+1)}{4mN^2}.$$

The knowledge of the distribution of $M_{(n,m)}$ under $\mathcal{H}_0$ allows to determine the rejection zone of the test. For $n, m \geqslant 30$, we can approximate the distribution of $M_{(n,m)}$ under $\mathcal{H}_0$ by the distribution $\mathcal{N}\left(\mathbb{E}_{\mathcal{H}_0}\left[M_{(n,m)}\right], \mathcal{V}ar\left(M_{(n,m)}\right)\right)$. We can then use a Fisher test (*cf.* Part II) or a χ^2 test (Chapter 4).

For an example using a chi-square test, see Subsection 4.5.1.

Remark 3.6 The Wilcoxon, Mann-Whitney, and median tests do not test two-sided alternatives.

3.4 Normality Tests

Normality tests are used to determine if a data set is well-modeled by a normal distribution (within some tolerance). These tests are all the more

important as many statistical tests, such as the Student's t-test or ANOVA (*cf.* Part II), require a normally distributed sample population.

In this section, we consider a random variable X with cumulative distribution function F and a n-sample $(X_1, \dots, X_n)$ of the same distribution. We note $\hat{F}_n$ the empirical distribution function associated to this sample.

To illustrate the different methods, we will consider the following three data sets of size $n = 200$:

- `data1`: simulated from the Gaussian distribution $\mathcal{N}(2, 1)$,
- `data2`: simulated from the uniform distribution $\mathcal{U}([2, 4])$,
- `data3`: simulated from the Cauchy distribution $\mathcal{C}(0, 1)$.

Listing 3.3: Different data sets.

```
> n=200
> data1=rnorm(n,2,1)
> data2=runif(n,min=2,max=4)
> data3=rcauchy(n)
```

3.4.1 Normal Probability Plot

The method, also called *quantile-quantile plot* (Q-Q plot),[5] consists of plotting the graph of points $\left(\Phi^{-1} \circ \hat{F}_n(x_{(i)}), x_{(i)}\right)$, where:

- $x_{(1)} \leqslant \dots \leqslant x_{(n)}$ is an ordered realization of the sample $(X_i)_{i \in [\![1,n]\!]}$,
- Φ represents the cumulative distribution function of the standard Gaussian law $\mathcal{N}(0, 1)$,[6]
- and $\hat{F}_n$ is the empirical distribution function associated to the sample $(X_i)_{i \in [\![1,n]\!]}$.

5: Or "doite de Henry" in French.

6: In other words, Φ^{-1} is the standard normal quantile function.

Figure 3.7 shows the Q-Q plot for the three data sets introduced beforehand.

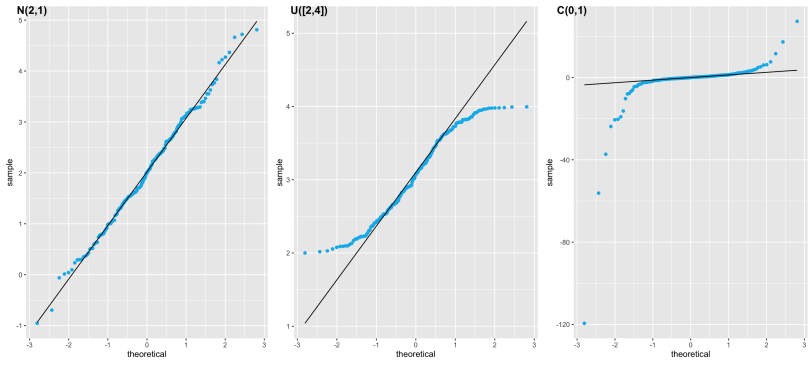


Figure 3.7: Q-Q plot for the 3 data sets: In order, $\mathcal{N}(2, 1)$, $\mathcal{U}([2, 4])$ and $\mathcal{C}(0, 1)$.

3.4.1.1 Principle

Let $X \sim \mathcal{N}(\bar{x}, \sigma^2)$ be a Gaussian variable of mean $\bar{x}$ and variance σ^2. If $N \sim \mathcal{N}(0, 1)$ is a centered normal distribution variable, we have the following equations:

$$\forall x \in \mathbb{R}, \quad \mathbb{P}(X < x) = \mathbb{P}\left(\frac{X - \bar{x}}{\sigma} < \frac{x - \bar{x}}{\sigma}\right) = \mathbb{P}(N < t)\,,$$

where $t = \frac{x - \bar{x}}{\sigma}$.

Hence, for each observation x_i of a variable X, we can compute the probability $\mathbb{P}(X < x_i)$ and, using a table of the Φ function, derive t_i such

that $\phi(t_i) = \mathbb{P}(X < x_i)$. Then, if the variable X is Gaussian, the points of coordinates (x_i, t_i) are practically aligned on the line of equation $t = \frac{x-\bar{x}}{\sigma}$. We can then easily conclude about the normality of a variable but also read its mean and standard deviation in the equation of the line.

In practice, the first step is usually to standardize our observations, *i.e.* to subtract their mean and renormalize them by their standard deviation (statistical software such as R or Python do it automatically).

Now we have to focus on the ends of the curve formed from the points. Suppose that the points at the ends of the curve do not fall on a straight line but are instead very far apart. In this case, we reject with certainty the hypothesis of Gaussianity; in other words, our observations are not normally distributed. On the contrary, if all the points plotted on the graph are perfectly aligned, the assumption of Gaussianity is reasonable. Without being able to assert it completely (it is a purely graphical tool), it is reasonable to assume that our observations are normally distributed.

3.4.1.2 Skewed and Tailed Q-Q Plots

Let standardized observations. Q-Q plots are also used to determine distributions' *skewness* (a measure of "asymmetry").

- If the bottom end of the Q-Q plot deviates from the straight line, but the upper end does not, then we can clearly state that the distribution has a heavier tail to its left, is skewed to the left, or is negatively skewed.
- On the other hand, if the upper end of the Q-Q graph deviates from the straight line but the lower end does not, then we can say that the observed distribution has a heavier tail on its right, or in other words, that it is skewed to the right, or positively skewed.

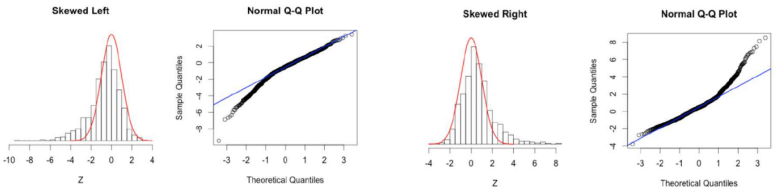


Figure 3.8: Influence of skewness/asymmetry on Q-Q plots

Similarly, we can talk about distributions' *kurtosis* (a measure of the "tailedness") by simply looking at their Q-Q plot. In the case of a thick-tailed distribution, both ends of the Q-Q plot deviate from the straight line while its center follows the line. In contrast, a thin-tailed distribution forms a Q-Q plot with very little or negligible deviation at the ends, which actually makes it a good fit for the normal distribution.

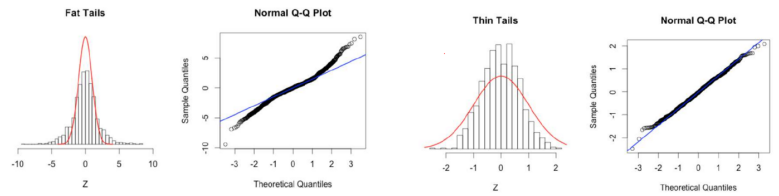


Figure 3.9: Influence of kurtosis/tailedness on Q-Q plots

Remark 3.7 Here we used Q-Q plots to check the fit of our observations to the normal distribution. We can also use such diagrams to evaluate the relevance of the fit of a given distribution to any theoretical model and, more generally, to compare two distributions that we consider similar.

To do this, we compare the position of certain quantiles in the observed population with that in the theoretical population. In the case of a good modeling hypothesis, the points thus generated should be roughly aligned with the first bisector.

3.4.2 Kolmogorov–Smirnov Test – Lilliefors Test

We want to test the null hypothesis

$\mathcal{H}_0$: "X follows a normal distribution",

against the alternative hypothesis

$\mathcal{H}_1$: "X does not follow a normal distribution".

The Lilliefors test is a normality test adapted from the Kolmogorov-Smirnov test to test the normality of a sample when the parameters of the assumed normal distribution are unknown, *i.e.* when neither the expectation μ nor the standard deviation σ are known.

We note

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \,,$$

the *empirical* mean and the *empirical* variance of X respectively.

Definition 3.10 (Kolmogorov–Smirnov test) *The* Kolmogorov normality test *is based on the test statistic*

$$D_n = \sup_{x\in\mathbb{R}} \left| \hat{F}_n(x) - \Phi\left(x; \bar{X}, S_X^2\right) \right| ,$$

where $\Phi\left(x; \bar{X}, S_X^2\right)$ *is the cumulative distribution function of the normal distribution* $\mathcal{N}\left(\bar{X}, S_X^2\right)$.

The reject region at level α *is of the form* $\mathcal{R}_\alpha = \{D_n \geqslant d_{n,1-\alpha}\}$

Proposition 3.16 *Under asummption* $\mathcal{H}_0$*, i.e.* X *follows a normal distribution* $\mathcal{N}(\mu, \sigma^2)$*,*

1. *The distribution of* D_n *does not depend on the unknown parameters* μ *and* σ*;*
2. *The law of* D_n *is then given by:*

$$D_n = \sup_{x\in\mathbb{R}} \left| \hat{\Phi}_n(x) - \Phi\left(\frac{x - \bar{Z}}{S_Z}\right) \right| ,$$

Listing 3.4: Kolmogorov-Smirnov test on the three simulated samples.

```
> library(nortest)
> lillie.test(data1)

    Lilliefors (Kolmogorov-
    Smirnov) normality test

data:  data1
D = 0.043152, p-value =
    0.4835

> lillie.test(data2)

    Lilliefors (Kolmogorov-
    Smirnov) normality test

data:  data2
D = 0.087607, p-value =
    0.00076

> lillie.test(data3)

    Lilliefors (Kolmogorov-
    Smirnov) normality test

data:  data3
D = 0.35484, p-value < 2.2e
    -16
```

where :

- $Z = (Z_1, \ldots, Z_n)$ *is i.i.d of law* $\mathcal{N}(0,1)$,
- $\hat{\Phi}_n$ *the empirical standard distribution function,*
- $\bar{Z}$ *the empirical mean, and*
- S_Z^2 *the empirical variance of* Z.

The distribution of D_n is tabulated (we can for example simulate it with $\mu = 0$ and $\sigma = 1$ to estimate the quantiles). However, until now, the tables for this distribution have been computed only by Monte Carlo methods.

Moreover, since the hypothesised cumulative distribution function has been moved closer to the data by estimation based on those data, the maximum discrepancy has been made smaller than it would have been if the null hypothesis had distinguished only one normal distribution. Thus, the "null distribution" of the test statistic, i.e., its probability distribution assuming the null hypothesis, is stochastically smaller than the classical Kolmogorov-Smirnov distribution. In other words, this test is not very powerful: a large number of observations is required to reject the hypothesis of normality.

Remark 3.8 A variant of the test can be used to test the null hypothesis that data come from an exponentially distributed population, when the null hypothesis does not specify which exponential distribution.

3.4.3 Shapiro-Wilk Test

This is a test based on the L-statistic (linear combination of order statistics), which relies on a comparison of the empirical variance with an estimator of the variance of X that has good properties under the normality assumption.

To date, the Shapiro-Wilk test remains the most efficient test for normality and can handle samples of up to 5000 observations.

3.4.3.1 Estimation of the Mean and Variance using Order Statistics for Symmetric Laws

Let $X_1, \ldots, X_n$ i.i.d. Let $\mu = \mathbb{E}[X_i]$ and $\sigma^2 = \mathcal{V}ar(X_i)$. Let $Y_i = {}^{(X_i - \mu)}/_{\sigma}$. Assume that Y_i is symmetrically distributed, *i.e.* that $-Y_i$ and Y_i have the same distribution. We denote by $X_{(1)} \leqslant \ldots \leqslant X_{(n)}$ the ordered sample of X_i and $Y_{(1)} \leqslant \ldots \leqslant Y_{(n)}$ that of Y_i. In particular,

$$Y_{(i)} = \frac{X_{(i)} - \mu}{\sigma}.$$

For all $i, j \in [\![1, n]\!]$, let $\alpha_i = \mathbb{E}[Y_{(i)}]$ and $B_{i,j} = Cov\big(Y_{(i)}, Y_{(j)}\big)$. We then have

$$X_{(i)} = \mu + \alpha_i \sigma + \varepsilon_i,$$

with $\mathbb{E}[\varepsilon_i] = 0$. Note that the ε_i are not independent: the variance-covariance matrix of the vector $\varepsilon = {}^t(\varepsilon_1 \ \ \varepsilon_2 \ \ \dots \ \ \varepsilon_n)$ is $\sigma^2 B$. Let $\mathbb{1}_n$ and α be the vectors of $\mathbb{R}^n$ defined by

$$\mathbb{1}_n = {}^t(1 \quad 1 \quad \dots \quad 1) \quad \text{and} \quad \alpha = {}^t(\alpha_1 \quad \alpha_2 \quad \dots \quad \alpha_n).$$

We denote A the matrix of size $(n, 2)$ defined by $A = (\mathbb{1}_n, \alpha)$. Finally, we note $X_{(\cdot)} = {}^t(X_{(1)} \quad X_{(2)} \quad \dots \quad X_{(n)})$. We then have the relation

$$X_{(\cdot)} = A \begin{pmatrix} \mu \\ \sigma \end{pmatrix} + \varepsilon .$$

The weighted least squares estimator of (μ, σ) is obtained by minimizing in the parameters (μ, σ) the criterion

$${}^t\left(X_{(\cdot)} - A \begin{pmatrix} \mu \\ \sigma \end{pmatrix}\right) B^{-1} \left(X_{(\cdot)} - A \begin{pmatrix} \mu \\ \sigma \end{pmatrix}\right) .$$

The solution of this system is

$$\begin{pmatrix} \hat{\mu}_n \\ \hat{\sigma}_n \end{pmatrix} = ({}^tAB^{-1}A)^{-1} \, {}^tAB^{-1}X_{(\cdot)} ,$$

and

$${}^tAB^{-1}A = \begin{pmatrix} {}^t\mathbb{1}_n B^{-1}\mathbb{1}_n & {}^t\mathbb{1}_n B^{-1}\alpha \\ {}^t\alpha B^{-1}\mathbb{1}_n & {}^t\alpha B^{-1}\alpha \end{pmatrix} .$$

Lemma 3.17 *When the law of Y_i is symmetric, ${}^t\mathbb{1}_n B^{-1}\alpha = 0$. So, the matrix ${}^tAB^{-1}A$ is diagonal.*

As a result,

$$\hat{\mu}_n = \frac{{}^t\mathbb{1}_n B^{-1} X_{(\cdot)}}{{}^t\mathbb{1}_n B^{-1}\mathbb{1}_n} \quad \text{and} \quad \hat{\sigma}_n = \frac{{}^t\alpha B^{-1} X_{(\cdot)}}{{}^t\alpha B^{-1}\alpha} .$$

It can be shown that, if the law of Y_i is not symmetric, then $\hat{\sigma}_n$ underestimates σ.

3.4.3.2 Test Procedure

Let $Z = (Z_1, \dots, Z_n)$ be an i.i.d. sample of distribution $\mathcal{N}(0, 1)$ and $Z_{(1)} \leqslant \dots \leqslant Z_{(n)}$ its ordered counterpart. Let α be the mean vector of the ordered statistics $Z_{(\cdot)}$, *i.e.* $\alpha_i = \mathbb{E}[Z_{(i)}]$,[7] and B be the covariance matrix of the ordered statistics $Z_{(\cdot)}$, *i.e.* $B_{i,j} = \mathbb{E}\left[\left(Z_{(i)} - \alpha_i\right)\left(Z_{(j)} - \alpha_j\right)\right]$.

7: Each expectation α_i depends of n!

Proposition 3.18 (Order statistics) *Let a sample $(X_1, X_2, \dots X_n)$ distributed with probability density function f and cumulative distribution F, then the probability density of the k-th order statistic is given by*

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1} (1 - F(x))^{n-k} f(x) .$$

Moreover, the joint probability density of the n order statistics is

$$f\left(x_{(1)}, x_{(2)}, \ldots, x_{(n)}\right) = n!\Big(\prod_{i=1}^{n} f\left(x_{(i)}\right)\Big)\mathbb{1}_{x_{(1)}<x_{(2)}<\ldots<x_{(n-1)}<x_{(n)}}\,.$$

The idea of the Shapiro-Wilk test is to consider the correlation of $\left(X_{(1)}, \ldots, X_{(n)}\right)$ with $(\alpha_1, \ldots, \alpha_n)$, in other words, to ask whether the order statistics of $(X_1, \ldots, X_n)$ are well correlated with expected standard normal order statistics. A correlation close to 1 would suggest a good fit to normality, whereas a correlation much less than 1 would suggest non-normality.

Definition 3.11 (Shapiro-Wilk test) *The* Shapiro-Wilk test *for testing the normality assumption of X_i is based on the test statistic*

$$W_n = \frac{\hat{\sigma}_n\left({}^t\alpha B^{-1}\alpha\right)^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2\left({}^t\alpha B^{-2}\alpha\right)}\,.$$

It can be written as

$$W_n = \frac{\left(\sum_{i=1}^{n}\alpha_i X_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}\,,$$

where

$$(\alpha_1, \ldots, \alpha_n) = \frac{{}^t\alpha B^{-1}}{\left({}^t\alpha B^{-2}\alpha\right)^{1/2}}$$

is a unit row vector.

The rejection region is of the form $\mathcal{R}_\alpha = \{W_n \leqslant c_{n,\alpha}\}$.

The a_i are tabulated, which makes it easy to calculate W_n; the quantiles $c_{n,1-\alpha}$ are also tabulated.

Listing 3.5: Shapiro-Wilk test on the three simulated samples.

```
> shapiro.test(data1)

    Shapiro-Wilk normality
    test

data:  data1
W = 0.99521, p-value = 0.7804

> shapiro.test(data2)

    Shapiro-Wilk normality
    test

data:  data2
W = 0.95623, p-value = 7.897e
    -06

> shapiro.test(data3)

    Shapiro-Wilk normality
    test

data:  data3
W = 0.33683, p-value < 2.2e
    -16
```

3.5 Very Important Remark: Interpretation of Non-Parametric Tests

The hypotheses of a test can be considered as parts of the set of probability measures on a certain space. In our case, $\mathcal{H}_0$ represents a singleton and $\mathcal{H}_1$ its complement. The non-parametric tests give real information only if the hypothesis $\mathcal{H}_0$ is rejected. Indeed, as soon as the distribution p of the sample is close to p_0, even if it is in $\mathcal{H}_1$, we will accept $\mathcal{H}_0$. This is all the more true when one is obliged to group classes together because the sample is too small or to create classes for continuous laws: many laws then provide the same probability vectors on the finite set. We use a non-parametric test to invalidate a model. If $\mathcal{H}_0$ is rejected, then the model must be modified. If not, then the model (although simplistic, approximate... and probably wrong) is satisfactory.

Chi-Squared Tests 4

The family of chi-squared tests, also written as χ^2 tests, gathers tests with various objectives: adjustment, independence, homogeneity, *etc.* Although, they all have in common that they measure the deviation from the null hypothesis via a "chi-squared divergence", and they are all associated with an asymptotically chi-squared distributed test statistic. The underlying idea is to compare observed numbers or frequencies in a sample with theoretical frequencies derived from statistical/modeling assumptions. Chi-squared tests are valid for the study of qualitative (or discrete) data with finite support. However, in practice, these tests are also applied to discrete data with infinite support or to continuous data after grouping into classes.

4.1 Reminders on the χ^2 Distribution

Definition 4.1 *We consider n independent variables of a reduced centered normal distribution: $Z_1, \dots, Z_n \sim \mathcal{N}(0,1)$. The quantity $\sum_{i=1}^{n} Z_i^2$ is a random variable whose distribution is that of a χ^2 with n degrees of freedom*

Let χ_n^2 be a chi-squared distribution of degree of freedom n. We then have:

- Expectation: $\mathbb{E}\left[\chi_n^2\right] = n$,
- Variance: $\mathcal{V}ar\left(\chi_n^2\right) = 2n$.

The degree of freedom n is the number of linearly independent observations appearing in the sum of squares.

The variable χ^2 is tabulated according to its degree of freedom n. An example of the chi-squared distribution is given in Figure 4.1.

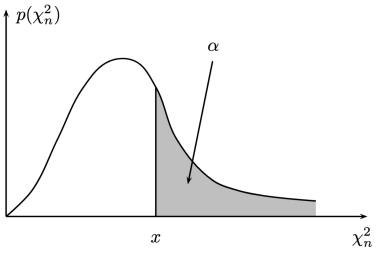


Figure 4.1: The chi-squared distribution.

Proposition 4.1 *Let $n, m \in \mathbb{N}$. We have the relation $\chi_n^2 + \chi_m^2 = \chi_{n+m}^2$.*

Proof. Direct application of the addition theorem for *independent* random variables. □

4.2 Chi-Squared Goodness of Fit Test

The chi-squared goodness of fit test is a statistical hypothesis test used to determine whether a variable is likely to come from a specified distribution or not. It is often used to evaluate whether sample data is representative of the full population.

Let X be a discrete qualitative or quantitative random variable with $K > 1$ modalities $\{a_1, \dots, a_K\}$[1] , of unknown distribution $\pi = (\pi_1, \dots, \pi_K)$, where

$$\forall k \in [\![1, K]\!], \quad \pi_k = \mathbb{P}(X = a_k) > 0\,.$$

1: or K distinct values, or with values in K classes...

Let $X_1, \dots, X_n$ i.i.d distributed according to the same law as X.

Let the probability law $p^0 = (p_1^0, \dots, p_K^0)$ on $\{a_1, \dots, a_K\}$ *known* and such that for all k, $p_k^0 \in]0, 1[$. The problem we aim to solve is the following: given this law p^0, how to know if X is p^0-distributed, from the n-sample $X_1, \dots, X_n$? In other words, we want to test:

$$\mathcal{H}_0 : \text{“}\,\forall k \in [\![1, K]\!],\ \pi_k = p_k^0\,\text{”} \quad \text{versus} \quad \mathcal{H}_1 : \text{“}\,\exists k \in [\![1, K]\!],\ \pi_k \neq p_k^0\,\text{”}\,.$$

A natural idea is to estimate the probability distribution of X using the n-sample $(X_1, \dots, X_n)$ and to compare this estimator with the distribution p^0. We therefore denote $N_k = \sum_{i=1}^n \mathbb{1}_{X_i = a_k}$ the number of times we get the value a_k in the sample and estimate π_k by $\hat{\pi}_k = \frac{N_k}{n}$.[2] We seek to establish whether the difference between the observed and theoretical values is significant or only due to random variation. To this end, we consider the statistic

2: In other words, its unbiased empirical statistic.

$$T_n = n \sum_{k=1}^K \frac{\left(\hat{\pi}_k - p_k^0\right)^2}{p_k^0} = \sum_{k=1}^n \frac{(N_k - n\,p_k^0)^2}{n\,p_k^0}\,.$$

A naive idea would have been to consider the difference $\sum_{k=1}^K \left(\hat{\pi}_k - p_k^0\right)$. But, in this case, the statistic is always zero:

$$\sum_{k=1}^K \left(\hat{\pi}_k - p_k^0\right) = \sum_{k=1}^K \hat{\pi}_k - \sum_{k=1}^K p_k^0 = 1 - 1 = 0\,.$$

Hence the presence of the square to overcome this issue. Finally, to avoid giving too much weight to small values of N_k, we consider a relative error. This statistic is called the chi-squared divergence between the π and p^0 distributions. It measures the “distance” between the observed and theoretical proportions under $\mathcal{H}_0$. Note that it is *not* a distance because it does not check the symmetry property.

Link with the multinomial distribution.

Proposition 4.2 *The random variable $N = (N_1, \dots, N_K)$ follows a multinomial distribution $\mathcal{M}(n, \pi)$ on $\mathbb{N}^K$, i.e. for all $(n_1, \dots, n_K) \in \mathbb{N}^K$ we have*

$$\mathbb{P}(N_1 = n_1, \dots, N_K = n_K) = \begin{cases} \dfrac{n!}{n_1! \dots n_K!}\,\pi_1^{n_1} \dots \pi_K^{n_K} & \text{if } \displaystyle\sum_{k=1}^K n_j = n \\ 0 & \text{else.} \end{cases}$$

Thus, we can reformulate the test by:

$$\mathcal{H}_0 : \text{“}\,N \sim \mathcal{M}(n, p^0)\,\text{”} \qquad \text{versus} \qquad \mathcal{H}_1 : \text{“}\,N \nsim \mathcal{M}(n, p^0)\,\text{”}\,.$$

Proposition 4.3 *Let* $\sqrt{\pi} = \left(\sqrt{\pi_1}, \dots, \sqrt{\pi_K}\right)$. *Then,*

$$Y_n = \left(\frac{N_1 - n\pi_1}{\sqrt{n\pi_1}}, \dots, \frac{N_K - n\pi_K}{\sqrt{n\pi_K}}\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \Gamma) ,$$

where $\Gamma = I_K - {}^t\sqrt{\pi}\sqrt{\pi}$ *is the orthogonal projection matrix on* $\mathcal{V}ect(\sqrt{\pi})^\perp$

Exercise 4.4 *Prove Proposition 4.3. To this end, you may introduce the variables* $S_i = (\mathbb{1}_{X_i=a_1}, \dots, \mathbb{1}_{X_i=a_K})$ *and note that they are i.i.d according to a multinomial distribution of parameters* $(1, \pi)$.

Theorem 4.5 *Assume that* $X_1, \dots, X_n$ *are i.i.d. distributed according to* $\pi = (\pi_1, \dots, \pi_K)$. *Then,*

$$Z_n = \sum_{k=1}^{K} \frac{(N_k - n\pi_k)^2}{n\pi_k} \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2(K-1) .$$

Proof. This theorem is a consequence of the previous proposition and of Cochran's theorem. □

We are now able to define the test procedure.

The Pearson's chi-squared test.

Using the previous results, we get the asymptotic behavior of T_n:

$$T_n = n \sum_{k=1}^{K} \frac{\left(\frac{N_k}{n} - p_k^0\right)^2}{p_k^0} \begin{cases} \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2(K-1) & \text{if } \mathcal{H}_0 \text{ is true,} \\ \xrightarrow[n\to+\infty]{a.s.} +\infty & \text{else.} \end{cases}$$

Thus, at a fixed α level, the Pearson's goodness-of-fit test consists of rejecting the null hypothesis $\pi = p^0$ if:

$$T_n = \sum_{k=1}^{K} \frac{(N_k - np_k^0)^2}{np_k^0} > x_{K-1,1-\alpha} ,$$

where $x_{K-1,1-\alpha}$ is the $1-a$ quantile of a χ^2 with $K-1$ degrees of freedom. According to the previous result, this test is of asymptotic level α.

According to the law of large numbers,

$$\frac{T_n}{n} \geqslant \left\| \frac{N}{n} - p^0 \right\|^2 \xrightarrow[n\to+\infty]{a.s.} \left\| \pi - p^0 \right\|^2 .$$

Hence $\lim_{n\to\infty} T_n = +\infty$ a.s, and the power of the test tends to 1 when n becomes infinite.

The chi-squared test is based on asymptotic results (a convergence in law under $\mathcal{H}_0$ and an almost sure convergence under $\mathcal{H}_1$). However, we only ever have a finite number of observations. Therefore, the challenge is to know to what extent one can act as if this limit is equality. In practice, the literature recommends the following recipe: for the test to be valid, np_k^0 must be greater or equal to 5 for all k. When this is not the case, the classes are grouped until these conditions are verified. However, be careful: the rejection region changes when we group the modalities because *the limit law depends on the number of modalities*.

The χ^2 test can also be used to test the fit of a law on $\mathbb{N}$, $\mathbb{R}$ or even $\mathbb{R}^d$. To do this, it is sufficient to divide the space into a finite number of classes. For a law on $\mathbb{N}$, we use for example the following division:

$$\mathbb{N} = \{0\} \cap \ldots \cap \{k\} \cap \{l \geqslant k+1\}.$$

Listing 4.1: Mendel's example.

```
> Nk=c(315,101,108,32)
> ptheo = c(9,3,3,1)/16
> chisq.test(Nk,p=ptheo)

Chi-squared test for given
    probabilities

data:  Nk
X-squared = 0.47002, df = 3,
    p-value = 0.9254

> n = sum(Nk)
> sum(((Nk-(n*ptheo))^2)/(n*
    ptheo))
[1] 0.470024
> qchisq(0.95,3)
[1] 7.814728
```

Example 4.1 (Mendel's example.) The color trait in peas is encoded by a gene with two allelic forms Y and g corresponding to the colors yellow and green. Yellow is dominant and green recessive. The shape character, round or wrinkled, is carried by another gene with two alleles R (dominant) and w (recessive). We cross 2 populations (pure) of peas: one yellow and round, the other green and wrinkled. According to Mendel's prediction, after 2 crosses, the proportion of peas

- YR: yellow and round is 9/16,
- Yw: yellow and wrinkled is 3/16,
- gR: green and round is 3/16,
- gw: green and wrinkled is 1/16.

In his experiments, Mendel obtained the following results $N_{YR} = 315$, $N_{Yw} = 101$, $N_{gR} = 108$, $N_{gw} = 32$. Here, $K = 4$ and we obtain that $T_n = 0.47$ and $x_{3,0.95} = 7.82$ (See Listing 4.1). Mendel's hypothesis is therefore widely accepted.

4.3 Chi-Squared Goodness of Fit Test to a Family of Laws

Let Θ be an open of $\mathbb{R}^d$, where $d \in [\![1, K]\!]$. Let a family of probability laws $\big(\mathcal{L}(\theta)\big)_{\theta\in\Theta}$ indexed by a parameter θ, and defined on a finite set $\{a_1, \ldots, a_K\}$. Let assume that the maximum likelihood estimator of θ is available.

We want to test if the law of X belongs to the family $\big(\mathcal{L}(\theta)\big)_\theta$, *i.e.*:

$$\mathcal{H}_0 : \text{“}\exists\theta \in \Theta,\ X \sim \mathcal{L}(\theta)\text{”} \quad \text{versus} \quad \mathcal{H}_1 : \text{“}\forall\theta \in \Theta,\ X \nsim \mathcal{L}(\theta)\text{”}.$$

The laws $\big(\mathcal{L}(\theta)\big)_{\theta\in\Theta}$ are characterized by the probability vectors

$$\mathcal{P}(\theta) \;=\; \big\{\, p(\theta) = \big(p_1(\theta),\dots,p_K(\theta)\big)\,;\ \theta\in\Theta\,\big\}$$

on $\{a_1,\dots,a_K\}$. Thus, still denoting π the law of X, we want to test

$$\mathcal{H}_0 : \text{“}\,\pi\in\mathcal{P}\,\text{”} \qquad \text{versus} \qquad \mathcal{H}_1 : \text{“}\,\pi\notin\mathcal{P}\,\text{”}\,.$$

The idea here is to replace p_0 in the T_n statistic defined in the previous section with the distribution of $\mathcal{P}(\Theta)$ that is “closest” to π given the data. To this end, we will replace p_0 with $p(\hat\theta_n)$, where $\hat\theta_n$ is the maximum likelihood estimator of the parameter θ based on the sample $(X_1,\dots,X_n)$, under $\mathcal{H}_0$. All put together, we therefore consider the following statistic:

$$\widehat{T}_n \;=\; \sum_{k=1}^{K}\frac{\big(\hat\pi_k - p_k(\hat\theta_n)\big)^2}{p_k(\hat\theta_n)} \;=\; \sum_{k=1}^{K}\frac{\big(N_k - np_k(\hat\theta_n)\big)^2}{np_k(\hat\theta_n)}\,.$$

Let us apply the following (admitted[3]) result:

3: which is difficult to prove because it involves the properties of maximum likelihood estimators, in particular their strong consistency in most cases: namely $\hat\theta_n$ converges a.s. towards θ, and $\sqrt{n}(\hat\theta_n-\theta)$ converges in law towards a Gaussian distribution.

Theorem 4.6 *Assume that:*

- *For any $k\in[\![1,K]\!]$, $\theta\mapsto p_k(\theta)$ is of class $\mathscr{C}^2$ on Θ and such that for any $\theta\in\Theta$, $p_k(\theta)\neq 0$;*
- *For any $\theta\in\Theta$, the vectors $v_i = {}^t\big(\partial_i p_1(\theta),\dots,\partial_i p_K(\theta)\big)$, where $i\in[\![1,d]\!]$, form a linearly independent family of $\mathbb{R}^K$;*
- *For any $\theta\in\Theta$, if the $X_1,\dots,X_n$ are i.i.d. of distribution $p(\theta)$ then the maximum likelihood estimator $\hat\theta_n$ is consistent towards θ.*

Under these conditions, if $X_1,\dots,X_n$ are i.i.d. of law $p(\theta)$ then

$$\widehat{T}_n \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \chi^2(K-d-1)\,.$$

In particular, the asymptotic behavior of $\widehat{T}_n$ is given by

$$\widehat{T}_n \;=\; n\sum_{k=1}^{K}\frac{\big(\frac{N_k}{n} - p_k(\hat\theta_n)\big)^2}{p_k(\hat\theta_n)} \begin{cases} \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \chi^2(K-d-1) & \text{if } \mathcal{H}_0 \text{ is true,}\\ \underset{n\to+\infty}{\xrightarrow{a.s.}} +\infty & \text{else.}\end{cases}$$

We then construct the χ^2 test of goodness of fit to the family $\mathcal{P}(\Theta)$ as follows: We reject the hypothesis if

$$\widehat{T}_n \;=\; \sum_{k=1}^{K}\frac{\big(N_k - np_k(\hat\theta_n)\big)^2}{np_k(\hat\theta_n)} \;>\; x_{K-d-1,1-\alpha}\,,$$

where $x_{K-d-1,1-\alpha}$ is the $1-a$ quantile of a χ^2 with $K-d-1$ degrees of freedom.

The p-value is

$$p\left(\widehat{T}_n^{\,\text{obs}}\right) \;=\; \mathbb{P}_{\mathcal{H}_0}\left(\widehat{T}_n \geqslant \widehat{T}_n^{\,\text{obs}}\right) \underset{n\to\infty}{\longrightarrow} \mathbb{P}\left(\chi^2(K-d-1)\geqslant \widehat{T}_n^{\,\text{obs}}\right).$$

Under the alternative assumption:

$$\frac{\widehat{T_n}}{n} \geqslant d^2\left(\frac{N}{n}, \mathcal{P}(\Theta)\right) \xrightarrow{a.s.} d^2\big(\pi, \mathcal{P}(\Theta)\big)\,,$$

and therefore the power tends towards 1 as soon as $d^2\big(\pi, \mathcal{P}(\Theta)\big) > 0$.

Remark 4.1 (Degrees of freedom) In Theorem 4.6, the degree of the limit law is smaller the more we test the fit to a large family. Moreover, this degree of freedom is bounded by $k-1$. Intuitively, this makes sense because we no longer compare the empirical frequencies to a fixed law but to the most probable law in a parameterized family, given the observations. We say that $\widehat{T_n}$ is stochastically smaller than T_n.

Example 4.2 For 10000 siblings of (exactly) 4 children, the number of boys composing these siblings is reported in Table 4.1

We decide to model the births by assuming that they are independent and that the probability of having a boy is equal to $\theta \in]0,1[$. We note X_i the number of boys in the i-th sibling. We therefore want to test

$$\mathcal{H}_0 : \text{“ } \exists\theta,\ X_i \sim \mathcal{B}in(4,\theta)\text{ ”} \quad \text{vs.} \quad \mathcal{H}_1 : \text{“ } \forall\theta,\ X_i \nsim \mathcal{B}in(4,\theta)\text{ ”}.$$

Under $\mathcal{H}_0$, the maximum likelihood estimator for θ is given by $\hat{\theta}_n = \frac{1}{4}\overline{X_n}$. We can therefore compute $p(\hat{\theta}_n) = \big(p_0(\hat{\theta}_n), \dots, p_4(\hat{\theta}_n)\big)$ with $p_k(\hat{\theta}_n) = \mathbb{P}(U = k)$ for $U \sim \mathcal{B}in(4,\theta)$. Moreover, still under $\mathcal{H}_0$, the test statistic is

$$\widehat{T_n} = \sum_{k=0}^{4} \frac{\big(N_k - np_k(\hat{\theta}_n)\big)^2}{np_k(\hat{\theta}_n)} \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2(5-1-1) = \chi^2(3)\,.$$

Exercise 4.7 *We study the number of connections to Google during the unit time of one second. We make 200 measurements, reported in Table 4.2. Let X be the $\mathbb{N}$-valued random variable counting the number of connections per second. Can it be considered as a Poisson distribution at the 5% level?*

Table 4.1: *Boys in siblings.* Number of boys in a sibling of 4 children.

Boys (k)	Class size (N_k)
0	572
1	2329
2	3758
3	2632
4	709

Listing 4.2: Boys in siblings.

```
> classes = c(0,1,2,3,4)
> Nk = c
    (572,2329,3758,2632,709)
> n = sum(Nk)
> pihat = Nk/n
> thetahat = sum(Nk*classes) /
     (n*4)
> ptheo = dbinom(0:4,4,
    thetahat)
> Tobs = sum( ((Nk-(n*ptheo))
    ^2) / (n*ptheo) )
> print(Tobs)
[1] 0.9882779
> val = chisq.test(Nk,p=ptheo)

# Beware of degrees of freedom
> print(val)

    Chi-squared test for given
     probabilities

data:  Nk
X-squared = 0.98828, df = 4,
    p-value = 0.9116

> pval = 1-pchisq(val$
    statistic, 3)
> print(pval)
X-squared
0.8040883
```

4.4 Chi-Squared Test of Independence

Let X and Y be two random variables with a finite number of states, $\{a_1, \dots, a_K\}$ and $\{b_1, \dots, b_L\}$ respectively. Let n pairs of random variables $(X_1, Y_1), \dots, (X_n, Y_n)$ be independent and of the same law than (X, Y).

We want to test the independence of the variables X and Y. To do this, we pose:

$$\mathcal{H}_0 : \text{“ } X \perp\!\!\!\perp Y \text{ ”} \qquad \text{versus} \qquad \mathcal{H}_1 : \text{“ } X \not\perp\!\!\!\perp Y \text{ ”} .$$

We start by giving an idea of the construction of the test statistic. First, we recall that the joint probabilities

$$\forall k \in [\![1, K]\!] , \ \forall \ell \in [\![1, L]\!], \quad \mathbb{P}(X = a_k, \, Y = b_\ell)$$

characterize the law of the couple (X, Y). Under $\mathcal{H}_0$,

$$\forall (k, \ell) \in [\![1, K]\!] \times [\![1, L]\!] , \quad \mathbb{P}(X = a_k, Y = b_\ell) = \mathbb{P}(X = a_k) \times \mathbb{P}(Y = b_\ell) .$$

On the other hand, under $\mathcal{H}_1$,

$$\exists (k, \ell) \in [\![1, K]\!] \times [\![1, L]\!] , \quad \mathbb{P}(X = a_k, Y = b_\ell) \neq \mathbb{P}(X = a_k) \times \mathbb{P}(Y = b_\ell) .$$

We introduce the following random variables:

- $N_{k,\ell} = \sum_{i=1}^{n} \mathbb{1}_{X_i = a_k, Y_i = b_\ell}$;
- $N_{k,\cdot} = \sum_{i=1}^{n} \mathbb{1}_{X_i = a_k} = \sum_{\ell=1}^{L} N_{k,\ell}$;
- $N_{\cdot,\ell} = \sum_{i=1}^{n} \mathbb{1}_{Y_i = b_\ell} = \sum_{k=1}^{K} N_{k,\ell}$.

Hence, we can estimate $\mathbb{P}(X = a_k, Y = b_\ell)$ by $\frac{N_{k,\ell}}{n}$ and $\mathbb{P}(X = a_k) \times \mathbb{P}(Y = b_\ell)$ by $\frac{N_{k,\cdot} N_{\cdot,\ell}}{n^2}$. Using the same reasoning as in the previous sections, we obtain the following test statistic:

$$I_n = n \sum_{k=1}^{K} \sum_{\ell=1}^{L} \frac{\left(\frac{N_{k,\ell}}{n} - \frac{N_{k,\cdot} N_{\cdot,\ell}}{n^2} \right)^2}{\frac{N_{k,\cdot} N_{\cdot,\ell}}{n^2}} = \sum_{k=1}^{K} \sum_{\ell=1}^{L} \frac{\left(N_{k,\ell} - \frac{N_{k,\cdot} N_{\cdot,\ell}}{n} \right)^2}{\frac{N_{k,\cdot} N_{\cdot,\ell}}{n}} .$$

Theorem 4.8 *We assume that for all k and all ℓ, $\mathbb{P}(X = a_k) > 0$ and $\mathbb{P}(Y = b_\ell) > 0$. Then, under $\mathcal{H}_0$,*

$$I_n \xrightarrow[n \to +\infty]{\mathcal{L}} \chi^2\big((K-1)(L-1)\big) .$$

This result can be seen as a corollary of Theorem 4.8. Indeed, we test the adequacy of the couple's distribution to the parametric family of product laws on $[\![1, K]\!] \times [\![1, L]\!]$, by estimating the parameters through a maximum likelihood.

Table 4.2: *Internet traffic.* Number of connections to Google per second.

Connections	Class size
0	6
1	15
2	40
3	42
4	37
5	30
6	10
7	9
8	5
9	3
10	2
11	1

Remark 4.2 To quickly find the number of degrees of freedom, note that the number of modes of the couple (X, Y) is KL.Moreover, under $\mathcal{H}_0$ (independence of the variables), to know the distribution of (X, Y), it is sufficient to estimate the first $K-1$ modalities of X, that is to say $\mathbb{P}(X = a_k)$ for $k \in [\![1, K-1]\!]$, and the same for Y. Thus the number of degrees of freedom is given by : $(KL-1) - [(K-1) + (L-1)] = (K-1)(L-1)$.

Everything put together, the asymptotic behavior of I_n is given by

$$I_n = n \sum_{k=1}^{K} \sum_{\ell=1}^{L} \frac{\left(\frac{N_{k,\ell}}{n} - \frac{N_{k,\cdot} N_{\cdot,\ell}}{n^2}\right)^2}{\frac{N_{k,\cdot} N_{\cdot,\ell}}{n^2}} \begin{cases} \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2\big((K-1)(L-1)\big) & \text{if } \mathcal{H}_0 \text{ is true,} \\ \xrightarrow[n\to+\infty]{a.s.} +\infty & \text{else.} \end{cases}$$

Proposition 4.9 *Let $\alpha \in]0, 1[$. The test of rejection region*

$$\mathcal{R}_\alpha = \left\{ I_n > x_{(K-1)(L-1),1-\alpha} \right\}$$

is a test of asymptotic level α to test $\mathcal{H}_0$ against $\mathcal{H}_1$.

Example 4.3 A survey was conducted with a sample of 250 French people about lowering the voting age to 16. In Table 4.3, the responses are ranked according to the respondents' level of education.

Can we say, with a 5% risk of error, that there is a relationship between a person's opinion on this issue and their level of education?

Table 4.3: *Voting age.* For or against lowering the voting age to 16, based on education level.

Educ.	Pros	Against	N_k
Brevet	10	15	25
Bac	20	85	105
Bac+2	20	100	120
$N_{\cdot,\ell}$	50	200	250

Listing 4.3: Voting age.

```
> contingence = matrix(
    c(10,20,20,15,85,100),
    ncol=2)
> chisq.test(contingence)

    Pearsons Chi-squared test

data:  contingence
X-squared = 7.1429, df = 2, p
    -value = 0.02812
```

Exercise 4.10 *Table 4.4 shows the eye and hair colors of 124 individuals. Are the two criteria independent at the 5% level?*

Table 4.4: *Eyes vs. Hair.* Colors of the eyes and hair of 124 individuals.

Hair \ Eyes	Blue	Gray	Brown
Blonds	25	13	7
Brown	9	17	13
Red	7	7	5
Black	3	10	8

4.5 Homogeneity Test

Chi-squared tests can also test the homogeneity of multiple samples.

We study a character that can take K values $\{a_1, \ldots, a_k\}$. We observe $L > 1$ independent samples $E_1, \ldots, E_L$. Last, we denote π_ℓ the discrete distribution of the sample $E\ell = (X_{\ell,1}, \ldots, X_{\ell,n_\ell})$ of size n_ℓ.[4]

We want to test if the L distinct samples come from the same distribution or not, *i.e.*:

$$\mathcal{H}_0 : \text{“ } \pi_1 = \ldots = \pi_L \text{ ”} \qquad \text{versus} \qquad \mathcal{H}_1 : \text{“ } \exists j \neq \ell,\ \pi_j \neq \pi_\ell \text{ ”}.$$

4: In particular, the samples are not necessarily all of the same size.

We know the number of times $N_{k,\ell} = \sum_{i=1}^{n_\ell} \mathbb{1}_{X_{\ell,i}=a_k}$ that the value a_k is observed in the sample E_ℓ. The practical implementation of the test is the same as for the independence test. We define:

$$N_{k,\cdot} = \sum_{\ell=1}^{L} N_{k,\ell} \qquad \text{and} \qquad N_{\cdot,\ell} = \sum_{k=1}^{K} N_{k,\ell} = n_\ell \,.$$

We then consider the test statistic

$$J_n = n \sum_{k=1}^{K} \sum_{\ell=1}^{L} \frac{\left(\frac{N_{k,\ell}}{n} - \frac{N_{k,\cdot} N_{\cdot,\ell}}{n^2}\right)^2}{\frac{N_{k,\cdot} N_{\cdot,\ell}}{n^2}} = \sum_{k=1}^{K} \sum_{\ell=1}^{L} \frac{\left(N_{k,\ell} - \frac{N_{k,\cdot} N_{\cdot,\ell}}{n}\right)^2}{\frac{N_{k,\cdot} N_{\cdot,\ell}}{n}} \,,$$

where $n = \sum_{k=1}^{K} \sum_{\ell=1}^{L} N_{k,\ell}$.

Theorem 4.11 *We assume that for all k and all ℓ, $\pi_{\ell,k} = \mathbb{P}(X_{\ell,i} = a_k) > 0$. Then, under $\mathcal{H}_0$,*

$$J_n \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2\big((K-1)(L-1)\big) \,.$$

As previously, the asymptotic behavior of J_n is given by

$$J_n = n \sum_{k=1}^{K} \sum_{\ell=1}^{L} \frac{\left(\frac{N_{k,\ell}}{n} - \frac{N_{k,\cdot} N_{\cdot,\ell}}{n^2}\right)^2}{\frac{N_{k,\cdot} N_{\cdot,\ell}}{n^2}} \begin{cases} \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2\big((K-1)(L-1)\big) & \text{if } \mathcal{H}_0 \text{ is true,} \\ \xrightarrow[n\to+\infty]{a.s.} +\infty & \text{else.} \end{cases}$$

Proposition 4.12 *Let $\alpha \in]0,1[$. The test of rejection region*

$$\mathcal{R}_\alpha = \left\{ J_n > x_{(K-1)(L-1),1-\alpha} \right\}$$

is a test of asymptotic level α to test $\mathcal{H}_0$ against $\mathcal{H}_1$.

Example 4.4 In this example, we want to know if the participation rate in a sports club of students from two secondary school A and B is identical or not. We have two samples $E_1 = (X_{1,1}, \ldots, X1_{1,n_1})$ and $E_2 = (X_{2,1}, \ldots, X_{2,n_2})$, where $X_{\ell,i}$ is the participation of the i-th student of school ℓ. In other words: $X_{\ell,i} \in \{a_1, a_2\} = \{\text{“yes”}, \text{“no”}\}$. We want to test:

$\mathcal{H}_0$ “ The two pop. are homogeneous (same participation rate) ”

against

$\mathcal{H}_1$ “ The two populations are not homogeneous) ”.

The observed and theoretical sample sizes are given in Table 4.5 and Table 4.6 respectively. The observed test statistic is therefore

$$J_n^{\text{obs}} = \frac{(12 - 17.27)^2}{17.27} + \ldots + \frac{(34 - 39.27)^2}{39.27} = 4.504 > x_{1,0.95} = 3.84 \,.$$

Table 4.5: *Sports club attendance rate.* **Observed** attendance for each of the two schools A and B.

Attend.	A	B	$N_{k,\cdot}$
Yes	12	26	38
No	38	34	72
$N_{\cdot,\ell}$	50	60	110

Table 4.6: *Sports club attendance rate.* **Theoretical** attendance for each of the two schools A and B.

Attend.	School A	School B
Yes	17.27	20.73
No	32.73	39.27

Hence, we reject $\mathcal{H}_0$: the participation rate in the sports club is different between the two schools.

Exercise 4.13 *We seek to invalidate the commonplace that all detergents are equal. Three detergents are used: A, B and C. We sort the clothes at the end of the wash into three categories: very dirty (VD), slightly dirty (SD) and clean (C). The different results are reported in Table 4.7.*

Can we say, at the 5% level, that all detergents are the same?

Table 4.7: *Laundry*. Cleanliness of the laundry at the end of the wash depending on the detergent used.

Detergent \ Clothes	VD	SD	C
A	30	65	205
B	23	56	121
C	75	125	300

Table 4.8: Number for each level of the Osgood scale, for each region.

Score	Brittany	Alsace
1		
2	1	
3	4	1
4	6	3
5	12	4
6	12	4
7	11	11
8	16	13
9	9	10
10	5	10
11	4	9
12	3	5
13	2	5
14	5	10
Total	90	85

Table 4.9: Observed numbers

Score	$>$ Med.	$\leqslant$ Med.	$N_{k,\cdot}$
Brit.	28	62	90
Al.	49	36	85
$N_{\cdot,\ell}$	77	98	175

Table 4.10: Theoretical numbers

Score	$>$ Med.	$\leqslant$ Med.	$N_{k,\cdot}$
Brit.	39.6	50.4	90
Al.	37.4	47.6	85
$N_{\cdot,\ell}$	77	98	175

4.5.1 Back to the Median Test

As we saw in Subsection 3.3.3, the idea of the median test is to test the hypothesis that two populations have the same median. This is done by counting the number of observations above the overall average for each sample. Then, using a chi-square test, we test whether the difference from the median is significant.

Example 4.5 (Cheese factory) A cheese factory commissioned a survey of its customers in two regions: Brittany and Alsace. Respondents give their opinion on a new cheese using fourteen-level Osgood scales (Table 4.8). Before launching the cheese on the market, the marketing manager would like to check the consistency between the scores given by each sample.

The median of the total sample is 8. For each region, we can therefore split our observations according to this quantile. We obtain Table 4.9 of observed numbers and Table 4.10 of theoretical numbers.

All computations done, we find that the observed chi-square statistic is 12.493, and that the p-value is 4.09×10^{-4}: The difference is therefore significant.

4.5.1.1 Yates Correction

The chi-square statistic is overestimated in the one degree of freedom situation. For this reason, we usually proceed to a correction of the median test: in each cell of the table where the theoretical number of individuals is greater than the observed number, we add 0.5; and, in the opposite case, we subtract 0.5.

In our example, this would modify the contingency table (Table 4.9) to retain the values 28.5, 61.5, 48.5, and 36.5. The p-value computed from these new numbers is then 7.19×10^{-4}. This change does not affect the conclusion.

Linear Model

Principle of the Linear Model and First Examples

5

Let Y be a random variable valued in $\mathbb{R}^n$ of which we observe a sample. Most often, we call this variable the *response variable*. The objective of the following chapters is to build a model which explains "as well as possible" this variable according to *explanatory variables* observed in the same sample.

5.1 Regular Linear Model

Definition 5.1 *A variable Y consisting of n observations Y_i is said to follow a* statistical linear model *if Y can be written in the form*

$$Y = X\theta + \varepsilon\,, \tag{5.1}$$

where:

- *$X \in \mathcal{M}_{n,k}\mathbb{R}$ is a known real matrix with n rows and k columns, such that $k < n$,*
- *$\theta \in \mathbb{R}^k$ is an unknown real vector of size k,*
- *the random vector $\varepsilon \in \mathbb{R}^n$ represents the error of the model.*

This definition is very general and goes far beyond the regression and variance analysis framework. The hypothesis $k < n$ means that the number of observations must be greater than the number of parameters to be estimated. This is a kind of identifiability assumption.

Definition 5.2 *The linear model 5.1 is called* regular *if the matrix X is regular, i.e. of rank k. Otherwise, i.e. if X is of rank $r < k$, we speak of* singular *models.*

5.1.1 Reminders About the Rank

Proposition 5.1 (Link between Injectivity and Rank) *Let $X \in \mathcal{M}_{n,k}\mathbb{R}$. The following propositions are equivalent:*

- *X is a matrix of rank k,*
- *The application $X : \mathbb{R}^k \to \mathbb{R}^n$ is injective,*
- *The matrix ${}^tXX$ is invertible*

Thus, if X is regular, then by injectivity of the application X, the equation $X\theta = 0_n$ has for unique solution $\theta = 0_k$. In particular, the columns of X are linearly independent in $\mathbb{R}^n$.

In some situations, the considered matrix X cannot be regular. However, it is sometimes possible to overcome this problem by adding so-called

identifiability constraints on the parameters to be estimated (See Section 8.1). From now on, unless explicitly stated otherwise, the model is assumed to be regular.

Proposition 5.2 (Hat Matrix) *Let $X \in \mathcal{M}_{n,k}\mathbb{R}$ be a regular matrix. Then the projection matrix on $\mathcal{I}m(X)$ is given by $P_{[X]} = X({}^tXX)^{-1}\,{}^tX$.*

Proof. Let $H := X({}^tXX)^{-1}\,{}^tX$ where $X \in \mathcal{M}_{n,k}\mathbb{R}$ is a regular matrix. For any $u \in \mathbb{R}^n$, we have $u = Hu + u - Hu$ and $Hu = X({}^tXX)^{-1}\,{}^tXu \in \mathcal{I}m(X)$. Let us show that $u - P_{[X]}u \in \mathcal{I}m(X)^\perp$. Let $v \in \mathbb{R}^k$. We have

$$\begin{aligned} {}^t(Xv)(u - P_{[X]}u) &= {}^tv\,{}^tX(u - X({}^tXX)^{-1}\,{}^tXu) \\ &= {}^tv\,{}^tXu - {}^tv({}^tXX)({}^tXX)^{-1}\,{}^tXu = 0\,. \end{aligned}$$

Hence the result. □

This matrix is called the *hat matrix* and is most often noted H.

5.1.2 Fundamental Assumptions

In order to be able to work more simply and to go further in the study of this model, we will now impose some restrictions on the vector ε.

Assumption (A1): Errors are centered, *i.e.* for all $i \in [\![1, n]\!]$, $\mathbb{E}[\varepsilon_i] = 0$.

This assumption is very important and ensures that the model is correctly defined, in that no relevant effects have been missed. Indeed, in the case where $\mathbb{E}[\varepsilon] \neq 0_n$, it would mean that part of the information was omitted when modeling $\mathbb{E}[Y]$. More precisely, this hypothesis amounts to assuming that

$$\mathbb{E}[Y] = X\theta = \sum_{j=1}^{k} \theta_j x^{(j)}\,,$$

where $x^{(j)}$ denotes the j-th column of the matrix X. In other words, the variables $x^{(j)}$ make it possible to explain Y by a cause and effect relationship. A counter-example is given in Figure 5.1. In this example it is clear that a curvature has been forgotten and that a better model would be

$$\forall i \in [\![1, n]\!] \quad Y_i = \theta_1 + \theta_2 Z_i + \theta_3 Z_i^2 + \varepsilon_i\,.$$

Moreover, this relationship is linear in nature: on average, Y writes as a linear combination of $x^{(j)}$. The variables $x^{(j)}$ are called explanatory variables or *predictors*, and the matrix X the "design matrix".

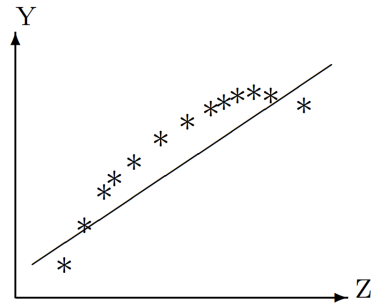


Figure 5.1: *Counter-example* to fundamental hypothesis (A1): The points are not aligned along a straight line but along a parabola.

Remark 5.1 The linear nature of the relationship between $x^{(j)}$ and Y justifies the term "linear model".

Assumption (A2): The variance of the errors is constant: For all $i \in [\![1, n]\!]$, $\mathcal{V}ar(\varepsilon_i) = \sigma^2$, where σ^2 is an unknown parameter to be estimated.

This amounts to imposing on Y that, for any $i \in [\![1, n]\!]$, $\mathcal{V}ar(Y_i) = \sigma^2$. In practice this assumption in one of the most difficult to check. However, it is often reasonable to assume that we meet (A2). If this is not the case, we can set up a statistical treatment of the linear model. However, this requires much more work.

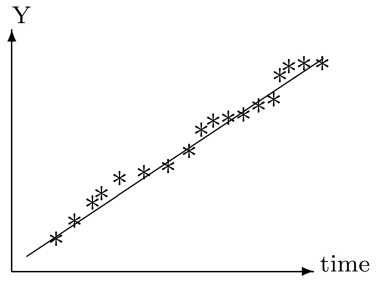


Figure 5.2: *Counter-example* to fundamental hypothesis (A3): Presence of inertia in the process: The curve tends to stay above the line for some time when it crosses it, and vice versa.

Assumption (A3): The variables ε_i are independent.

We will consider that this hypothesis is checked when each observation (statistical unit) corresponds to an independent sampling or a physical experiment under independent conditions. This is not always the case. For example, consider time series:[1] some inertia may occur, and the error of the past can influence the future error. Hence, temporal problems require particular statistical treatments (ARMA process, for instance).

1: We will not deal here with this type of data, which would require one, or even several, dedicated chapters.

Hypothesis (H4): The variables ε_i are distributed according to Gaussian laws: For all $i \in [\![1, n]\!]$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

This assumption is the least important since we can get by without it when the number of data is large.

The normality of errors assumption can be justified:

- *By a theoretical argument:* Errors can be described as measurement errors. They are an accumulation of small, uncontrollable, and independent hazards. For example, an animal's weight measurement may be subject to fluctuations due to measurement errors during weighing, its state of health, its genetic baggage, or even its natural propensity to gain more or less weight. According to the central limit theorem, if all these effects are independent with the same zero mean and the same "small" variance, their sum tends towards a Gaussian variable. The Gaussian distribution models all situations where chance results from several causes independent of each other. Notably, measurement errors generally follow the Gaussian distribution quite well.
- *By a practical argument:* It is easy to check if a random variable follows a normal distribution. By studying the a posteriori distribution of the computed residuals (estimated errors) and comparing it to the theoretical (normal) distribution, one often finds that the Gaussian assumption is reasonable.

It follows from the hypotheses (A1 – 4) the normality of Y:

$$Y \sim \mathcal{N}_n\left(X\theta, \sigma^2 I_n\right) .$$

This last equality could have been chosen as a definition of a linear model, this is formally correct, but in practice, it is better to distinguish the four hypotheses. In particular, as we have seen, the hypothesis of Gaussianity (H4) is less critical, especially for large data sets. In several cases, we shall consider a non-Gaussian linear model where (H4) is simply removed or replaced by a weaker form: i.i.d errors with finite fourth moments, for instance.

In the statistical literature, several methods, often graphical, are proposed to test (A1 – 4). We will discuss them in Section 9.7.

5.2 Example: Linear Gaussian Models

5.2.1 The Linear Regression Model

We try to model a quantitative variable Y as a function of quantitative explanatory variables $x^{(1)}, \dots, x^{(p)}$. Under the Gaussian assumption, the linear regression model is written as

$$Y_i = \theta_0 + \theta_1 x_i^{(1)} + \dots + \theta_p x_i^{(p)} + \varepsilon_i ,$$

where $\theta_0, \theta_1, \dots, \theta_p$ are unknown parameters and the $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d of laws $\mathcal{N}(0, \sigma^2)$, where σ^2 has to be estimated. We can rewrite the model in the following matrix form:

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

where $\theta = {}^t(\theta_0, \theta_1, \dots, \theta_p)$ and $X = (\mathbb{1}_n, x^{(1)}, \dots, x^{(p)}) \in \mathcal{M}_{n,p+1}\mathbb{R}$.

Exercise 5.3 *What is the law of Y_i? The one for Y?*

The linear regression model will be studied in detail in Chapter 9.

5.2.2 The Analysis of Variance Model

We want to model a quantitative variable Y as a function of one, or more, qualitative explanatory variable(s) called factor(s). Under the Gaussian assumption, the one-factor model with I modalities is written

$$\forall i \in [\![1, I]\!], \quad \forall j \in [\![1, n_i]\!], \quad Y_{i,j} = \mu_i + \varepsilon_{i,j} , \tag{5.2}$$

where $\mu_1, \dots, \mu_I$ are unknown parameters and where $\varepsilon_{1,1}, \dots, \varepsilon_{I,n_I}$ are independent samples of distribution $\mathcal{N}(0, \sigma^2)$, with σ^2 to be estimated.

Exercise 5.4 (Matrix writing of this model) *In order to write this model*

in matrix form, the observations are arranged by modality of the factor:

$$Y = {}^t(Y_{1,1}, \ldots, Y_{1,n_1}, Y_{2,1}, \ldots, Y_{2,n_2}, \ldots, Y_{I,1}, \ldots, Y_{I,n_I}) .$$

Let $n = \sum_{i=1}^{I} n_i$. *Write the model* (5.2) *in the form*

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

by specifying the design matrix $X \in \mathcal{M}_{n,I}\mathbb{R}$ *and* $\theta \in \mathbb{R}^I$. *What is the law of* $Y_{i,j}$, $Y_i = {}^t(Y_{i,1}, \ldots, Y_{i,n_i})$ *and* Y*?*

The analysis of variance (ANOVA) model will be studied in detail in Chapter **??**.

Estimation of the Parameters 6

In this chapter, we will focus on the estimation of parameters in a regular general linear model

$$Y = X\theta + \varepsilon\,, \quad \text{where} \quad \varepsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$$

and where $X \in \mathcal{M}_{n,k}\mathbb{R}$ is a matrix of rank $rank(X) = k$.

Note that the linear model is a statistical model with $k+1$ parameters : $\theta \in \mathbb{R}^k$ and $\sigma \in \mathbb{R}$.

6.1 Estimation of θ

In this section, we focus on the estimation of the parameter vector $\theta \in \mathbb{R}^k$. To do so, we use the least-squares method. We aim to find the vector θ that minimizes the distance between the image of the matrix X and the observations Y. In other words, the least-squares estimator of θ is defined by

$$\hat{\theta} \in \underset{\vartheta\in\mathbb{R}^k}{\operatorname{argmin}} \|Y - X\vartheta\|_2^2 := \underset{\vartheta\in\mathbb{R}^k}{\operatorname{argmin}} SSE(\vartheta)\,. \tag{6.1}$$

In the previous formula, the norm $\|\cdot\|_2$ is the one resulting from the usual scalar product in $\mathbb{R}^n$, *i.e.*

$$\forall u \in \mathbb{R}^n, \quad \|u\|_2^2 = \langle u \,|\, u \rangle = \sum_{i=1}^{n} u_i^2 = {}^t u u\,.$$

Hence, in matrix form, it is possible to write

$$\hat{\theta} \in \underset{\vartheta}{\operatorname{argmin}}\, {}^t(Y - X\vartheta)(Y - X\vartheta)\,.$$

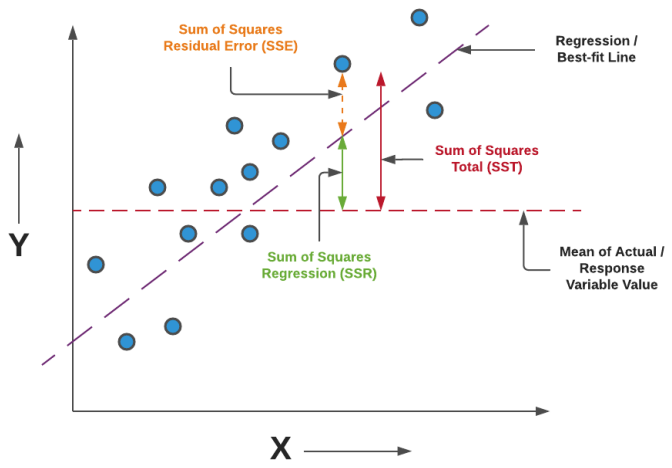


Figure 6.1: SSR, SSE and SST Representation in relation to Linear Regression.

SST (Total Sum of Squares): Sum of the squared difference between actual values related to the response variable and the empirical mean of actual values. It is also called Variance of the Response.
SSE (Error Sum of Squares): Sum of the squared difference between the actual and predicted values. It is also termed as Residual Sum of Squares.
SSR (Regression Sum of Squares): Sum of the squared difference between the predicted value and mean of actual values. It is also termed as Explained Sum of Squares.
For more details, refer to Section 6.7.

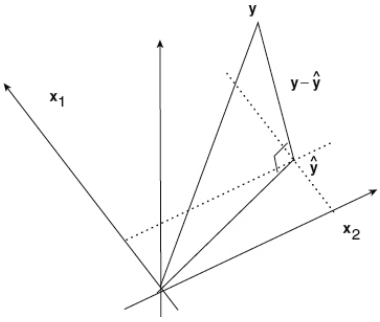


Figure 6.2: Geometric interpretation of the least-squares estimator.

Theorem 6.1 *Let Y follow a regular linear model. The estimator $\hat{\theta}$ obtained by the least-squares method is*

$$\hat{\theta} = ({}^tXX)^{-1}\,{}^tXY\,.$$

Proof. Let $P_{[X]}$ be the orthogonal projection on $\mathcal{I}m(X)$. Then:

$$\min_{\theta}\|Y - X\theta\|^2 = \min_{u\in\mathcal{I}m(X)}\|Y-u\|^2 = \|Y - P_{[X]}Y\|^2\,.$$

See Figure 6.2 for an illustration. Hence, $X\hat{\theta} = P_{[X]}Y = X({}^tXX)^{-1}\,{}^tXY$. As X is assumed to be regular, we deduce that $\hat{\theta} = ({}^tXX)^{-1}\,{}^tXY$ by uniqueness. □

This first theorem gives us an explicit formula for the least-squares estimator of the vector θ. Interestingly, this formula is purely geometrical and does not require any knowledge of the law of errors. Indeed, the least-squares estimator of the vector θ checks the following property :

$$X\hat{\theta} = P_{[X]}Y\,.$$

Moreover, since the solution is explicit, we can efficiently compute it, and at a relatively low numerical cost: solving k linear systems, which is usually straightforward. Thus, linear models can have large sizes and fit very well in reality.

Remark 6.1 In the particular case where the errors are Gaussian, the least-squares estimator $\hat{\theta}$ corresponds exactly to the maximum likelihood estimator. Indeed, in this case, we have :

$$\mathcal{L}(\theta,\sigma^2;y) = \prod_{i=1}^{n} f(y_i;\theta)\,,$$

where $f(y_i;\theta)$ is the density of the normal distribution of the random variable Y_i . In other words,

$$\mathcal{L}\left(\theta,\sigma^2;(Y_1,\ldots,Y_n)\right) = \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(-\frac{\|Y-X\theta\|^2}{2\sigma^2}\right).$$

To obtain the maximum likelihood estimator θ, we then maximize the above log-likelihood as a function of θ. However, by growth of the exponential function, this amounts to minimize $\|Y - X\theta\|^2$.

The following result explains the performance of the least-squares estimator.

Theorem 6.2 *Let Y follow a regular linear model and $\hat{\theta}$ be the least squares estimator defined by* (6.1)*. Then:*

1. *The least square estimator is unbiased:*

$$\mathbb{E}[\hat{\theta}] = \theta \qquad \textit{and} \qquad \mathcal{V}ar(\hat{\theta}) = \sigma^2({}^tXX)^{-1}\,;$$

2. *(Rao-Blackwell Theorem) Moreover, if the variables ε_i are i.i.d of centered Gaussian distribution, i.e. under (A3 − 4), $\hat{\theta}$ is the best estimator among all unbiased estimators of θ, i.e.*

$$\mathcal{V}ar({}^tC\tilde{\theta}) \geqslant \mathcal{V}ar({}^tC\hat{\theta}),$$

for any unbiased $\tilde{\theta}$ estimator of θ, and any linear combination ${}^tC\theta$, where $C \in \mathbb{R}^k$.

3. *Last, under the same assumptions, $\hat{\theta}$ is a Gaussian vector:*

$$\hat{\theta} \sim \mathcal{N}_k\left(\theta, \sigma^2({}^tXX)^{-1}\right).$$

Exercise 6.3 *Prove Theorem 6.2. Lets recall that $\mathbb{E}[Y] = X\theta$ and, for all matrix A, $\mathcal{V}ar(AY) = A\mathcal{V}ar(Y)\,{}^tA$.*

In addition to its unbiased nature, the strength of the least squares estimator lies in the control of the precision of the estimate, thanks to Theorem 6.2.1.

6.2 Adjusted Values and Residuals

Once we have estimated θ by $\hat{\theta}$, we can define $\hat{Y}_i$ the *predicted (or adjusted) values* by the model. For each observation Y_i,

$$\hat{Y} = {}^t(\hat{Y}_1, \ldots, \hat{Y}_n) = X\hat{\theta} = X({}^tXX)^{-1}\,{}^tXY = P_{[X]}Y = HY.$$

Moreover, the following residuals

$$\hat{\varepsilon} = {}^t(\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n) = Y - \hat{Y} = (I_n - P_{[X]})Y = (I_n - H)Y$$

provide an estimate of the errors ε_i.

Thus, given realizations y_i, we obtain the observed predicted values $\hat{y}_i = (\hat{Y}_i)^{\text{obs}} = (X\hat{\theta}^{\text{obs}})_i$ and the computed residuals $(\hat{\varepsilon}_i)^{\text{obs}} = y_i - \hat{y}_i$.

Proposition 6.4

1. $\hat{Y} \sim \mathcal{N}_n(X\theta, \sigma^2 H)$, *where* $H = X({}^tXX)^{-1}\,{}^tX$;
2. $\hat{\varepsilon} \sim \mathcal{N}_n(0_n, \sigma^2(I_n - H))$;
3. *The random variables* $\hat{Y}$ *and* $\hat{\varepsilon}$ *are independent;*
4. *The random variables* $\hat{\theta}$ *and* $\hat{\varepsilon}$ *are independent.*

Exercise 6.5 *Prove Proposition 6.4.*

Indications:

1. *Use the law of* $\hat{\theta}$*;*
2. *Note that* $\hat{\varepsilon} = (I_n - H)Y$ *and* $Y \sim \mathcal{N}(X\theta, \sigma^2 I_n)$*;*
3. *Cf. Cochran's theorem;*

6.3 Estimation of σ^2

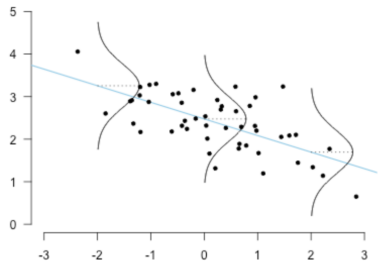


Figure 6.3: Interpretation of σ^2 as the variance of Y around the regression line

In this section, we are interested in the estimate of $\sigma^2 \in \mathbb{R}$, the variance of the errors, called the *residual variance*. By definition of the linear model, the residual variance σ^2 is also given as the variance of Y for X fixed. In the context of linear regression, this is interpreted as the variance of Y around the theoretical regression line (*cf.* Figure 6.3). This definition of σ^2 then suggests estimating it from the differences between the observed Y_i and the adjusted $\hat{Y}_i$ values.

Theorem 6.6 *Let* $\hat{\theta}$ *be the least squares estimator of* θ*. Under the assumptions (A1 – 4), and if* $X \in \mathcal{M}_{n,k}\mathbb{R}$*, then*

$$\hat{\sigma}^2 = \frac{\|\hat{\varepsilon}\|^2}{n-k} = \frac{\|Y - \hat{Y}\|^2}{n-k} = \frac{\|Y - X\hat{\theta}\|^2}{n-k} = \frac{SSE(\hat{\theta})}{n-k}$$

is an optimal unbiased estimator of σ^2*, independent of* $\hat{\theta}$*. Moreover,*

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n-k}\,\chi^2(n-k)\,.$$

Exercise 6.7 (Proof of Theorem 6.6)

1. *Show that* $SST(\hat{\theta}) := \|Y - X\hat{\theta}\|^2 = \|P_{[X]^\perp} Y\|^2$ *;*
2. *Using Cochran's theorem, show that* $SST(\hat{\theta}) \sim \sigma^2 \chi^2(n-k)$ *; Deduce that* $\hat{\sigma}^2$ *is an unbiased estimator of* σ^2 *;*
3. *Using Proposition 6.4, show that* $\hat{\theta}$ *and* $\hat{\sigma}^2$ *are independent.*

The estimation of σ^2 is therefore

$$(\hat{\sigma}^2)^{\text{obs}} = \frac{\|(\hat{\varepsilon})^{\text{obs}}\|^2}{n-k} = \frac{\|y - \hat{y}\|^2}{n-k}.$$

The denominator $n - k$ comes from the fact that we have already estimated k parameters in the model.

Remark 6.2 From a geometrical point of view,

- $\hat{\theta}$ depends on the projection of the data on $\mathcal{I}m(X)$, and
- $\hat{\sigma}^2$ on the projection of the data on $\mathcal{I}m(X)^\perp$.

6.4 Standard Errors

The standard error of the regression is the average distance that the observed values fall from the regression line.

- According to Theorem 6.2, the variance-covariance matrix of $\hat{\theta}$ is given by $\Gamma_{\hat{\theta}} = \sigma^2 ({}^tXX)^{-1}$, where σ is unknown. We estimate this matrix by

$$\hat{\Gamma}_{\hat{\theta}} = \hat{\sigma}^2 ({}^tXX)^{-1}.$$

  Thus, $\mathcal{V}ar(\hat{\theta}_j)$ is estimated by $\hat{\sigma}^2 [({}^tXX)^{-1}]_{jj}$ and, consequently, the *standard error* of $\hat{\theta}_j$, denoted se_j , is given by

$$se_j = \sqrt{\hat{\sigma}^2 [({}^tXX)^{-1}]_{jj}}.$$

  Hence, the correlation matrix of $\hat{\theta}$ has for element j, j':

$$r(\hat{\theta}_j, \hat{\theta}_{j'}) = \frac{\hat{\sigma}^2 [({}^tXX)^{-1}]_{jj'}}{se_j \times se_{j'}} = \frac{[({}^tXX)^{-1}]_{jj'}}{\sqrt{[({}^tXX)^{-1}]_{jj}} \sqrt{[({}^tXX)^{-1}]_{j'j'}}}.$$

- Likewise, the variance $\mathcal{V}ar(\hat{Y}) = \sigma^2 H = \sigma^2 X({}^tXX)^{-1}\,{}^tX$ is estimated by $\hat{\sigma}^2 H$. Therefore, $\sqrt{\hat{\sigma}^2 H_{ii}}$ is the standard error of $\hat{Y}_i$.

- Finally, $\sqrt{\hat{\sigma}^2(1-H_{ii})}$ is the error of $\hat{\varepsilon}_i$. We can then define $\hat{\varepsilon}_i/\sqrt{\hat{\sigma}^2}$ the *standardized residual* and $\hat{\varepsilon}_i/\sqrt{\hat{\sigma}^2(1-H_{ii})}$ the *studentized residual*.

6.5 Confidence Intervals

The confidence interval measures the degree of precision one has on the sample estimates. Two main sources of variation in the data can lead to a lack of precision in estimating a quantity.

- *Insufficient data*: For example, in the case of a survey, one does not interview the entire population but only a fraction of the population. Similarly, only a finite number of measurements are made for physical measurements, whereas, in theory, an infinite number of measurements is desirable to obtain a perfect result.
- There can also be *noise in the measurement*, which is almost always the case in practice.

Assume that we want to estimate a parameter denoted ϑ. The confidence interval I_y, at the confidence level $1-\alpha$, for an observation Y, is the interval in which, for any value ϑ

$$\mathbb{P}_\vartheta\left[Y \mid \vartheta \in I_y\right] \geqslant 1-\alpha\,.$$

This does *not* mean that "the probability that the true value of ϑ falls in I_y is $1-\alpha$", which would not make sense since this value is not a random variable. It means that "if the true value of ϑ is not in I_y, the *a priori* probability of the observation outcome y we obtained was less than α". For example, suppose the parameter ϑ is not in the interval. In that case, the observation y corresponds to a rare phenomenon for which the confidence interval does not contain the true value.

6.5.1 Confidence Interval for θ_j

Given that $\hat{\theta} \sim \mathcal{N}_k\left(\theta, \sigma^2({}^tXX)^{-1}\right)$, we have $\hat{\theta}_j \sim \mathcal{N}_k\left(\theta_j, \sigma^2[({}^tXX)^{-1}]_{jj}\right)$. Therefore,

$$\frac{\hat{\theta}_j - \theta_j}{\sqrt{\sigma^2[({}^tXX)^{-1}]_{jj}}} \sim \mathcal{N}(0,1)\,.$$

Moreover,

$$(n-k)\,\hat{\sigma}^2 \sim \sigma^2\chi^2(n-k)\,.$$

Otherwise, these two random variables are independent. Hence, Cochran's theorem ensures that:

$$T = \frac{\dfrac{\hat{\theta}_j - \theta_j}{\sqrt{\sigma^2 \left[({}^tXX)^{-1}\right]_{jj}}}}{\sqrt{\dfrac{(n-k)\hat{\sigma}^2}{(n-k)\sigma^2}}} = \frac{\hat{\theta}_j - \theta_j}{\hat{\sigma}\sqrt{\left[({}^tXX)^{-1}\right]_{jj}}} \sim \mathcal{T}(n-k)\,.$$

If we denote by $t_{n-k,1-\alpha/2}$ the $(1-\alpha/2)$-quantile of the Student's law with $(n-k)$ degrees of freedom, then the confidence interval of the parameter θ_j of security $1-\alpha$ is defined by :

$$\begin{aligned} CI_{1-\alpha}(\theta_j) &= \left[\hat{\theta}_j \pm t_{n-k,1-\alpha/2} \times \hat{\sigma}\sqrt{\left[({}^tXX)^{-1}\right]_{jj}}\right] \\ &= \left[\hat{\theta}_j \pm t_{n-k,1-\alpha/2}\, se_j\right]\,. \end{aligned}$$

6.5.2 Confidence Interval for $(X\theta)_i$

Let $\mathbb{E}[Y_i] = (X\theta)_i$ be the average response of Y_i. We estimate it by $\hat{Y}_i = (X\hat{\theta})_i$. Since $\hat{\theta} \sim \mathcal{N}_k\left(\theta, \sigma^2({}^tXX)^{-1}\right)$, according to the properties of Gaussian vectors, the distribution of $\hat{Y}_i$ is

$$\hat{Y}_i = (X\hat{\theta})_i \sim \mathcal{N}\left((X\theta)_i, \sigma^2\left[X({}^tXX)^{-1}\,{}^tX\right]_{ii}\right) = \mathcal{N}\left((X\theta)_i, \sigma^2 H_{ii}\right)\,,$$

with the notations introduced previously: $H = X({}^tXX)^{-1}\,{}^tX$. Moreover, $(n-k)\hat{\sigma}^2 \sim \sigma^2\chi^2(n-k)$, and $\hat{Y}$ is independent of $\hat{\sigma}^2$. Hence,

$$\frac{\hat{Y}_i - (X\theta)_i}{\hat{\sigma}\sqrt{\left[X({}^tXX)^{-1}\,{}^tX\right]_{ii}}} \sim \mathcal{T}(n-k)\,.$$

The confidence interval of $(X\theta)_i$ at the $1-\alpha$ confidence level is therefore given by:

$$\begin{aligned} CI_{1-\alpha}\left((X\theta)_i\right) &= \left[\hat{Y}_i \pm t_{n-k,1-\alpha/2} \times \hat{\sigma}\sqrt{\left[X({}^tXX)^{-1}\,{}^tX\right]_{ii}}\right] \\ &= \left[\hat{Y}_i \pm t_{n-k,1-\alpha/2} \times \hat{\sigma}\sqrt{H_{ii}}\right]\,. \end{aligned}$$

Remark 6.3 Point estimation is possible without any assumption on the distribution of the errors ε, thanks to the method of least squares. However, it is not the same for the estimation by confidence interval (and for the tests): in this case, the (H4) assumption of Gaussianity is mandatory.

6.6 Prediction

A linear model can also be used to make predictions, *i.e.* to predict the expected value for the response Y_0 when the explanatory variables take given values X_0. This question has two facets:

1. One can be interested in the average behavior of Y_0, *i.e.* $\mathbb{E}[Y_0]$ for these given values X_0 of the explanatory variables.
2. We can also be interested in the real value that Y_0 will take if only one trial is performed in these explanatory variables X_0.

First, it is crucial to understand the difference between this two settings. First one is related to the confidence interval of $\mathbb{E}[Y_0]$, and second one is related to the prediction interval. Figure 6.4 illustrates this difference. While confidence intervals represent the range of uncertainty associated with the estimator of an unknown parameter, prediction intervals are ranges of values that may contain future individual observations. More precisely, we assume a new set of given values of the explanatory variables X_0 in both cases. Except that, in the first case, we want to predict an average response corresponding to these explanatory variables. And in the second case, we want to predict a new "individual" value.

For example, suppose one is studying the relationship between the weight and age of an animal. In that case, one can predict the value of the 20-day weight either as the average weight of the animals at 20 days or as the 20-day weight of a new animal. For the new animal, individual variability must be taken into account, which increases the estimator's variance and, thus, the interval's width.

More generally, the prediction interval is always wider than the confidence interval because of the additional uncertainty associated with predicting an individual value. The prediction interval also depends on the quality of the model and its adequacy to the region of interest.

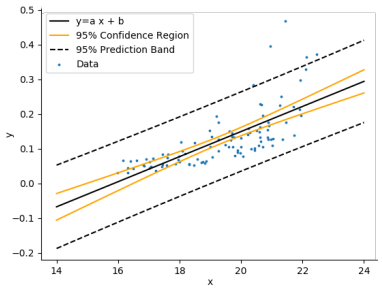


Figure 6.4: Confidence Interval and Prediction interval bands in linear regression

6.6.1 Confidence Interval for $X_0\theta$

We consider new values for the explanatory variables, gathered in the linear vector $X_0 \in \mathcal{M}_{1,k}\mathbb{R}$. The average response is then $X_0\theta$.

The estimator of $X_0\theta$ is $\hat{Y}_0 = X_0\hat{\theta}$. By the same arguments as before, the distribution of $\hat{Y}_0$ is:

$$\hat{Y}_0 = X_0\hat{\theta} \sim \mathcal{N}\left(X_0\theta, \sigma^2 X_0({}^tXX)^{-1}\,{}^tX_0\right) .$$

So, the confidence interval of $X_0\theta$ at the confidence level of $1-\alpha$ is written:

$$CI_{1-\alpha}(X_0\theta) = \left[X_0\theta \pm t_{n-k,1-\alpha/2} \times \hat{\sigma}\sqrt{X_0({}^tXX)^{-1}\,{}^tX_0}\right] .$$

6.6.2 Prediction Interval

Let the same new set of explanatory variables $X_0 \in \mathcal{M}_{1,k}\mathbb{R}$. A new observation Y_0, corresponding to X_0, is written

$$Y_0 = X_0\theta + \varepsilon_0\,,$$

where ε_0 is assumed to be independent of ε_i for all $i \in [\![1, n]\!]$, and $\varepsilon_0 \sim \mathcal{N}(0, \sigma^2)$.

To predict in which interval the result of a new trial will lie, we have to consider two uncertainty factors:

- the uncertainty in the estimate of the average test result $X_0\theta$,
- the uncertainty on the error term ε_0.

In the context of linear model, the parameter vector θ is estimated by

$$\hat{\theta} = ({}^tXX)^{-1}\,{}^tXY\,,$$

where $Y = {}^t(Y_1, \ldots, Y_n)$. The linear model then predicts the value

$$\hat{Y}_0 = X_0\hat{\theta} \sim \mathcal{N}\left(X_0\theta, \sigma^2\, X_0({}^tXX)^{-1}\,{}^tX_0\right)\,.$$

According to the assumptions on ε_0, we have that $Y_0 \sim \mathcal{N}(X_0\theta, \sigma^2)$, and Y_0 is independent of $\hat{Y}_0$. Hence,

$$Y_0 - \hat{Y}_0 \sim \mathcal{N}_k\left(0, \sigma^2\,(1 + X_0({}^tXX)^{-1}\,{}^tX_0)\right)\,.$$

Moreover, according to Theorem 6.6,

$$(n-k)\,\hat{\sigma}^2 \sim \sigma^2\,\chi^2(n-k)\,.$$

Since $\hat{\sigma}^2$ is independent of $\hat{\theta}$ and ε_0 (because ε_0 is independent of all the ε_i, $i \in [\![1, n]\!]$), it comes

$$\frac{Y_0 - \hat{Y}_0}{\hat{\sigma}\,\sqrt{1 + X_0({}^tXX)^{-1}\,{}^tX_0}} \sim \mathcal{T}(n-k)\,.$$

Finally, if we denote by $t_{n-k,1-\alpha/2}$ the $(1 - \alpha/2)$-quantile of the Student's law with $(n-k)$ degrees of freedom, we obtain

$$\mathbb{P}\left(Y_0 \in \left[\hat{Y}_0 \pm t_{n-k,1-\alpha/2} \times \hat{\sigma}\,\sqrt{1 + X_0({}^tXX)^{-1}\,{}^tX_0}\right]\right) = 1 - \alpha\,.$$

Therefore, the prediction interval of the variable Y for a new observation at point X_0 is defined by

$$CI_{1-\alpha}(Y_0) = \left[\hat{Y}_0 \pm t_{n-k,1-\alpha/2} \times \hat{\sigma}\,\sqrt{1 + X_0({}^tXX)^{-1}\,{}^tX_0}\right]\,.$$

In particular, we can notice an increase of the variance with respect to

$$CI_{1-\alpha}(X_0\theta) = \left[\hat{Y}_0 \pm t_{n-k,1-\alpha/2} \times \hat{\sigma}\,\sqrt{X_0({}^tXX)^{-1}\,{}^tX_0}\right]\,.$$

Example 6.1 A light bulb manufacturer wants to estimate the burn time of his bulbs. He takes a random sample of 100 bulbs and records their burn time to failure in a spreadsheet. He finds a 95% confidence interval of the mean of [1230, 1265] hours. Therefore, he is 95% sure that the true average for the whole population of bulbs is within this interval. He also calculates the prediction interval and finds [1350, 1500] hours (for specific levels of the input manufacturing parameters). Thus, he is 95% certain that the next bulb produced under the same conditions will burn between 1350 and 1500 hours.

Remark 6.4 Do not risk learning these formulas by heart! You need to understand (and be able to redo) the construction of confidence intervals for a parameter, an average response, and a prediction interval for a new response.

6.7 Decomposition of the Variance

The purpose of implementing a linear model is to explain the variability of a variable Y by other variables.

Definition 6.1 (Empirical variance) *Let Z be a real continuous random variable. Suppose that we observe an n-sample $(Z_1, \dots, Z_n)$ having the same distribution as Z. We define the* empirical variance *of Z as the variance of the n-sample $(Z_i)_{i\in[\![1,n]\!]}$, taken as a discrete variable. Namely, given the empirical mean $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$ of Z,*

$$\widehat{\mathcal{V}ar}(Z) = \frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \bar{Z}\right)^2 .$$

Likewise, we can define the empirical covariance between two continuous variables, or between a continuous variable and a discrete variable, as long as the associated samples/discrete variables are observed the same number of times. Due to abuse of notation, the "hat" is sometimes omitted.

We note:[1]

- the *total variability* of Y:

$$SST = \|Y - \bar{Y}\mathbb{1}_n\|^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = n\,\widehat{\mathcal{V}ar}(Y);$$

- the *variability explained* by the model, *i.e.* by the predictors, or regression sum of squares:

$$SSR = \|\hat{Y} - \bar{Y}\mathbb{1}_n\|^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = n\,\widehat{\mathcal{V}ar}(\hat{Y});$$

1: Attention! In French, we use the notations SCT for "Somme des Carrés Totale", SCE for "Somme des Carrés Expliquée" (by regression), and SCR for "Somme des Carrés Résiduelle". In particular, the French notation SCE ("Somme des Carrés Expliquée" by the regression) corresponds to the English notation SSR (regression sum of squares). And conversely, the French notation SCR corresponds to the English SSE.

- the *residual variability*, not explained by the model also called error sum of squares:

$$SSE = \|Y - \hat{Y}\|^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = n\,\widehat{Var}(\hat{\varepsilon})\,.$$

All these quantities are shown in Figure 6.1 and Figure 6.5.

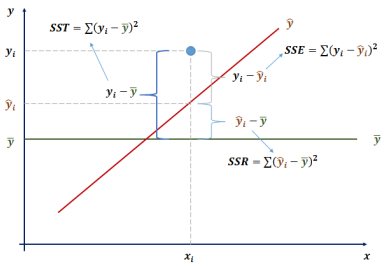


Figure 6.5: Decomposition of the Variance.

Proposition 6.8 (Decomposition of the Variance) *The total variance of Y admits the following decomposition:*

$$\widehat{Var}(Y) = \widehat{Var}(\hat{Y}) + \widehat{Var}(\hat{\varepsilon}) \qquad i.e. \qquad SST = SSR + SSE\,.$$

Exercise 6.9 *Prove this result.*

We will see later that this decomposition leads to definitions specific to each model depending on the model studied.

According to the least squares criterion used to estimate the parameters, the objective is to minimize the residual variability SSE and thus maximize the explained variability SSR. To judge the fit of the model to the data, we define the R^2 *criterion*, which represents the proportion of the variance of Y explained by the model:[2]

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{\widehat{Var}(\hat{Y})}{\widehat{Var}Y} \in [0,1]\,.$$

The closer R^2 is to 1, the better the model fits the data. We will discuss the efficiency of this criterion in the following chapters.

2: Be careful ⚠
Here, the formula is "reversed" in French and in English...

Fisher-Snedecor Test 7

This chapter will focus on several tests that can be implemented on the linear model. We will assume during all this part that the hypotheses (A1 – 4) are verified. The tests presented below cannot be used if these constraints are not satisfied.

7.1 Nested Models

We consider a linear Gaussian model

$$Y = X\theta + \varepsilon\,, \quad \text{where} \quad \varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n)\,. \tag{7.1}$$

We are interested in investigating the nullity of some components of the parameter θ or some linear combinations of its components, for example $\theta_j = 0$; $\theta_j = \theta_k = 0$ or $\theta_j = \theta_k$. These assumptions rely on the notion of nested models.

Definition 7.1 *Two models are said to be* nested *if one can be considered as a particular case of the other. This is equivalent to comparing a reference model to a reduced or constrained model.*

This approach aims at determining whether the model used can be simplified or not. Two examples of submodels are:

General model of simple linear regression:	$Y_i = a + bX_i + \varepsilon_i$,
Submodel with slope nullity:	$Y_i = a + \varepsilon_i$.
General model of the 1-factor analysis of variance:	$Y_{i,j} = \mu_i + \varepsilon_{i,j}$,
Submodel with group equality	$Y_{i,j} = \mu + \varepsilon_{i,j}$.

In the following, we will consider two equivalent writings of the null hypothesis $\mathcal{H}_0$: the first is more practical, while the second is more theoretical.

Writing 1: To specify the nullity of some components of the θ parameter, we introduce a matrix $C \in \mathcal{M}_{qk}\mathbb{R}$ where k denotes the number of parameters of the reference model and $q \in [\![1, k]\!]$ the number of constraints tested. We try to find out if $C\theta = 0_q$. In other words, C represent the coefficient of a linear combination, and we want to test

$$\mathcal{H}_0 : \text{“}C\theta = 0_q\text{”} \qquad \text{against} \qquad \mathcal{H}_1 : \text{“}C\theta \neq 0_q\text{”}$$

The matrix $C \in \mathcal{M}_{q,k}$ is assumed to be of rank q.

Exercise 7.1 *We assume a model with $k = 3$ parameters. In the following three cases, specify the matrix C:*

1. *Test the hypothesis $\mathcal{H}_0$: "$\theta_2 = 0$",*
2. *Test the hypothesis $\mathcal{H}_0$: "$\theta_2 = \theta_3$",*
3. *Test the hypothesis $\mathcal{H}_0$: "$\theta_2 = \theta_3 = 0$".*

Writing 2: Let us consider the general framework of the linear model. Let the model (7.1) and X_0 be a matrix such that

$$\mathcal{I}m(X_0) \subset \mathcal{I}m(X) \quad \text{and} \quad k_0 = dim\big(\mathcal{I}m(X_0)\big) \; < \; k = dim\big(\mathcal{I}m(X)\big)\,.$$

The model defined by

$$Y \; = \; X_0\beta + \varepsilon \tag{7.2}$$

is called a sub-model of the linear model defined in (7.1). Most often, X_0 is a matrix consisting of k_0 column vectors of X with $k_0 < k$ and β is a vector of length k_0. We then note SSE_0 the sum of squares of the residuals of this sub-model, associated to $n - k_0$ degrees of freedom and defined as follows

$$SSE_0 \; = \; \|Y - X_0\hat{\beta}\|^2\,,$$

where $\hat{\beta}$ is the least squares estimator from model (7.2) for β. Insofar as $\mathcal{I}m(X0) \subset \mathcal{I}m(X)$ and by definition of the least squares estimators, we can notice that $SSE_0 \geqslant SSE$.

In order to try to know if the observations are from model (7.1) or (7.2), we introduce the model

$$Y \; = \; R + \varepsilon\,.$$

Therefore, testing for the presence of a submodel is equivalent to testing

$$\mathcal{H}_0\colon \text{“}R \in \mathcal{I}m(X_0) \qquad \text{against} \qquad \mathcal{H}_1\colon \text{“}R \in \mathcal{I}m(X) \setminus \mathcal{I}m(X_0)\text{”}\,.$$

7.2 Fisher-Snedecor Test

The Fisher-Snedecor test is the decision rule for rejecting, or not,

$$\mathcal{H}_0\colon \text{“}C\theta = 0_q\text{”}\,, \qquad i.e. \qquad \mathcal{H}_0\colon \text{“}R \in \mathcal{I}m(X_0)\text{”}\,.$$

- Rejecting $\mathcal{H}_0$ means deciding that $C\theta \neq 0_q$, *i.e.* that some components of $C\theta$ are not null. Therefore, we do not have confidence in the sub-model, and we prefer to continue working with the reference model;

- Not to reject $\mathcal{H}_0$ is not to exclude that all the components of $C\theta$ are null. In this case, keeping a too complicated model is unnecessary, and we prefer to keep the constrained and simpler model to explain the data.

7.2.1 Test Statistics and Decision Rule

Consider the framework of the general linear model (7.1), under the (A1 – 4) assumptions.

Theorem 7.2 *Under the null hypothesis $\mathcal{H}_0$, i.e. assuming that the sub-model (7.2) is true,*

$$F = \frac{\frac{SSE_0 - SSE}{k - k_0}}{\frac{SSE}{n-k}} = \frac{\frac{\|\hat{Y} - \hat{Y}_0\|^2}{k-k_0}}{\frac{\|Y-\hat{Y}\|^2}{n-k}} \sim \mathcal{F}(k-k_0, n-k)\,,$$

where $\mathcal{F}(k-k_0, n-k)$ denotes the Fisher distribution with parameters $(k-k_0, n-k)$.

Moreover, F is independent of $\hat{Y}_0 = X_0\hat{\beta}$.

Exercise 7.3
1. *Show that $SSE = \|P_{[X]^\perp}\varepsilon\|^2 \sim \sigma^2\chi^2(n-k)$;*
2. *Let A be a vector subspace of $\mathcal{I}m(X) = [X]$ such that $\mathcal{I}m(X) = A \overset{\perp}{\oplus} \mathcal{I}m(X_0)$.*
   Show that $SSE_0 - SSE = \|P_A\varepsilon\|^2 \underset{\mathcal{H}_0}{\sim} \sigma^2\chi^2(k-k_0)$;
3. *Deduce that $F \underset{\mathcal{H}_0}{\sim} \mathcal{F}(k-k_0, n-k)$;*
4. *Show that F is independent of $\hat{Y}_0$ and $\hat{\beta}$.*

One can use Figure 7.1 as a basis for reasoning.

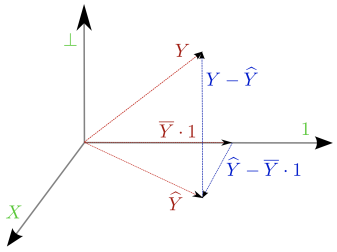


Figure 7.1: Geometric interpretation of the F-test

Proposition 7.4 *This Fisher-Snedecor test statistic can be rephrased as follows:*

$$F = \frac{{}^t[C\hat{\theta}][C({}^tXX)^{-1}\,{}^tC]^{-1}[C\hat{\theta}]}{q\,\hat{\sigma}^2}, \quad \textit{where} \quad q = k - k_0\,.$$

The demonstration does not present any conceptual difficulty but is very calculating. At first reading, one can admit it.

Proof. Let $\Delta = C({}^tXX)^{-1}\,{}^tC$ and $\bar{F} = \dfrac{{}^t[C\hat{\theta}][C({}^tXX)^{-1}\,{}^tC]^{-1}[C\hat{\theta}]}{q\,\hat{\sigma}^2}$.
The demonstration is in four steps.

1. *Let us show that Δ is invertible.*

   Since $\mathrm{rk}(C) = q$, the application $C\colon \mathbb{R}^k \to \mathbb{R}^q$ is surjective. Especially, ${}^tC$ is injective.

   The matrix $({}^tXX)^{-1}$ being invertible, there exists $A \in \mathcal{M}_k\mathbb{R}$ invertible such that $({}^tXX)^{-1} = A\,{}^tA$, and

   $$\mathrm{rk}(\Delta) = \mathrm{rk}\left(CA\,{}^tA\,{}^tC\right) = \mathrm{rk}\left({}^tA\,{}^tC\right) = q - \dim\left(\ker({}^tA\,{}^tC)\right)\,.$$

   However, using the invertibility of ${}^tA$ and the injectivity of ${}^tC$,

   $${}^tA\,{}^tCx = O_k \iff {}^tCx = 0_k \implies x = 0_q\,.$$

   Hence $\ker({}^tA\,{}^tC) = \{O_q\}$ and $\mathrm{rk}(\Delta) = q$, *i.e.* Δ is invertible.

2. *Let us show that $\bar{F} \sim \mathcal{F}(q, n-k)$ under $\mathcal{H}_0$.*

   Given that $\hat{\theta} \sim \mathcal{N}_k\left(\theta, \sigma^2({}^tXX)^{-1}\right)$, we have $C\hat{\theta} \sim \mathcal{N}_q\left(C\theta, \sigma^2\Delta\right)$. In particular, under $\mathcal{H}_0$, $C\hat{\theta} \sim \mathcal{N}_q\left(0_q, \sigma^2\Delta\right)$.

   As above, Δ being invertible, there exists $\Lambda \in \mathcal{M}_q\mathbb{R}$ invertible such that $\Delta = \Lambda\,{}^t\Lambda$. Hence, under $\mathcal{H}_0$, ${}^t\Lambda C\hat{\theta} \sim \mathcal{N}_q\left(0_q, \sigma^2 I_q\right)$. We therefore deduce that

   $$\frac{{}^t[{}^t\Lambda C\hat{\theta}][{}^t\Lambda C\hat{\theta}]}{\sigma^2} \sim \chi^2(q)\,.$$

   Moreover, $(n-k)\,\hat{\sigma}^2 \sim \sigma^2\chi^2(n-k)$ and $\hat{\theta}$ and $\hat{\sigma}^2$ are independent. Hence, by definition of the Fisher distribution,

   $$\frac{\dfrac{{}^t[{}^t\Lambda C\hat{\theta}][{}^t\Lambda C\hat{\theta}]}{\sigma^2\,q}}{\dfrac{(n-k)\,\hat{\sigma}^2}{(n-k)\,\sigma^2}} = \frac{{}^t[C\hat{\theta}][C({}^tXX)^{-1}\,{}^tC]^{-1}[C\hat{\theta}]}{q\,\hat{\sigma}^2} \sim \mathcal{F}(q, n-k)\,.$$

3. *Let us show that $F = \bar{F}$.*

   Note that

   $$\begin{aligned}\|Y - X_0\hat{\beta}\|^2 &= \min_{\beta\in\mathbb{R}^q}\|Y - X_0\beta\|^2 = \min_{u\in\mathcal{I}m(X_0)}\|Y - u\|^2\\ &= \min_{u\in X(\ker(C))}\|Y - u\|^2 = \min_{\theta\in\ker(C)}\|Y - X\theta\|^2\\ &:= \|Y - X\bar{\theta}\|^2\,.\end{aligned}$$

The vector $\bar{\theta}$ minimizes $\|Y - X\theta\|^2$ under the constraint $\theta \in \ker(C)$. To determine $\bar{\theta}$, we then solve the following constrained optimization problem: Let $\lambda \in \mathbb{R}$,

$$\begin{aligned} & \frac{\partial}{\partial\theta}\left[{}^t(Y - X\theta)(Y - X\theta) + \lambda\,{}^tC\theta\right] = 0_k \\ \iff \quad & \frac{\partial}{\partial\theta}\left[{}^tYY - {}^t\theta\,{}^tXY - {}^tYX\theta + {}^t\theta\,{}^tXX\theta + \lambda\,{}^tC\theta\right] = 0_k \\ \iff \quad & -2\,{}^tXY + 2\,{}^tXX\theta + \lambda\,{}^tC = 0_k\,. \end{aligned}$$

Hence,

$$\bar{\theta} = ({}^tXX)^{-1}\,{}^tXY - \frac{\lambda}{2}({}^tXX)^{-1}\,{}^tC\,.$$

Using the constraint $\bar{\theta} \in \ker(C)$ and the invertibility of Δ, we get ${}^{\lambda}/_2 = \Delta^{-1}C({}^tXX)^{-1}\,{}^tXY$, and by putting all the pieces together,

$$\begin{aligned} \bar{\theta} &= ({}^tXX)^{-1}\,{}^tXY - ({}^tXX)^{-1}\,{}^tC\Delta^{-1}C({}^tXX)^{-1}\,{}^tXY \\ &= \hat{\theta} - ({}^tXX)^{-1}\,{}^tC\Delta^{-1}C\hat{\theta}\,. \end{aligned}$$

Therefore,

$$\begin{aligned} \|X\hat{\theta} - X_0\hat{\beta}\|^2 = \|X\hat{\theta} - X\bar{\theta}\|^2 &= \|({}^tXX)^{-1}\,{}^tC\Delta^{-1}C\hat{\theta}\|^2 \\ &= {}^t(C\hat{\theta})\Delta^{-1}C({}^tXX)^{-1}\,{}^tXX({}^tXX)^{-1}\,{}^tC\Delta^{-1}(C\hat{\theta}) \\ &= {}^t(C\hat{\theta})\Delta^{-1}(C\hat{\theta})\,, \end{aligned}$$

and by definition $\hat{\sigma}^2 = {}^{\|Y - \hat{Y}\|^2}/_{n-k}$.

4. *Let us show that* $q = k - k_0$.

   Let $(e_1, \ldots, e_{k-q})$ be a basis of $\ker(C)$. So $(Xe_1, \ldots, Xe_{k-q})$ is a generating family of $X(\ker(C))$. It is also a free family as X is injective. Thus $\dim(X(\ker(C))) = \dim(\mathcal{I}m(X_0)) = k - q = k_0$.

□

This last expression has the advantage of not requiring the estimation of the constrained model to test $\mathcal{H}_0$: "$C\theta = 0_q$" against $\mathcal{H}_1$: "$C\theta \neq 0_q$".[1]

1: However, a matrix has to be inverted. There is no magic formula!

In the following, we will note F^{obs} the observed value of the random variable F.

The quantity of interest in the construction of our Fisher test is $\Delta(SSE) = SSE_0 - SSE$. Intuitively, if the observed value of $\Delta(SSE)$ is substantial, there is little chance that the observations of Y are "from" the sub-model. On the other hand, if the observed value of $\Delta(SSE)$ is small, the original model can most likely be simplified: the sub-model explains the observations insofar as SSE_0 is comparable to SSE. Therefore, the rejection zone with a first-order risk α writes

$$\mathcal{R}_\alpha = \{F > f_{q,n-k,1-\alpha}\}\,,$$

where $f_{q,n-k,1-\alpha}$ is the $(1-\alpha)$-quantile of the Fisher distribution of degrees of freedom $q = k - k_0$ and $n - k$.

7.3 Special Case Where $q = 1$: Student's Test

In the particular case where we test the nullity of a single linear combination of the components of the parameter, *i.e.* $q = 1$ and $C \in \mathcal{M}_{1,k}\mathbb{R}$, then the null hypothesis can be written as

$$\mathcal{H}_0 : \text{“}C\theta = 0\text{”}\,.$$

We then have $C({}^tXX)^{-1}\,{}^tC \in \mathbb{R}$, and the random variable F writes as follows:

$$F \;=\; \frac{(C\hat{\theta})^2}{\hat{\sigma}^2\, C({}^tXX)^{-1}\,{}^tC}\,.$$

F follows a Fisher distribution with 1 and $n-k$ degrees of freedom. However, a property of the Fisher-Snedecor distribution ensures that a Fisher-Snedeor distribution with 1 and m degrees of freedom is nothing but the square of a Student distribution with m degrees of freedom. Therefore, we obtain the following equality: If $\Phi \sim \mathcal{F}(1, n-k)$ and $T \sim \mathcal{T}(n-k)$, then

$$\mathbb{P}\big[\Phi \leqslant f_{1,n-k,1-\alpha}\big] \;=\; 1-\alpha \;=\; \mathbb{P}\left[T^2 \leqslant f_{1,n-k,1-\alpha}\right]\,.$$

We therefore deduce the following property on quantiles:

$$f_{1,n-k,1-\alpha} \;=\; t^2_{n-k,1-\alpha/2}\,.$$

According to Fisher's test, we reject the hypothesis $\mathcal{H}_0$ if $F > f_{1,n-k,1-\alpha}$. Yet, we have the following equivalence

$$F \leqslant f_{q,n-k,1-\alpha} \quad\Longleftrightarrow\quad |C\hat{\theta}| \leqslant t_{n-k,1-\alpha/2} \times \hat{\sigma}\sqrt{C({}^tXX)^{-1}\,{}^tC}\,.$$

Hence the confidence interval at security level $1-\alpha$ of $C\theta$ is

$$CI_{1-\alpha}(C\theta) \;=\; \left[C\hat{\theta} \,\pm\, t_{n-k,1-\alpha/2} \times \hat{\sigma}\sqrt{C({}^tXX)^{-1}\,{}^tC}\right]\,.$$

Finally, the test consists of rejecting the null hypothesis if and only if 0 does not belong to the confidence interval of $C\theta$.

Exercise 7.5 *Directly construct the Student's t-test of nullity of the parameter θ_j at the α level.*

7.4 Confidence Interval (Region) for $C\theta$

7.4.1 Confidence Interval for $C\theta \in \mathbb{R}$

Let's start with the confidence interval for a linear combination $C\theta \in \mathbb{R}$. We keep to the notations of Section 7.3. As $\hat{\theta} \sim \mathcal{N}_k(\theta, \sigma^2({}^tXX)^{-1})$, we have $C\hat{\theta} \sim \mathcal{N}(C\theta, \sigma^2\Delta)$, where $\Delta = C({}^tXX)^{-1}\,{}^tC \in \mathbb{R}$. Moreover, $(n-k)\hat{\sigma}^2 \sim \sigma^2\chi^2(n-k)$, and $\hat{\theta}$ and $\hat{\sigma}^2$ are independent. Then,

$$\frac{C\hat{\theta} - C\theta}{\hat{\sigma}\sqrt{\Delta}} \sim \mathcal{T}(n-k)\,.$$

This gives the following confidence interval at confidence level $1-\alpha$:

$$CI_{1-\alpha}(C\theta) = \left[C\hat{\theta} \pm t_{n-k,1-\alpha/2} \times \hat{\sigma}\sqrt{C({}^tXX)^{-1}\,{}^tC}\right].$$

7.4.2 Confidence Region for $C\theta \in \mathbb{R}^q$

Suppose now that, as in Section 7.2, $C\theta$ is of dimension $q > 1$.

Recall that the set of c_0 accepted for a test $\mathcal{H}_0$: "$C\theta = c_0$" against $\mathcal{H}_1$: "$C\theta \neq c_0$", at the α-level, defines a confidence interval at the $1-\alpha$ confidence level. In particular, this definition does not require c_0 to be in $\mathbb{R}$. We can therefore generalize the construction of confidence intervals to any dimension.

Let c_0 be any particular value of $\mathbb{R}^q$. We have $C\hat{\theta} - C\theta \sim \mathcal{N}_q(0_q\sigma^2\Delta)$, where $\Delta = C({}^tXX)^{-1}\,{}^tC \in \mathcal{M}_q\mathbb{R}$. Then,

$$\frac{{}^t(C\hat{\theta} - C\theta)\,\Delta^{-1}\,(C\hat{\theta} - C\theta)}{\sigma^2} \sim \chi^2(q)\,.$$

Morover, $(n-k)\hat{\sigma}^2 \sim \sigma^2\chi^2(n-k)$, and the two statistics are independent. Hence, still using the same arguments, we deduce that

$$\Phi := \frac{{}^t(C\hat{\theta} - C\theta)\,\Delta^{-1}\,(C\hat{\theta} - C\theta)}{q\hat{\sigma}^2} \sim \mathcal{F}(q, n-k)\,.$$

Last,

$$\begin{aligned}
&\mathbb{P}\left(\Phi \leqslant f_{q,n-k,1-\alpha}\right) = 1-\alpha \\
&\quad\iff \mathbb{P}\left({}^t(C\hat{\theta} - C\theta)\,\Delta^{-1}\,(C\hat{\theta} - C\theta)] \leqslant q\,\hat{\sigma}^2\,f_{q,n-k,1-\alpha}\right) = 1-\alpha \\
&\quad\iff \mathbb{P}\left(C\theta \in \mathcal{E}_{1-\alpha}(C\theta)\right) = 1-\alpha\,,
\end{aligned}$$

where $\mathcal{E}_{1-\alpha}(C\theta)$ is the confidence ellipsoid defined by

$$\mathcal{E}_{1-\alpha}(C\theta) = \left\{u \in \mathbb{R} \mid {}^t(C\hat{\theta} - u)\,(C({}^tXX)^{-1}\,{}^tC)^{-1}\,(C\hat{\theta} - u) \leqslant q\,\hat{\sigma}^2\,f_{q,n-k,1-\alpha}\right\}.$$

The set of $c_0 \in \mathbb{R}^q$ accepted by the test $\mathcal{H}_0$: "$C\theta = c_0$" against $\mathcal{H}_1$: "$C\theta \neq c_0$", at the α-level, forms the $\mathcal{E}_{1-\alpha}(C\theta)$ confidence ellipsoid defined above.

Singular Models and Orthogonality

8

8.1 Singular Models

Up to now, we have restricted ourselves to studying regular linear models. However, some models cannot be parameterized in a regular way: they are naturally *over-parameterized*. A typical example is the additive model in the analysis of variance (ANOVA) with two factors (*cf.* Chapter 12).

Consider, for instance, the case where the two factors both have two levels, and we observe the four combinations once and only once. So, with the notations seen in Chapter 1, we have :

$$\forall i \in \{1,2\}\,, \quad \forall j \in \{1,2\}\,, \quad Y_{i,j} = \mu + a_i + b_j + \varepsilon_{i,j}\,.$$

In particular, the parameters of the model and the design matrix write respectively

$$\theta = \begin{pmatrix} \mu \\ a_1 \\ a_2 \\ b_1 \\ b_2 \end{pmatrix} \qquad \text{and} \qquad X = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

Here, it is obvious that X is singular. Indeed, its kernel is not restricted to the null vector: ${}^t(2,-1,-1,-1,-1) \in \mathrm{Ker}(X)$ for instance. The values $\mu, a_i, b_i, i \in \{1,2\}$, are therefore not uniquely identifiable. Actually, the model is over-parameterized: we have five unknown parameters for only four observations.

Definition 8.1 *A linear model is said to be* singular *or non-regular when the matrix X is non-injective, i.e. if there exists $\theta \neq 0_k$ such that $X\theta = 0_n$.*

In this case :

- $X\hat{\theta}$ remains unique, since it is the orthogonal projection of Y onto $\mathcal{I}m(X)$;
- On the other hand, $\hat{\theta}$ is *not* unique. Indeed, if $\hat{\theta}$ is a solution, then for all $u \in \mathrm{Ker}(X)$, $\hat{\theta} + u$ is still a solution.

Moreover, if X is not regular, then the matrix $trXX$ is not invertible. To handle this issue, we define hereafter the notion of generalized inverse.

Definition 8.2 (Generalized inverse) *Let M be a matrix. We define* a *generalized inverse of M, denoted $M^\dagger$, by $MM^\dagger M = M$.*

This construction is always possible. Indeed, $X \mapsto {}^tXX$ defines a bijective application of $\mathrm{Ker}(X)^\perp$ on itself. It is therefore sufficient to neglect the

part contained in the kernel: We define $({}^tXX)^\dagger$ as the true inverse on $\mathrm{Ker}(X)^\perp$, arbitrarily completed on $\mathrm{Ker}(X)$. Hence, the definition of $({}^tXX)^\dagger$ is far from unique!

It is then possible to generalize the results of the regular case.

Proposition 8.1 *If $({}^tXX)^\dagger$ is a generalized inverse matrix of ${}^tXX$, then $\hat{\theta} = ({}^tXX)^\dagger\,{}^tXY$ is a solution of the normal equations*

$$({}^tXX)\theta = {}^tXY\,.$$

Proof. By definition of the transposition operation, pour tout $v \in \mathbb{R}^k$,

$$\left\langle Xv \,\middle|\, P_{[X]^\perp}Y \right\rangle = \left\langle v \,\middle|\, {}^tXP_{[X]^\perp}Y \right\rangle = 0$$

and in particular

$${}^tXY = {}^tXP_{[X]}Y + {}^tXP_{[X]^\perp}Y = {}^tXP_{[X]}Y\,.$$

Thus, there exists $u \in \mathbb{R}^k$ such that ${}^tXY = {}^tXXu$. Last

$$\begin{aligned}\left({}^tXX\right)\hat{\theta} &= ({}^tXX)({}^tXX)^\dagger\,{}^tXY \\ &= ({}^tXX)({}^tXX)^\dagger\,{}^tXXu = {}^tXXu = {}^tXY\,.\end{aligned}$$

□

Remark 8.1 This estimator is not unique and depends on the definition chosen for $({}^tXX)^\dagger$. On the other hand, as said before, the vector $X\hat{\theta}$ remains unique, even if the matrix X is singular.

In general, we prefer to remove the indeterminacy on $\hat{\theta}$ by setting constraints to give a more intuitive meaning to the estimated parameters composing θ.

8.1.1 Constraints on Identifiability

Suppose the matrix X is singular of rank $r < k$ so that there are $k - r$ redundant parameters.

Proposition 8.2 *Let $M \in \mathcal{M}_{k-r,k}\mathbb{R}$ be a matrix of rank $k - r$ such that $\mathrm{Ker}(M) \cap \mathrm{Ker}(X) = \{0_k\}$. Then:*

- *The matrix $({}^tXX + {}^tMM)$ is invertible and its inverse is a generalized inverse matrix of ${}^tXX$;*
- *The vector $\hat{\theta} = ({}^tXX + {}^tMM)^{-1}\,{}^tXY$ is the unique solution of the system* $\begin{cases} {}^tXX\theta = {}^tXY \\ {}^tM\theta = 0_{k-r}\,. \end{cases}$

Exercise 8.3 *Prove Proposition 8.2.*

1. *To show that* ${}^tXX + {}^tMM$ *is invertible, show that the matrix*

$$A = \begin{pmatrix} X \\ M \end{pmatrix} \in \mathcal{M}_{n+k-r,k}\mathbb{R}$$

   is injective. Thus, ${}^tAA$ *is invertible.*
2. *Consider the following minimization problem:*

$$g\colon \theta \mapsto \|Y - X\theta\|^2 + \|M\theta\|^2.$$

   Write $g(\theta)$ *as* $g(\theta) = \|\tilde{Y} - A\theta\|^2$ *with* $\tilde{Y}$ *to be specified. Deduce that* $\hat{\theta}$ *is a solution of the system* $\begin{cases} {}^tXX\theta = {}^tXY \\ {}^tM\theta = 0_{k-r} \end{cases}$.
3. *Show the uniqueness of the solution.*

The choice of the constraints is not always obvious. Moreover, we obtain a corresponding estimator for each constraint, which can sometimes be confusing.

Example 8.1 Let's take the example of the one-factor analysis of variance with differential effect. For simplicity, we assume that $I = 4$. The model writes as follows:

$$\forall i \in [\![1,4]\!], \forall j \in \{1\}, \quad Y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}.$$

Or, in matrix form

$$\theta = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

In particular, the X matrix associated with the model is singular. We must therefore impose a constraint (called identifiability) on the vector θ. This constraint can be stated by choosing a matrix with 1

row (number of redundant parameters) and k columns (number of parameters in the model). One possibility is to consider

$$M = \begin{pmatrix} 0 & 1 & 1 & 1 \end{pmatrix}.$$

The corresponding constraint is

$$M\theta = 0 \quad \Longleftrightarrow \quad \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 0\,.$$

In other words, we impose that the sum of the differential effects is zero. We can check that the conditions of the previous proposition are well satisfied: the suggested estimator can then be constructed.

See Chapter 11 for some explanations on the vocabulary used in the previous example.

8.1.2 Estimable Functions and Contrasts

Therefore, it is always possible to construct an estimator in the presence of a singular matrix. And so, what about the tests? In particular, are these constraints systematically necessary?

Most of the quantities we wanted to test are θ functions with no chosen identifiability constraints. We say that they are estimable because they are intrinsic.

Definition 8.3 (Estimable function) *A linear combination $C\theta$ is said to be an estimable function (of parameter θ) if it does not depend on the particular choice of a solution of the normal equations.*

We characterize these functions as those written $C\theta = DX\theta$, where D is a matrix of full rank.

Definition 8.4 (Contrast) *We call contrast an estimable function $C\theta$ such that $C\mathbb{1} = 0$, where $\mathbb{1}$ denotes the unit vector.*

In analysis of variance, most of the linear combinations we test are actually contrasts (see Chapter 11). In the previous example, $\alpha_1 - \alpha_2$ is a contrast.

8.2 Orthogonality

8.2.1 Orthogonality for Regular Models

Orthogonality is a notion that can significantly simplify the resolution and understanding of a linear model. A linear model usually admits a natural decomposition of the θ parameters (see example below) and, consequently, a decomposition of the X associated matrix. We focus here on the possible orthogonality of the different spaces associated with this decomposition (orthogonality will always be understood hereafter in the sense of orthogonality related to the usual Euclidean scalar product).

The problem will be more or less delicate depending on whether the model is regular. First, let us illustrate by two examples what parameter decomposition means.

Example 8.2 (Multiple linear regression) Let the multiple linear regression model on three variables $x^{(1)}$, $x^{(2)}$ and $x^{(3)}$: Given $n > 4$,

$$\forall i \in [\![1, n]\!], \quad Y_i = \mu + \theta_1 x_i^{(1)} + \theta_2 x_i^{(2)} + \theta_3 x_i^{(3)} + \varepsilon_i .$$

The vector θ has four coordinates – μ, θ_1, θ_2, θ_3 – and the matrix X has four columns. Hence, we naturally want to decompose X according to its column vectors. More precisely, we speak in this case of partition (of the matrix) into four elements, which amounts to writing the latter as a concatenation of 4 column vectors. The orthogonality of the partition then corresponds strictly to the orthogonality of the four vector spaces $[\mathbb{1}]$, $[x^{(1)}]$, $[x^{(2)}]$ and $[x^{(3)}]$.

Example 8.3 (Quadratic regression) Consider the quadratic regression model on $x^{(1)}$ and $x^{(2)}$: Given $n > 6$, for all $i \in [\![1, n]\!]$,

$$Y_i = \mu + \theta_1 x_i^{(1)} + \theta_2 x_i^{(2)} + \gamma_1 \left(x_i^{(1)}\right)^2 + \gamma_2 \left(x_i^{(2)}\right)^2 + \delta x_i^{(1)} x_i^{(2)} + \varepsilon_i .$$

Here, rather than asking, as before, for the orthogonality of each of the regressors (which would be a lot to ask), we can define the natural partition corresponding to:

- the constant μ;
- the linear effects θ_1, θ_2;
- the squared effects γ_1, γ_2;
- the product effect δ.

The orthogonality of the partition is then defined as the orthogonality of the vector subspaces: $[\mathbb{1}]$, $\left[\left(x^{(1)}, x^{(2)}\right)\right]$, $\left[\left(\left(x_i^{(1)}\right)^2, \left(x_i^{(2)}\right)^2\right)\right]$ and $\left[x^{(1)} x^{(2)}\right]$

Consequently, it is clear from these two examples that we should speak of a model with an orthogonal partition rather than an orthogonal model.

The following definition formalizes these examples.

Definition 8.5 *Consider a regular general linear model $Y = X\theta + \varepsilon$, and a partition into m terms of X and θ, i.e.*

$$Y = X_1\theta_1 + \ldots + X_m\theta_m + \varepsilon ,$$

where the matrix for all $j \in [\![1, m]\!]$, there exists $k_j \in [\![1, k]\!]$ such that $\sum_{j=1}^m k_j = k$, and $X_j \in \mathcal{M}_{n,k_j}\mathbb{R}$ and $\theta_j \in \mathbb{R}^{k_j}$ We say that this partition is orthogonal if the vector subspaces of $\mathbb{R}^n$, $[X_1], \ldots, [X_m]$, are orthogonal.

A simple consequence of the orthogonality of a linear model is that the information matrix ${}^tXX$ has a diagonal block structure, where each

block is associated with an element of the partition. Most often, the partition of the parameter vector θ into different effects comes from:

- in regression, from the different variables;
- in analysis of variance, from the decompositions into interactions.

Orthogonality gives statistical models the following two properties:

Proposition 8.4 *Let be a regular linear model with an orthogonal partition:* $Y = X_1\theta_1 + \ldots + X_m\theta_m + \varepsilon$ *Then:*

1. *The least squares estimators of the different effects* $\hat{\theta}_1, \ldots, \hat{\theta}_m$ *are uncorrelated and independent under the Gaussian hypothesis (H4);*
2. *For* $\ell \in [\![1, m]\!]$*, the expression of the* $\hat{\theta}_\ell$ *estimator does not depend on the presence or absence of the other* θ_j *terms in the model.*

Thus, orthogonality simplifies the computations, making it easy to obtain an explicit expression for the estimators. Moreover, it gives an approximate independence between the tests of the different effects: the tests on orthogonal effects are linked only by the estimate of σ^2.

8.2.2 Orthogonality for Singular Models

When the model is singular, it is necessary to consider the identifiability constraints. Therefore, we carry out the same partition procedure in orthogonal spaces, but while taking into account the system of constraints, *i.e.* the partition $C_j\theta_j = 0$ such that the applications$X_j|_{\mathrm{Ker}(C_j)}$ are injective.

Definition 8.6 *Consider the following partition of a linear model:*

$$Y = X_1\theta_1 + \ldots + X_m\theta_m + \varepsilon\,.$$

Let a system of constraints $C_1\theta_1 = 0, \ldots, C_m\theta_m = 0$ *which make the model idenifiable. We say that these constraints make the partition orthogonal if the vector subspaces*

$$\forall j \in [\![1, m]\!]\,, \quad V_j = \left\{X_j\theta_j \mid \theta_j \in \mathrm{Ker}(C_j)\right\}$$

are orthogonal.

This notion is close to the regular case. However, the notion of orthogonality depends here on the chosen constraints. The idea is to choose constraints that make the model orthogonal. We will see that this definition makes sense with the essential example of the analysis of variance model with two crossed factors (see Chapter 12).

9 Linear Regression

9.1 Introduction

9.1.1 Illustrative Example

To illustrate the concepts discussed in this chapter, we will consider the following example: We are interested in the possible relationship between a man's weight and various physical characteristics. For 22 healthy men aged 16 to 30, we have:

- Y: weight in kg;
- X_1: maximum circumference of the forearm in cm;
- X_2: maximum circumference of the biceps in cm;
- X_3: distance around the chest directly under the armpits in cm;
- X_4: distance around the neck, measured approximately halfway up, in cm;
- X_5: distance around the shoulders, measured at the point of the shoulder blades, in cm;
- X_6: distance around the waist at the trouser line, in cm;
- X_7: height from head to feet, in cm;
- X_8: maximum circumference of the calf in cm;
- X_9: circumference of the thigh, measured halfway between the knee and the top of the leg, in cm;
- X_{10}: circumference of the head in cm.

The dataset `mensurations.txt` illustrating this chapter is available on the moodle page of the course.

Figure 9.1 shows some descriptive statistics for our data set. We refer to Chapter 1 for an explanation of the different ways to represent quantitative variables.

9.1.2 Regression

Regression is one of the best-known and most widely applied statistical methods for analyzing quantitative data. It establishes a relationship between a quantitative variable and one or more other quantitative variables. Suppose we are interested in the relationship between two variables (e.g., weight versus maximum forearm circumference X_1). In that case, we speak of *simple regression* by expressing one variable as a function of the other. We refer to *multiple regression* if the relationship is between one variable and several other variables (e.g., weight as a function of all other quantitative variables). Implementing a regression requires a causal relationship between the variables considered in the model.

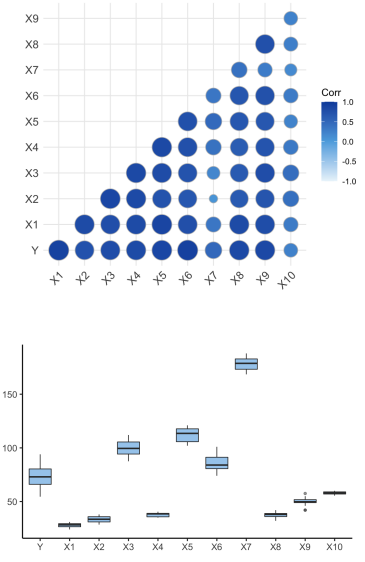


Figure 9.1: *Description of the data*: Bottom, boxplot of the different quantitative variables. Top, graphical representation of the two by two correlations of the quantitative variables.

This method can be implemented on quantitative data observed for n individuals and presented as

- a quantitative variable Y taking the value Y_i for the individual $i \in [\![1, n]\!]$ called the variable to be explained or *response variable*,
- p quantitative variables $x^{(1)}, \dots, x^{(p)}$ taking respectively the values $x_i^{(1)}, \dots, x_i^{(p)}$ for the individual i, called *explanatory variables* or *predictors*.

If $p = 1$, we are in the case of simple regression. When the values taken by an explanatory variable are chosen by the experimenter, we say that the explanatory variable is *controlled*.

In our example, $n = 22$, Y is the weight variable and $p = 10$

Consider a pair of quantitative random variables (X, Y). If there is a relationship between these two variables, the knowledge of the value taken by X modifies our uncertainty concerning the realization of Y: it generally decreases it. If we admit that there is a cause and effect relationship between X and Y, the random phenomenon represented by X can be used to predict the one represented by Y and the link is written in the form $\hat{Y} = f(X)$. We say that we regress Y on X. The challenge is to choose f wisely, so that the estimation of Y is unbiased, *i.e.* $\mathbb{E}[\hat{Y} - Y] = 0$, and with a minimal prediction error $\varepsilon = \hat{Y} - Y$.

In the most frequent cases, we choose the set of affine functions, *i.e.*

$$x \mapsto \theta_0 + \theta_1 x \quad \text{or} \quad \left(x^{(1)}, \dots, x^{(p)}\right) \mapsto \theta_0 + \theta_1 x^{(1)} + \dots + \theta_p x^{(p)},$$

and we speak of *linear regression*.

9.1.3 Simple Linear Regression Model

Consider a sample of n individuals. For an individual $i \in [\![1, n]\!]$, we have observed:

- Y_i the value of the quantitative variable Y (*e.g.* the weight),
- x_i the value of the quantitative variable x (*e.g.* the maximum circumference of the forearm)

We want to study the relationship between these two variables, and in particular, the effect of x (explanatory variable) on Y (response variable). First, we can represent this relationship graphically by drawing the cloud of n points with coordinates $(x_i, Y_i)_{i \in [\![1,n]\!]}$ (see figure 6.2). In the case where the point cloud is of "linear" form, we will try to fit this point cloud by a line. The relationship between Y_i and x_i is then written as a simple linear regression model:

$$\begin{cases} Y_i = \theta_0 + \theta_1 x_i + \varepsilon_i, \quad i \in [\![1, n]\!], \\ \varepsilon_1, \dots, \varepsilon_n \ \textit{i.i.d.} \text{ of law } \mathcal{N}(0, \sigma^2). \end{cases} \tag{9.1}$$

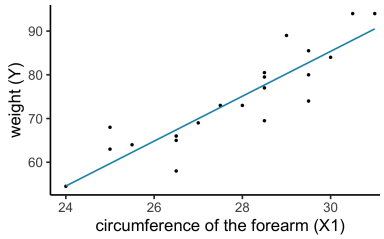


Figure 9.2: Plot of weight as a function of maximum forearm circumference. In blue, the fitted simple linear regression line.

The first part of the model $\theta_0+\theta_1 x_i$ represents the mean of Y_i given x_i and the second part ε_i the difference between this mean and the value of Y_i. The scatterplot is summarized by the line of equation $y = \hat{\theta}_0^{\text{obs}} + \hat{\theta}_1^{\text{obs}} x$.

Below, we present the results obtained with the `lm` command for this example of simple linear regression. In particular, $\hat{\theta}_0^{\text{obs}} = -68.644$ and $\hat{\theta}_1^{\text{obs}} = 5.134$. We can also read their respective standard error in the second column.

```
> reg.simple = lm(Y~X1,data=Data)
> summary(reg.simple)

Call:
lm(formula = Y ~ X1, data = Data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3981 -1.9234 -0.3646  2.8012  8.7678

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -68.644     15.589  -4.403 0.000274 ***
X1             5.134      0.560   9.167 1.34e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.926 on 20 degrees of freedom
Multiple R-squared:  0.8078,    Adjusted R-squared:  0.7981
F-statistic: 84.03 on 1 and 20 DF,  $p$-value: 1.338e-08
```

Listing 9.1: Simple linear regression

9.1.4 Multiple Linear Regression Model

Consider a sample of n individuals. For an individual $i \in [\![1, n]\!]$, we have observed:

- Y_i the value of the quantitative response variable Y (*e.g.* the weight),
- $x_i^{(1)}, \ldots, x_i^{(p)}$ the values of p other quantitative variables $x^{(1)}, \ldots, x^{(p)}$.

We want to explain the quantitative variable Y by the p quantitative variables $x^{(1)}, \ldots, x^{(p)}$. The model is written

$$\begin{cases} Y_i = \theta_0 + \theta_1 x_i^{(1)} + \ldots + \theta_p x_i^{(p)} + \varepsilon_i, \quad i \in [\![1, n]\!], \\ \varepsilon_1, \ldots, \varepsilon_n \ \textit{i.i.d.} \text{ of law } \mathcal{N}(0, \sigma^2). \end{cases} \tag{9.2}$$

Hereafter, the results obtained with the `lm` command for the multiple linear regression example. The first two columns correspond respectively to the estimates and the standard errors for each parameter.

```
> reg = lm(Y~.,data=Data)
> summary(reg)

Call:
lm(formula = Y ~ ., data = Data)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.5523 -0.9965  0.0461  1.0499  4.1719

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -69.51714   29.03739  -2.394 0.035605 *
X1            1.78182    0.85473   2.085 0.061204 .
X2            0.15509    0.48530   0.320 0.755275
X3            0.18914    0.22583   0.838 0.420132
X4           -0.48184    0.72067  -0.669 0.517537
X5           -0.02931    0.23943  -0.122 0.904769
X6            0.66144    0.11648   5.679 0.000143 ***
X7            0.31785    0.13037   2.438 0.032935 *
X8            0.44589    0.41251   1.081 0.302865
X9            0.29721    0.30510   0.974 0.350917
X10          -0.91956    0.52009  -1.768 0.104735
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.287 on 11 degrees of freedom
Multiple R-squared:  0.9772,    Adjusted R-squared:  0.9565
F-statistic: 47.17 on 10 and 11 DF,  p-value: 1.408e-07
```

Listing 9.2: Multiple linear regression

9.2 Estimation

9.2.1 General Results

The model (9.2) can be rewritten in the matrix form can be rewritten in the matrix form

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(p)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(p)} \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(p)} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{pmatrix}}_{\theta} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\varepsilon},$$

where $X \in \mathcal{M}_{n,p+1}\mathbb{R}$ (*i.e.* $k = p+1$). If the model is regular, we can estimate the vector of the parameters θ by the least-squares method. Hence,

$$\hat{\theta} = ({}^tXX)^{-1}\,{}^tXY \sim \mathcal{N}(0, \sigma^2({}^tXX)^{-1}).$$

We then deduce $\hat{Y}_i = (X\hat{\theta})_i = \hat{\theta}_0 + \sum_{j=1}^{p} \hat{\theta}_j x_i^{(j)}$ the adjusted value of Y_i and the residual $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$, of observed value $(\hat{\varepsilon}_i)^{\text{obs}} = y_i - \hat{y}_i$. The variance σ^2 is estimated by

$$\hat{\sigma}^2 = \frac{\|Y - X\hat{\theta}\|^2}{n-(p+1)} = \frac{1}{n-(p+1)} \sum_{i=1}^{n} (\hat{\varepsilon}_i)^2.$$

Moreover,

- the standard error of $\hat{\theta}_j$ is $se(\hat{\theta}_j) = \sqrt{\hat{\sigma}^2 \, [({}^t XX)^{-1}]_{j+1,j+1}}$,
- the standard error of $\hat{Y}_i$ is $se(\hat{Y}_i) = \sqrt{\hat{\sigma}^2 \, [X({}^t XX)^{-1} \, {}^t X]_{ii}} = \sqrt{\hat{\sigma}^2 \, H_{ii}}$,
- the standard error of $\hat{\varepsilon}_i$ is $se(\hat{\varepsilon}_i) = \sqrt{\hat{\sigma}^2 \, (1 - H_{ii})}$.

Exercise 9.1 *We assume the simple linear regression framework of equation* (9.1). *Show that the least squares estimators of θ_0 and θ_1 are given by:*

$$\begin{cases} \hat{\theta}_1 = \dfrac{\widehat{Cov}(Y, x)}{\widehat{Var}(x)} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} , \\ \hat{\theta}_0 = \bar{Y} - \hat{\theta}_1 \bar{x} , \end{cases}$$

where $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ *and* $\bar{Y} = \dfrac{1}{n}\sum_{i=1}^{n} Y_i$.

***Indications**: Minimize the least squares function*

$$(a, b) \mapsto \sum_{i=1}^{n}(Y_i - a - bx_i)^2 .$$

9.2.2 Properties in Simple Linear Regression

In this section, we consider the framework of a simple linear regression (Equation (9.1)). The following proposition gives properties between the residuals and the values predicted by the model.

Proposition 9.2 *Consider a simple linear regression model.*

1. $\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$ *and* $\sum_{i=1}^{n} \hat{Y}_i = \sum_{i=1}^{n} Y_i$,
2. *The regression line passes through the point with coordinates* $(\bar{x}, \bar{Y})$,
3. *The vector of residuals is not correlated with the explanatory variable:*

$$\widehat{Cov}(x, \hat{\varepsilon}) = 0 ,$$

4. *The vector of residuals is not correlated with the fitted variable:*

$$\widehat{Cov}(\hat{Y}, \hat{\varepsilon}) = 0\,,$$

5. *The variance of Y admits the decomposition*

$$\widehat{Var}(Y) = \widehat{Var}(\hat{Y}) + \widehat{Var}(\hat{\varepsilon})\,,$$

6. *The square of the correlation coefficient of x and Y is written in the following forms:*

$$r^2(x, Y) = \frac{\widehat{Var}(\hat{Y})}{\widehat{Var}(Y)} = 1 - \frac{\widehat{Var}(\hat{\varepsilon})}{\widehat{Var}(Y)}\,.$$

We deduce that the empirical variance of Y is the sum of an explained variance $\widehat{Var}(\hat{Y})$ and a residual variance $\widehat{Var}(\hat{\varepsilon})$, and that $r^2(x, Y)$ is the ratio between the explained variance and the total variance.

Exercise 9.3 *Prove Proposition 9.2. You may freely use the closed forms for θ_0 and θ_1 obtained in Exercise 9.1.*

9.2.3 The R^2 Coefficient

9.2.3.1 Definition

The R^2 coefficient, defined as the square of the correlation coefficient of x and Y, is a measure of the goodness of fit, equal to the ratio of the

variance actually explained to the variance to be explained:

$$R^2 = r^2(x, Y) = \frac{\widehat{\mathcal{V}ar}(\hat{Y})}{\widehat{\mathcal{V}ar}(Y)}.$$

Thus, $R^2 \in [0, 1]$ is interpreted as the *proportion of variance explained by the regression*.

Note that the decomposition (5) in Proposition 9.2 is identical to that introduced in Section 6.7, one being obtained by multiplying by n the other. As a reminder:

$$SST = SSE + SSR,$$

where[1]

- $SST = \|Y - \bar{Y}\mathbb{1}_n\|^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = n\,\widehat{\mathcal{V}ar}(Y)$ is the total sum of the (corrected) squares of Y,
- $SSR = \|\hat{Y} - \bar{Y}\mathbb{1}_n\|^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = n\,\widehat{\mathcal{V}ar}(\hat{Y})$ is the sum of squares explained by the model, or regression sum of squares,
- $SSE = \|Y - \hat{Y}\|^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = n\,\widehat{\mathcal{V}ar}(\hat{\varepsilon})$ is the sum of squares of the residuals, or error sum of squares.

1: whose acronyms in English are always so tricky compared to the French. *Cf.* Note 1 page 70.

Hence, to compute R^2, we often use the expression

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

In the simple linear regression example, the R^2 value is 0.8078. We read this value in the corresponding R code (Listing 21), second to last line. To find the values of SST, SSR and SSE, we can use the `anova` command.

```
> anova(reg.simple)
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
X1         1 2038.88 2038.88  84.032 1.338e-08 ***
Residuals 20  485.27   24.26
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the case of a multiple regression of Y by $x^{(1)}, \ldots, x^{(p)}$, the multiple correlation coefficient denoted $r(Y, x^{(1)}, \ldots, x^{(p)})$ is defined as the empirical linear correlation coefficient of Y by $\hat{Y}$:

$$r(Y, x^{(1)}, \ldots, x^{(p)}) = r(Y, \hat{Y}).$$

Hence, the coefficient R^2 of the multiple regression is equal to the square of the empirical $r(Y, x^{(1)}, \ldots, x^{(p)})$. Here, in the multiple linear regression example, the R^2 value is 0.9772 (See Listing 30).

Note that the R^2 is significantly better in the multivariate regression case than in the simple regression case. Actually, this observation is true overall, as described in the following graph.

9.2.3.2 Mechanical Increase of R^2

When an explanatory variable is added to a model, the sum of squares of the residuals decreases or at least remains stable. Indeed, if we consider a model with $p-1$ variables

$$Y_i = \theta_0 + \theta_1 x_i^{(1)} + \ldots + \theta_{p-1} x_i^{(p-1)} + \varepsilon_i$$

then the estimated coefficients $(\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_{p-1})$ minimize

$$\phi(\theta_0, \theta_1, \ldots, \theta_{p-1}) = \sum_{i=1}^{n} \left[Y_i - \left(\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_{p-1}\right) \right]^2 .$$

If we add a new explanatory variable $x^{(p)}$ to the model, we obtain

$$Y_i = \theta_0 + \theta_1 x_i^{(1)} + \ldots + \theta_{p-1} x_i^{(p-1)} + \theta_p x^{(p)} + \varepsilon_i$$

and the estimated coefficient, denoted $(\tilde{\theta}_0, \tilde{\theta}_1, \ldots, \tilde{\theta}_{p-1})$ minimize

$$\psi(\theta_0, \theta_1, \ldots, \theta_p) = \sum_{i=1}^{n} \left[Y_i - \left(\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_p\right) \right]^2 ,$$

which, by construction, satisfies the equality

$$\psi(\theta_0, \theta_1, \ldots, \theta_{p-1}, 0) = \phi(\theta_0, \theta_1, \ldots, \theta_{p-1}) .$$

Hence the inequality:

$$\psi(\tilde{\theta}_0, \tilde{\theta}_1, \ldots, \tilde{\theta}_p) \leqslant \psi(\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_{p-1}, 0) = \phi(\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_{p-1}) .$$

This proves the "mechanical" increase of R^2 but without improving the model, as we will see later.

9.3 Tests of the Nullity of the Model Parameters

In this paragraph, we investigate whether the proposed model can be simplified or not, *i.e.* whether some explanatory variables $x^{(j)}$ are negligible.

9.3.1 Nullity of *a* Model Parameter

We want to study the effect of the presence of a given explanatory variable $x^{(j)}$, where $j \in [\![1, p]\!]$. To do this, we test :

$$\mathcal{H}_0^{(j)} : \text{“}\theta_j = 0\text{”} \qquad \text{against} \qquad \mathcal{H}_1^{(j)} : \text{“}\theta_j \neq 0\text{”} ,$$

where θ_j is the parameter associated to the variable $x^{(j)}$. We set up for this purpose a classical Student's test.

Exercise 9.4 *Construct Student's statistical test to test* $\mathcal{H}_0^{(j)} : \text{“}\theta_j = 0\text{”}$ *against* $\mathcal{H}_1^{(j)} : \text{“}\theta_j \neq 0\text{”}$ *at level* α.

In the previous examples of simple and multiple linear regression, the last column of the R outputs shows the p-value associated with the nullity test for each of the θ_j coefficients; the second-to-last column displays the values of the test statistics. According to the R output shown on Listing 21, we strongly reject the nullity of each of the coefficients in the simple regression model for a 5% test. Similarly, according to the output shown on Listing 30, we reject the nullity of the coefficients θ_0, θ_6, and θ_7 in the multiple linear regression example for a 5% test.

Warning! Each nullity test is performed separately. So, *beware of quick conclusions!*

9.3.2 Nullity of *Some* Model Parameters

Consider a reference model with p explanatory variables. We want to study the influence of q explanatory variables (with $q \leqslant p$) on the variable to be explained. This amounts to testing the nullity hypothesis of q parameters of the model:

$$\mathcal{H}_0 : \text{“}\theta_1 = \theta_2 = \ldots = \theta_q = 0\text{”} , \qquad \text{where } q \leqslant p .$$

Under the alternative hypothesis, at least one of the parameters $\theta_1, \theta_2, \ldots, \theta_q$ is non-zero.

This test can be formulated as the comparison of two nested models, one with $p + 1$ parameters and the other with $p + 1 - q$ parameters:

- Under $\mathcal{H}_0$,

$$Y_i = \theta_0 + \theta_{q+1} x_i^{(q+1)} + \ldots + \theta_p x_i^{(p)} + \varepsilon_i \,, \qquad (M_0)$$

- Under $\mathcal{H}_1$,

$$Y_i = \theta_0 + \theta_1 x_i^{(1)} + \ldots + \theta_p x_i^{(p)} + \varepsilon_i \,. \qquad (M_1)$$

The $\mathcal{H}_0$ hypothesis can therefore be tested using the Fisher statistic:

$$F = \frac{\frac{SSE_0 - SSE_1}{q}}{\frac{SSE_1}{n-(p+1)}} \overset{\mathcal{H}_0}{\sim} \mathcal{F}\big(q, n-(p+1)\big) \,,$$

where SSE_0 is the error sum of squares of the "reduced" model (M_0) under $\mathcal{H}_0$ and SSE_1 is the error sum of squares of the reference model (M_1). We compare F to the quantile $f_{q,n-p-1,1-\alpha}$. If $F \geqslant f_{q,n-p-1,1-\alpha}$, then we reject $\mathcal{H}_0$.

Note that in the case where $q = 1$, we test the nullity of a single parameter of the model, and we find the same conclusions as with the previous Student's t-test.

In our multiple linear regression example, we want to test the submodel composed only of the variables X_1, X_6, and X_7. Using the `anova` function, we will perform a Fisher test between this sub-model and the full model.

```
> reg0 = lm(Y~X1+X6+X7,data=Data)
> anova(reg0,reg)
Analysis of Variance Table

Model 1: Y ~ X1 + X6 + X7
Model 2: Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10
Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     18 112.149
2     11  57.524  7    54.625 1.4922 0.2653
```

The p-value being 0.2653, we accept the sub-model (M_0).

Exercise 9.5 *In the output* `R` *of* `anova(reg0,reg)` *above, what does each of the numerical values correspond to?*

9.3.3 Nullity of *all* Model Parameters

In this section, we want to test the null hypothesis of all the parameters of the model (associated with the explanatory variables):

$$\mathcal{H}_0 : \text{“}\theta_1 = \theta_2 = \ldots = \theta_p = 0\text{”} \,.$$

This test compares the goodness of fit of the reference model with that of the "white model". This hypothesis, composed of p constraints, means that the p parameters associated with the p explanatory variables are zero, *i.e.* that no explanatory variable present in the model can explain the variable Y. Under $\mathcal{H}_0$, the model is written :

$$Y_i = \theta_0 + \varepsilon_i \qquad \text{and} \qquad \hat{\theta}_0 = \bar{Y}\,.$$

Moreover, the error sum of squares SSE_0 is equal to the total sum of squares SST.

Exercise 9.6 *Show that Fisher's test statistic in this case is written*

$$F = \frac{\frac{SSR_1}{p}}{\frac{SSE_1}{n} - (p+1)} = \frac{R^2}{1-R^2} \times \frac{n-p-1}{p} \underset{\sim}{\mathcal{H}_0} \mathcal{F}(p, n-p-1)\,,$$

where SSR_1 is the regression sum os squares of the reference model, and R^2 is the fit criterion of the reference model.

We compare F to the quantile $f_{p,n-p-1,1-\alpha}$: If $F \geqslant f_{p,n-p-1,1-\alpha}$, then we reject $\mathcal{H}_0$, and we conclude that there is at least one non zero parameter in the model.

In the example of the multiple linear regression, we can implement this test with the `anova` function. We can also notice that the result of this test is given directly in `summary(reg)` (see Listing 30).

```
> regwhite = lm(Y~1,data=Data)
> anova(regwhite,reg)
Analysis of Variance Table

Model 1: Y ~ 1
Model 2: Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     21 2524.15
2     11   57.52 10    2466.6 47.168 1.408e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, the p-value is $1.408\,\mathrm{e}^{-07}$, so we reject the hypothesis that all coefficients are zero.

9.4 Confidence Intervals

9.4.1 Confidence Interval for θ_j

Listing 9.3: Confidence interval for simple regression

```
>confint(reg.simple,level=.9)
                5 %    95 %
(Intercept) -95.530 -41.758
X1            4.168   6.099
```

Listing 9.4: Confidence interval for multiple regression

```
> confint(reg)
                 2.5%  97.5%
(Intercept) -133.428 -5.606
X1            -0.099  3.663
X2            -0.913  1.223
X3            -0.308  0.686
X4            -2.068  1.104
X5            -0.556  0.498
X6             0.405  0.918
X7             0.031  0.605
X8            -0.462  1.354
X9            -0.374  0.969
X10           -2.064  0.225
```

We follow the general construction made in Section 6.5.1. Here $k = p + 1$. Using that

- $\hat{\theta}_j \sim \mathcal{N}\left(\theta_j, \sigma^2 [{}^tXX]^{-1}_{j+1,j+1}\right)$,
- $\big(n-(p+1)\big)\,\hat{\sigma}^2 \sim \sigma^2\,\chi\big(n-(p+1)\big)$,
- $\hat{\theta}_j$ and $\hat{\sigma}_j$ are independent,

we get that

$$\frac{\hat{\theta}_j - \theta_j}{\hat{\sigma}\sqrt{[{}^tXX]^{-1}_{j+1,j+1}}} \sim \mathcal{T}\big(n-(p+1)\big)\,.$$

We then construct the confidence interval for the parameter θ_j at the $1-\alpha$ confidence level as follows:

$$CI_{1-\alpha}(\theta_j) = \left[\hat{\theta}_j \pm t_{n-(p+1),1-\alpha/2} \times \hat{\sigma}\sqrt{[{}^tXX]^{-1}_{j+1,j+1}}\right].$$

In R, we can easily obtain the confidence intervals for the θ_j coefficients using the `confint` function.

9.4.2 Confidence Interval for $(X\theta)_i$

Using the construction made in Section 6.5.2, the confidence interval of $(X\theta)_i$ at the confidence level of $1-\alpha$ is therefore given by:

$$CI_{1-\alpha}\big((X\theta)_i\big) = \left[\hat{Y}_i \pm t_{n-(p+1),1-\alpha/2} \times \hat{\sigma}\sqrt{[X({}^tXX)^{-1}\,{}^tX]_{i,i}}\right].$$

9.4.3 Confidence Interval for $X_0\theta$

For new observations $x_0^{(1)}, \ldots, x_0^{(p)}$ of the explanatory variables, we define $X_0 = (1, x_0^{(1)}, \ldots, x_0^{(p)}) \in \mathcal{M}_{1,p+1}\mathbb{R}$. The average response is then

$$X_0\theta = \theta_0 + \sum_{j=1}^{p} \theta_j x_0^{(j)}\,.$$

Using the construction made in Section 6.6.1, we obtain the confidence interval of $X_0\theta$ at the confidence level of $1-\alpha$:

$$CI_{1-\alpha}(X_0\theta) = \left[X_0\hat{\theta} \pm t_{n-(p+1),1-\alpha/2} \times \hat{\sigma}\sqrt{X_0({}^tXX)^{-1}\,{}^tX_0}\right].$$

In the simple linear regression example, see Figure 9.3.

9.5 Prediction Interval

We want to predict in which interval the result of a new trial $x_0^{(1)}, \ldots, x_0^{(p)}$ will lie. So, we want to construct a prediction interval for a new observation Y_0, corresponding to $X_0 = (1, x_0^{(1)}, \ldots, x_0^{(p)})$:

$$Y_0 = X_0\theta + \varepsilon_0 ,$$

where ε_0 is independent of all the ε_i, $i \in [\![1, n]\!]$, and distributed according to a law $\mathcal{N}(0, \sigma^2)$. Using the construction made in Section 6.6.2, we obtain that the prediction interval of the variable Y for a new observation X_0 is defined by

$$CI_{1-\alpha}(Y_0) = \left[X_0\hat{\theta} \pm t_{n-(p+1),1-\alpha/2} \times \hat{\sigma} \sqrt{1 + X_0({}^tXX)^{-1}\,{}^tX_0} \right] .$$

Carefully note the difference between $CI_{1-\alpha}(Y_0)$ and

$$CI_{1-\alpha}(X_0\theta) = \left[X_0\hat{\theta} \pm t_{n-(p+1),1-\alpha/2} \times \hat{\sigma} \sqrt{X_0({}^tXX)^{-1}\,{}^tX_0} \right] .$$

In the simple linear regression example, the confidence intervals $CI_{1-\alpha}(Y_0)$ and $IIC_{1-\alpha}(X_0\theta)$ are shown in Figure 9.3.

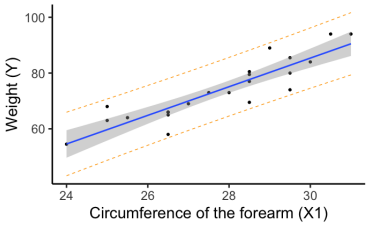


Figure 9.3: Prediction interval of Y, in dashed orange lines, and confidence interval of $X\theta_0$, in gray, for the simple linear regression model.

Remark 9.1 To make predictions using this linear regression model, we advise you to use this model only in the domain covered by the data. Indeed, the studied phenomenon can be linear in the observed domain and have a different behavior in another domain.

9.6 Selection of Explanatory Variables

In the presence of p explanatory variables for which we do not know which ones are really influential, we must look for a model to explain Y that is both efficient (smallest possible residuals) and economical (fewest possible explanatory variables).

We will now focus on the study of the X matrix, *i.e.* the explanatory variables themselves. In this part, we will see how to choose the model that best fits our data and eliminate certain variables that are not very explanatory to gain in interpretation. This problem of variable selection is, in fact, a model selection problem.

9.6.1 General Framework for Model Selection

For the sake of simplicity, we present this problem in the context of multiple linear regression. The tools developed here are more general and can be used in a broader framework, often without additional work.

We give ourselves a family of models $\mathcal{M}$ formally representing a family of subsets of $[\![1, p]\!]$. This choice is made a priori and may not be exhaustive. For example, we can consider

- *Exhaustive family*: $\mathcal{M} = \mathcal{P}(\{1, \dots, p\})$ *i.e.* the family of all subsets of $[\![1, p]\!]$,
- *Growing family*: $\mathcal{M} = \mathcal{P}(\{1, \dots, m\})_{m \in [\![1,p]\!]}$.

In the following, for $m \in \mathcal{M}$, we will note $|m|$ the cardinal of m. Let $X_{(m)}$ be the matrix consisting of the vectors $x^{(j)}$ for $j \in m$. We will also assume that for all $m \in \mathcal{M}$ the matrix $X_{(m)}$ is regular, *i.e.* of rank $|m| + 1$. Note that the "+1" comes from the constant (of the intercept) which is assumed to be systematically present in all models.

Assumptions about the true model: We assume that there exists $m^* \in \mathcal{M}$, unknown, such that the true model is written :

$$Y = \mu^* + \varepsilon^* = X_{(m^*)}\theta_{(m^*)} + \varepsilon^*, \text{ where } \varepsilon^* \sim \mathcal{N}_n(0_n, \sigma^{*2} I_n),$$

and where the vector $\theta_{(m^*)} \in \mathbb{R}^{|m^*|+1}$ has all its coordinates non-zero.

The idea of model selection techniques is to find the model from our given collection that best explains our data. In other words, we are looking for the best estimator, $m_0 \in \mathcal{M}$, for the collection of models we have chosen. Note that the modeler has chosen this family. This leaves the possibility to add modeling constraints. But, it also implies that a wrong choice concerning this collection of models will lead to an inaccurate estimation of the true model.

Analysis models: Let the family of models m defined by,

$$Y = \mu_{(m)} + \varepsilon = X_{(m)}\theta_{(m)} + \varepsilon, \text{ where } \varepsilon \sim \mathcal{N}(0_n, \sigma^2 I_n),$$

in correspondence with $\mathcal{M}$.

To specify the modeling, we will use the following vocabulary:

Definition 9.1 *Let an analysis model $m \in \mathcal{M}$*

- *If $m = m_p = [\![1, p]\!]$, the model is said to be* complete, *i.e. all available explanatory explanatory variables are significant;*
- *If $m^* \subset m$ with $m \neq m^*$, we say that the model is* over-fitted*;*
- *If $|m \cap m^*| < |m^*|$, we say that the model is* wrong*;*
- *If $m \subset m^*$ with $m \neq m^*$, we say that the model is* under-fitted.

Recall that each model corresponds to a choice among all the explanatory variables, and that there are therefore potentially superfluous explanatory variables. In case of *over-fitting*, *i.e.* if there are superfluous variables, an over-fitted model is a model containing all the variables of the true model plus a certain number of superfluous variables. A *false model* is typically a model where not all the variables of the true model have been chosen and some superfluous variables may have been chosen. A special case is the *sub-fitting* corresponding to a false model containing no superfluous variables.

9.6.2 Some Criteria to Select a Model

To give a meaning to "best", we need a criterion to quantify the quality of an estimator. In other words, we give ourselves a function R able to quantify the gap between m and the true model m^*, and we try to minimize this risk. In the following, we develop several ways to define these selection criteria. Note that these criteria do not allow to find m^*, but only to approach it.

This corresponds to the basics of *model selection*. For more details, see for example [7].

9.6.2.1 The Adjustment Coefficients

In the situation where only a small number of regressors are involved, there are already several approaches that are more or less directly inspired by the tools studied above. To "test" the validity of a sub-model m with respect to a larger model, there are two indices (or coefficients) whose calculation and interpretation are pretty immediate.

A first possibility is to focus on the coefficient of determination:

$$R^2_m = \frac{SST - SSE(m)}{SST} = 1 - \frac{\|Y - X_{(m)}\hat{\theta}_{(m)}\|^2}{\|Y - \bar{Y}\mathbb{1}_n\|^2}.$$

Therefore, this index compares the fitted values of Y with the observed values through $\|\hat{Y}_{(m)} - Y\|^2$, the denominator corresponding to a renormalization. The closer the R coefficient is to 1, the better the fit of the model to the data. If one has to choose between two explanatory models, one is easily tempted to select the one with the higher coefficient of determination.

However, it is important to temper this type of reasoning. Indeed, the maximization of this criterion R^2_m amounts to maximizing $\|Y - \hat{Y}_{(m)}\|^2$, and it is clear that the quantity $\|Y - \hat{Y}_{(m)}\|^2 = \|P_{[X^{\perp}_{(m)}]}\|^2$ decreases for a nested sequence of models. Therefore, maximizing R^2_m leads for sure to choose the complete model m_k. The use of this type of criterion thus favors the selection of strongly parameterized models. On the other hand, for models with the same cardinal $|m|$, this coefficient can be used to select an optimal model.

It is possible to improve the R^2_m coefficient to allow the selection of models with a different number of explanatory variables by defining the adjusted determination coefficient $\tilde{R}^2_m$. This coefficient enables to take into account the number of selected regressors and thus proposes a compromise between the adequacy and the parameterization of the model. This index is defined by:

$$\tilde{R}^2_m = 1 - \frac{n-1}{n-|m|-1}\frac{SSE(m)}{SST} = 1 - \frac{n-1}{n-|m|-1}\frac{\|Y - X_{(m)}\hat{\theta}_{(m)}\|^2}{\|Y - \bar{Y}\mathbb{1}_n\|^2}.$$

The interpretation is similar to that of R^2_m.

9.6.2.2 Bottom-Up and Top-Down Strategy

In the presence of a small number of models, the adjustment coefficient is a possible option. Otherwise, one can use a strategy based on Fisher's test and called top-down regression. The methodology is as follows: we start with the model using all possible regressors. At each step, we compute the Fisher statistic corresponding to the deletion of each of the variables still present. We then delete the variable with the smallest value, *i.e.* with the largest p-value. In fact, at each step, we remove the least significant variable in the sense of the Fisher test. We then repeat this process until all the statistics are above a predetermined threshold, *i.e.* when all the p-values are below a predetermined threshold, for example 5%.

Beware, this strategy can be extremely cumbersome to implement depending on the number of variables in question (we can go up to $|m|!$ Fisher tests).

 # Initialization: Let a threshold s and $m_{[0]} = \{1, \dots, p\}$.

 # Iteration t:

Step 1: For any $j \in m_{[t]}$, we compute the p-value p_j of the Fisher sub-model test of

$$M_0 : \ m_{[t]} \setminus \{j\} \quad \text{against} \quad M_1 : \ m_{[t]} \, ;$$

Step 2: $\hat{j} = \underset{j \in m_{[t]}}{\operatorname{argmax}} \, p_j$;

Step 3:

- If $p_j > s$, $m_{[t+1]} = m_{[t]} \setminus \{j\}$ and we go back to step 1,
- Else, `STOP`.

Model selection by bottom-up regression uses exactly the same arguments, except that we start with an empty model (without regressor, only the intercept), and we add the most significant variables (in the sense of Fisher's test) until the p-values exceed a previously fixed threshold.

9.6.2.3 Oracle Estimator

The ℓ^2-risk is an usual criterion to measure the difference between the true model m^* and an analysis model $m \in \mathcal{M}$.

Definition 9.2 *Let $m \in \mathcal{M}$. The ℓ^2-risk, or quadratic risk, between models m and m^* is defined by:*

$$\mathcal{R}(m, m^*) = \mathbb{E}\left[\left\|\mu^* - \hat{Y}_{(m)}\right\|^2\right] = \mathbb{E}\left[\left\|X_{(m^*)}\theta_{(m^*)} - X_{(m)}\hat{\theta}_{(m)}\right\|^2\right],$$

where $\mu^* = X_{(m^*)}\theta_{(m^*)}$ and $\hat{Y}_{(m)} = X_{(m)}\hat{\theta}_{(m)}$.

For any $m \in \mathcal{M}$, we define $\mu^*_{(m)} = P_{[X_{(m)}]}\mu^*$, the orthogonal project of μ^* on the vector space $\mathcal{I}m(X_{(m)})$. It is then possible to compute the ℓ^2-risk explicitly.

Proposition 9.7 *For all $m \in \mathcal{M}$,*

$$\mathcal{R}(m, m^*) = \sigma^{*2}(|m| + 1) + \|\mu^*_{(m)} - \mu^*\|^2 .$$

Proof. Let $m \in \mathcal{M}$ and $\mu^*_{(m)} = P_{[X_{(m)}]}\mu^*$. We have:

$$\begin{aligned}
\mathcal{R}(m, m^*) &= \mathbb{E}\left[\left\|X_{(m)}\hat{\theta}_{(m)} - \mu^*\right\|^2\right] \\
&= \mathbb{E}\left[\left\| \underbrace{X_{(m)}\hat{\theta}_{(m)} - \mu^*_{(m)}}_{\in \mathcal{I}m(X_{(m)})} + \underbrace{\mu^*_{(m)} - \mu^*}_{\in \mathcal{I}m(X_{(m)})^\perp} \right\|^2\right] \\
&= \mathbb{E}\left[\left\|X_{(m)}\hat{\theta}_{(m)} - \mu^*_{(m)}\right\|^2\right] + \mathbb{E}\left[\left\|\mu^*_{(m)} - \mu^*\right\|^2\right] \\
&\qquad \text{according to the Pythagorean theorem} \\
&= \mathbb{E}\left[\left\|X_{(m)}\hat{\theta}_{(m)} - \mu^*_{(m)}\right\|^2\right] + \left\|\mu^*_{(m)} - \mu^*\right\|^2 .
\end{aligned}$$

Yet,

$$X_{(m)}\hat{\theta}_{(m)} = P_{[X_{(m)}]}Y = P_{[X_{(m)}]}\left(X_{(m^*)}\theta_{(m^*)} + \varepsilon^*\right) = \mu^*_{(m)} + P_{[X_{(m)}]}\varepsilon^* .$$

Hence,

$$\left\|X_{(m)}\hat{\theta}_{(m)} - \mu^*_{(m)}\right\|^2 = \left\|P_{[X_{(m)}]}\varepsilon^*\right\|^2 \sim \sigma^{*2}\chi^2(|m| + 1)$$

from Cochran's theorem. Finally,

$$\mathbb{E}\left[\left\|X_{(m)}\hat{\theta}_{(m)} - \mu^*_{(m)}\right\|^2\right] = \sigma^{*2}(|m| + 1) .$$

□

Therefore, to minimize the distance between m and m^*, there is a compromise to be found. If $|m|$ is small, it will be the same for the variance term $\sigma^{*2}(|m| + 1)$, at the expense of the bias term $\|\mu^*_{(m)} - \mu^*\|^2$. On the contrary, for large values of $|m|$, one can hope to have a slight bias, but at the risk of having a more significant error, which is reflected in an increase of the $\sigma^{*2}(|m| + 1)$ term. This bias-variance trade-off is very classical in this model selection framework.

Remark 9.2 By definition of $\mu^*_{(m)}$, as soon as $m^* \subset m$, $\|\mu^*_{(m)} - \mu^*\|^2 = 0$.

To summarize, we have a measure of the quality of an estimator through its ℓ^2-risk $\mathcal{R}$ and our goal is to minimize this risk. Each model m has a risk $r_m = R(m)$, and the best model in term of the ℓ^2-risk is the so-called

oracle model

$$m_0 \in \operatorname*{argmin}_{m \in \mathcal{M}} r_m \,. \tag{9.3}$$

The oracle estimator is the best estimator in term of the risk $\mathcal{R}$, so we would like to use this estimator to estimate m^*. Unfortunately, we cannot use it in practice, since it cannot be computed from the data only. Actually, m_0 depends on the collection of risks $\{r_m, m \in \mathcal{M}\}$, which is unknown to the statisticians since it depends on the unknown signal m^*.

A natural idea to circumvent this issue is to replace the risk r_m in (9.3) by some estimator $\hat{r}_m$ of the risk and therefore estimate m^* by

$$\hat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \hat{r}_m \,.$$

The estimate $\hat{m}$ can be computed from the data only, but we have a priori no guarantee that it performs well. The main challenge now is to provide some suitable $\hat{r}_m$ for which we can guarantee that the selected estimator $\hat{m}$ performs almost as well as the oracle m_0.

9.6.2.4 Mallows' C_p Criterion

Let a model $m \in \mathcal{M}$. According to the Pythagorean theorem and the Cochran theorem, we have:

$$\begin{aligned}
\mathbb{E}\left[\left\|Y - \hat{Y}_{(m)}\right\|^2\right] &= \mathbb{E}\left[\left\|Y - \mu^*_{(m)}\right\|^2\right] - \mathbb{E}\left[\left\|\hat{Y}_{(m)} - \mu^*_{(m)}\right\|^2\right] \\
&= \mathbb{E}\left[\left\|Y - \mu^* + \mu^* - \mu^*_{(m)}\right\|^2\right] - \mathbb{E}\left[\left\|\hat{Y}_{(m)} - \mu^*_{(m)}\right\|^2\right] \\
&= \mathbb{E}\left[\|Y - \mu^*\|^2\right] + \|\mu^* - \mu^*_{(m)}\|^2 - (|m| + 1)\sigma^{*2} \\
&= n\sigma^{*2} - \|\mu^* - \mu^*_{(m)}\|^2 - (|m| + 1)\sigma^{*2} \,.
\end{aligned}$$

In other words

$$\|\mu^* - \mu^*_{(m)}\|^2 = \mathbb{E}\left[\left\|Y - \hat{Y}_{(m)}\right\|^2\right] + (|m| + 1)\sigma^{*2} - n\sigma^{*2} \,.$$

Since we want to find the optimal model m, we can neglect the term $-n\sigma^{*2}$, which does not depend on m. Therefore, Mallows propose to estimate the bias term $\|\mu^* - \mu^*_{(m)}\|^2$ by $\|Y - \hat{Y}_{(m)}\|^2 + (|m| + 1)\sigma^{*2}$ [8].

If the variance of the target model m^* is known, we then obtain the criterion:

$$C_p(m) = \|Y - \hat{Y}_{(m)}\|^2 + 2|m|\sigma^{*2} \,,$$

and we will select the model $\hat{m}_{C_p}$ satisfying:

$$\hat{m}_{C_p} \in \operatorname*{argmin}_{m \in \mathcal{M}} C_p(m) \,.$$

In case of unknown variance, we use the estimator $\hat{\sigma}^2 = \hat{\sigma}^2_{(m_p)}$, where $m_p = [\![1, p]\!]$ is the model taking into account all the regressors.

9.6.2.5 The AIC and BIC Criteria

Mallows' C_p criterion is based on the attempt to minimize the distance between m and the true model in the sense of a quadratic risk. The AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) criteria are constructed to minimize the Kullback-Leibler divergence between the two models.

Definition 9.3 (Kullback-Leibler) *Let μ and ν be two probability measures dominated by the same measure (in our case the Lebesgue measure). The* Kullback divergence *between these two measures is given by:*

$$KL(\mu|\nu) = \mathbb{E}_\mu\left[\log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right].$$

First, note the non-symmetry of $KL(\cdot,\cdot)$. This is why we prefer to speak of divergence rather than distance. However, this divergence verifies, like any "classical" distance, the following properties:

- $KL(\mu|\nu) \geqslant 0$ for any measure μ and ν,
- $KL(\mu|\nu) = 0$ if and only if $\mu = \nu$.

Convexity arguments can prove these properties.

Akaike Information Criterion. More precisely, AIC is founded in *information theory*. Let us consider two candidate modelsn $\hat{m}_1$ and $\hat{m}_2$, to estimate m^*. If we knew m^*, we could find the information lost by using $\hat{m}_1$ to represent m^* by computing the Kullback-Leibler divergence, $KL(m^*|\hat{m}_1)$; similarly, computing $KL(m^*|\hat{m}_2)$ allows us to quantify the information lost by using $\hat{m}_1$ to represent m^*. We would then choose the best model by minimizing the lost information. However, this calculation is inaccessible because, by nature, m^* is unknown. On the other hand, Akaike has shown that it is possible to estimate whether using $\hat{m}_1$ rather than $\hat{m}_2$ leads to a more or less critical loss of information [10, 11]. The criterion developed by Akaike for this purpose writes

$$AIC(m) = -2\log \mathcal{L}_{(m)}^{\max} + 2|m| \quad \text{and} \quad \hat{m}_{AIC} \in \operatorname*{argmin}_{m\in\mathcal{M}} AIC(m),$$

where $\mathcal{L}_{(m)}^{\max}$ is the maximum value of the likelihood function for the model m.

We will not present here the theoretical construction of this AIC criterion. However, the proof is available in [13], for instance.

Note that AIC tells nothing about the absolute quality of a model, only the quality relative to other models. Thus, if all the candidate models fit poorly, AIC will not give any warning of that. Hence, after selecting a model via AIC, it is usually good practice to validate the absolute quality of the model.

Moreover, the proposed estimate is only asymptotically valid. Also, if the number of data points is small, there is a substantial probability that AIC will select models with too many parameters, *i.e.* that AIC will

overfit. To address such potential overfitting, AICc was developed: AIC with a correction for small sample sizes, namely

$$AIC_c(m) = AIC + n\,\frac{n+|m|-1}{n-|m|-3}\,.$$

Bayesian Information Criterion. The BIC criterion introduced by Schwarz [12], extends the general writing of the AIC criterion using the Bayesian viewpoint. The unknown parameter is no longer considered as a vector but as a random variable. An a priori law is then applied to estimate the "parameter". The approach then consists in trying to exploit this information for estimation. This approach theoretically brings more richness since the range of possible solutions is extended.

In concrete terms, we obtain a criterion really similar to the AIC formula but with a different penalty for the number of parameters. With AIC, the penalty is $2|m|$, while with BIC, the penalty is $|m|\log(n)$. This approach leads to the BIC criterion defined by:

$$BIC(m) = -2\log \mathcal{L}_{(m)}^{\max} + 2|m|\log(n) \quad \text{and} \quad \hat{m}_{BIC} \in \operatorname*{argmin}_{m\in\mathcal{M}} BIC(m)\,,$$

where $\mathcal{L}_{(m)}^{\max}$ is the maximum value of the likelihood function for the model m.

9.6.3 Variable Selection Algorithms

In practice, once a model selection criterion has been chosen, determining the "best" model by an exhaustive search is impossible because of the number of models to be explored. Therefore, we resort to step-by-step methods.

9.6.3.1 Top-Down/Backward Methods

We start from the model using the p explanatory variables, and we look, at each step of the algorithm, for the most relevant variable to delete according to the chosen criterion. We iterate the algorithm until we reach the empty set. Among the variables visited during the algorithm, the best one is selected according to the criterion.

Some algorithms stop when a given threshold is reached.

 # Initialization: $m_{[0]} = [\![1, p]\!]$.

 # Iteration t:

Step 1: For all $j \in m_{[t]}$, compute $c_j = \texttt{CRIT}\left(m_{[t]} \setminus \{j\}\right)$;

Step 2: $\hat{j} = \operatorname*{argmin}_{j\in m_{[t]}} c_j$;

Step 3: $m_{[t+1]} = m_{[t]} \setminus \{\hat{j}\}$

- If $m_{[t+1]} \neq \emptyset$, we go back to step 1,
- Else, `STOP`.

9.6.3.2 Bottom-Up/Forward Methods

We start with an empty set of variables, and we look, at each step of the algorithm, for the most relevant variable to add according to the chosen criterion. We iterate the algorithm until all the variables are integrated. Among the variables visited during the algorithm, the best one is selected according to the criterion.

As before, some algorithms stop when a given threshold is reached.

 # Initialization: $m_{[0]} = \emptyset$.

 # Iteration t:

Step 1: For all $j \in [\![1, p]\!] \setminus m_{[t]}$,
compute $c_j = \texttt{CRIT}\left(m_{[t]} \cup \{j\}\right)$;

Step 2: $\hat{j} = \underset{j \in m_{[t]}}{\operatorname{argmin}}\, c_j$;

Step 3: $m_{[t+1]} = m_{[t]} \cup \{\hat{j}\}$

- If $m_{[t+1]} \neq [\![1, p]\!]$, we go back to step 1,
- Else, `STOP`.

9.6.3.3 Stepwise Methods

From a given model, one selects a new variable (as for a bottom-up method). Then, one tries to eliminate one of the variables from the model (as for a top-down method), and so on. It is necessary to define for such a method an input and an output criterion.

We can also quote the method of the "*s best subsets*": We search exhaustively among all the subsets of s variables, the s best, in the sense of the considered criterion.

9.6.4 Back to our Dataset

In this section, we will illustrate in our example some variable selection strategies. Thanks to the `regsubsets` function, we can set up a bottom-up, top-down, or stepwise method. We can also choose a criterion among Mallows' C_p, the adjusted R^2, and the BIC criterion. We can also use the `stepAIC` function.

```
> library(leaps)
> select_bwd = regsubsets(Y~.,data=Data,nbest=1,nvmax=10,
                                              method="backward")
> summary(select_bwd)

Subset selection object
Call: regsubsets.formula(Y ~ ., data = Data, nbest = 1,
                                 nvmax = 10, method = "backward")
10 Variables  (and intercept)
    Forced in Forced out
X1      FALSE      FALSE
```

```
X2      FALSE      FALSE
X3      FALSE      FALSE
X4      FALSE      FALSE
X5      FALSE      FALSE
X6      FALSE      FALSE
X7      FALSE      FALSE
X8      FALSE      FALSE
X9      FALSE      FALSE
X10     FALSE      FALSE
1 subsets of each size up to 10
Selection Algorithm: backward
          X1  X2  X3  X4  X5  X6  X7  X8  X9  X10
1  ( 1 )  " " " " " " " " " " "*" " " " " " " " "
2  ( 1 )  "*" " " " " " " " " "*" " " " " " " " "
3  ( 1 )  "*" " " " " " " " " "*" "*" " " " " " "
4  ( 1 )  "*" " " " " " " " " "*" "*" " " " " "*"
5  ( 1 )  "*" " " " " " " " " "*" "*" " " " " "*"
6  ( 1 )  "*" " " " " " " " " "*" "*" "*" "*" "*"
7  ( 1 )  "*" " " "*" " " " " "*" "*" "*" "*" "*"
8  ( 1 )  "*" " " "*" "*" " " "*" "*" "*" "*" "*"
9  ( 1 )  "*" "*" "*" "*" " " "*" "*" "*" "*" "*"
10  ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" "*" "*"
```

Similar results are obtained if we execute the command

```
select_fwd = regsubsets(Y~.,data=Data,nbest=1,nvmax=10,
                                                method="forward")
```

on the second line.

In the above code, `nvmax` corresponds to the maximum number of predictors to incorporate in the model. For example, if `nvmax` = 10, as is the case here, the function will return up to the best 10-variables model, that is: the best 1-variable model, the best 2-variables model, *etc.*

The function `summary` reports the best variables for each model size. From the output above, an asterisk specifies that a given variable is included in the corresponding model. For example, it can be seen that the best 2-variables model contains only X_1 and X_6: "$Y \sim X_1 + X_6$". The best 3-variables model is "$Y \sim X_1 + X_6 + X_7$", and so forth. A natural question then arises: which of these best models should we ultimately choose for our predictive analysis?

To answer this question, we will review the BIC, Mallows' C_p criteria, and the adjusted R^2. To do this, we can look at the graphical tables of best subsets given by `regsubsets`. Figure Figure 9.4 displays, for each criterion, a table of models showing which variables are in each model. In addition, the models are ranked by the specified model selection statistic. Thus, the model in the first row is the optimal model for the corresponding criterion.

We can also search for the value of the optimal criterion, and refer to the diagram of the different models to read the optimal model.

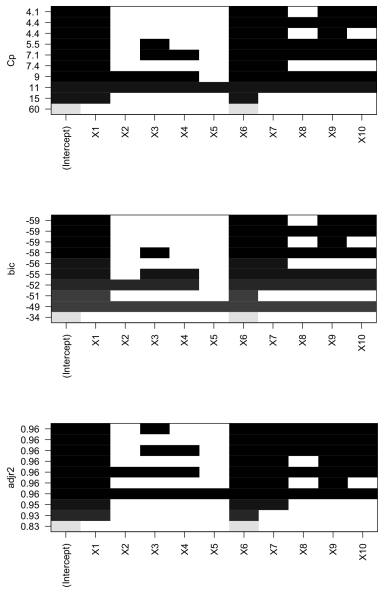


Figure 9.4: From top to bottom: result of the variable selection process with the Mallows' C_p criterion, the BIC criterion and the adjusted R^2.

```
> res = summary(select_bwd)
> str(data.frame(AdjR2 = which.max(res$adjr2),
        +  Cp = which.min(res$cp),
        +  BIC = which.min(res$bic)))
'data.frame': 1 obs. of  3 variables:
$ AdjR2: int 7
$ Cp   : int 5
$ BIC  : int 5
```

Finally, in our example, with the Mallows' C_p and the BIC criteria, we retain the model composed of the variables X_1, X_6, X_7, X_9, and X_{10}. The test of the sub-model confirms that this sub-model is sufficient to explain the variable Y. We obtain the same result with the AIC criterion. See the (long) output just after.

```
> reg.fin = lm(Y~X1 + X6 + X7 + X9 + X10, data=Data)
> anova(reg.fin,reg)
Analysis of Variance Table

Model 1: Y ~ X1 + X6 + X7 + X9 + X10
Model 2: Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     16 73.955
2     11 57.524  5    16.432 0.6284 0.6822
```

With the adjusted R^2, the selected model contains more variables, as expected. Namely: X_1, X_3, X_6, X_7, X_8, X_9, and X_{10}.

```
> library(MASS)
> modselect_aic = stepAIC(reg,trace=TRUE,direction=c("backward")
    )
Start:  AIC=43.15
Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10

       Df Sum of Sq     RSS    AIC
- X5    1     0.078  57.602 41.175
- X2    1     0.534  58.058 41.349
- X4    1     2.338  59.861 42.022
- X3    1     3.668  61.192 42.505
- X9    1     4.963  62.486 42.966
<none>               57.524 43.145
- X8    1     6.110  63.634 43.366
- X10   1    16.348  73.871 46.648
- X1    1    22.726  80.250 48.470
- X7    1    31.085  88.608 50.650
- X6    1   168.627 226.150 71.263

Step:  AIC=41.18
Y ~ X1 + X2 + X3 + X4 + X6 + X7 + X8 + X9 + X10

       Df Sum of Sq     RSS    AIC
- X2    1     0.586  58.188 39.398
- X4    1     2.367  59.969 40.061
- X3    1     4.689  62.291 40.897
<none>               57.602 41.175
- X8    1     6.426  64.028 41.502
```

```
- X9    1     6.538  64.140 41.541
- X10   1    18.606  76.208 45.333
- X1    1    33.697  91.299 49.308
- X7    1    36.863  94.465 50.058
- X6    1   174.761 232.363 69.860

Step:  AIC=39.4
Y ~ X1 + X3 + X4 + X6 + X7 + X8 + X9 + X10

       Df Sum of Sq     RSS    AIC
- X4    1     1.785  59.974 38.063
<none>               58.188 39.398
- X9    1     6.278  64.467 39.652
- X3    1     6.529  64.718 39.738
- X8    1     7.253  65.441 39.982
- X10   1    18.143  76.331 43.369
- X1    1    41.943 100.132 49.340
- X7    1    47.012 105.201 50.426
- X6    1   174.827 233.016 67.921

Step:  AIC=38.06
Y ~ X1 + X3 + X6 + X7 + X8 + X9 + X10

       Df Sum of Sq     RSS    AIC
- X3    1     4.748  64.722 37.739
<none>               59.974 38.063
- X9    1     7.028  67.002 38.501
- X8    1    10.607  70.581 39.646
- X10   1    17.091  77.065 41.579
- X1    1    43.614 103.588 48.086
- X7    1    46.538 106.512 48.699
- X6    1   178.038 238.011 66.388

Step:  AIC=37.74
Y ~ X1 + X6 + X7 + X8 + X9 + X10

       Df Sum of Sq     RSS    AIC
<none>               64.722 37.739
- X8    1     9.233  73.955 38.673
- X10   1    12.772  77.494 39.701
- X9    1    15.559  80.281 40.479
- X7    1    42.815 107.538 46.910
- X1    1    59.005 123.727 49.995
- X6    1   196.988 261.710 66.476
```

9.7 Validation of the Model

Once the model has been implemented, the "statistical soundness" of this model must be checked *a posteriori* regarding the normality of the residuals, the adequacy of the fitted value $\hat{Y}_i$ to the observed value Y_i. We can also ensure that there are no outliers.

9.7.1 Graphical Post Control

As a first step, a simple but effective technique consists in using a graphical control to empirically test the four basic postulates (at least the assumptions (A1 – 3), since the (H4) one is not so important as soon as sufficient data are available).

A first check we can make is to observe the graph of the n points $(y_i, \hat{y}_i)$. This graph is indeed very informative: if the points are aligned along the first bisector (see Figure 9.5), we can think that the linear regression model fits our situation.

In *simple linear regression*, the graphical comparison between the scatterplot (x_i, y_i) and the ordinary least squares regression line of Y by x gives almost exhaustive information (see Figure 9.2).
On this graph, if one observes a curvature of the "true" regression curve of Y, we can think that the model is inadequate and does not allow for testing assumption (A1).

However, in the case of *multiple regression*, this type of graph cannot be used because of the multiplicity of regressors. We must therefore check the different hypotheses one by one on the ε_i error terms. These ε_i are unfortunately unobservable. Hence, we will use their natural predictors, namely the residuals $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$.

Therefore, in the following paragraphs, we present several approaches to ensure the legitimacy of the conclusions, and adequate procedures for any multiple linear regression. These techniques are mainly based on the (graphical) analysis of the residuals. More precisely, we try to verify that the estimated residuals $\hat{\varepsilon}_i = Y_i - X_i\hat{\theta}$ behave in accordance with the model's hypotheses, *i.e.* that they have a random behavior close to iid random variables of Gaussian distribution.

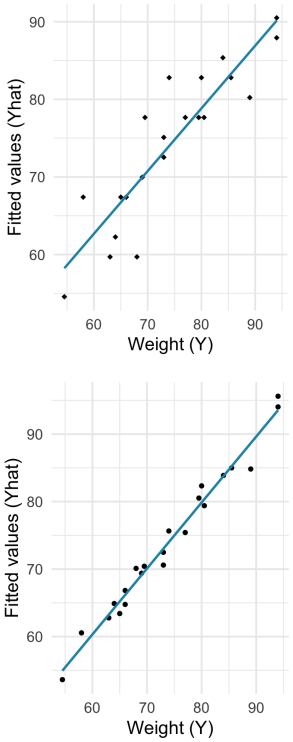


Figure 9.5: *Graphical post control*: Points $y_i, \hat{y}_i)$ for the example in simple (left) and multiple (right) linear regression.

9.7.2 (A1 – 2) Goodness of Fit & Homoscedasticity

The most classical graph used to check the adequacy of the model and the homoscedasticity is the graph of the residuals versus the fitted

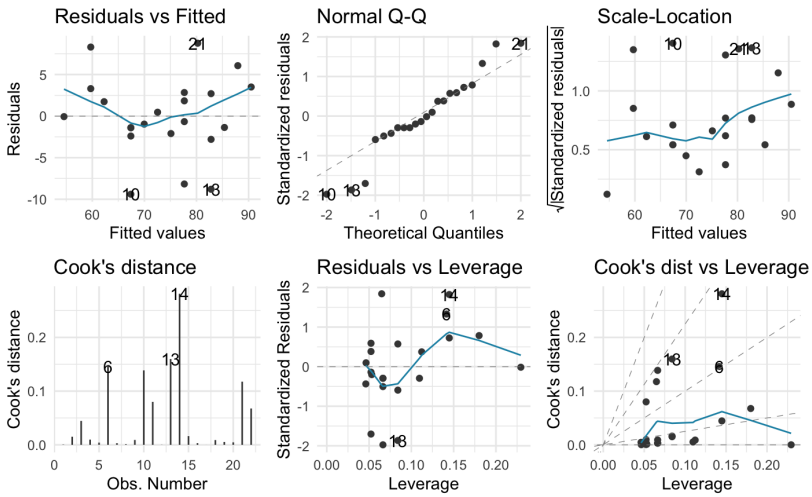


Figure 9.6: Diagnostics for the simple linear regression example

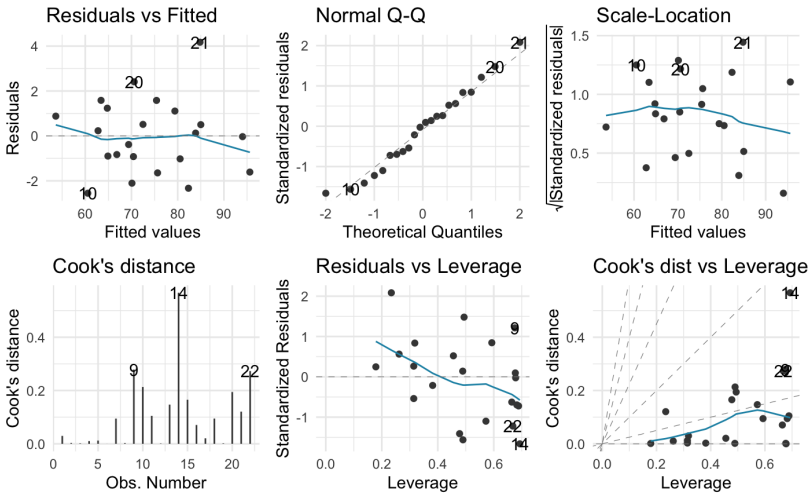


Figure 9.7: Diagnostics for the multiple linear regression example

values $(\hat{Y}_i)_i$. This graph should be done almost systematically. For an illustration, see the graph at the top left of Figure 9.6 and Figure 9.7.

This amounts to plotting the coordinates of vector $P_{[X]^\perp}Y$ as a function of those of vector $P_{[X]}Y$. According to the Cochran theorem, if we meet the four assumptions $(A1-4)$, these two vectors are independent since they are centered and Gaussian. Therefore, we seek to visually validate the independence and the Gaussianity of the two vectors. However, from the graph alone, we can only see the possible deficiency of the assumptions (A1) and (A2). Practically speaking, if we see nothing notable on the graph, that is, if we observe a cloud of points centered and aligned in any way, this is a very good sign: The residuals do not seem to have any interesting property, and this is what we are asking for the error.

Roughly speaking two main pathological patterns can be detected.

- The first one is *"banana shape"* as in Figure 9.8.
  In this case, we can think that the model does not fit the data. Indeed, there does not seem to be any independence between the $\hat{\varepsilon}_i$ and the $\hat{Y}_i$. Therefore, it is necessary to improve the analysis of the problem to propose other relevant regressors or to transform the regressors $x^{(j)}$ by a function of type (log, sin).
- The other typical pathological pattern is the *"trumpet shape"* as in Figure 9.9.
  In this example, there is strong evidence that the variance is not homogenous. One possibility is to set up a change of variable for Y to "make" the variance of the noise constant (see next paragraph).

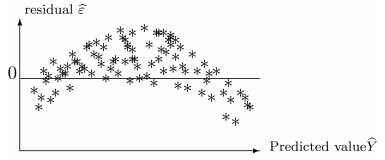


Figure 9.8: *Residuals vs Fitted values*: Banana shape

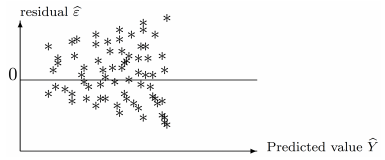


Figure 9.9: *Residuals vs Fitted values*: Trumpet shape

Can we transform the model ?

- We can freely transform the regressors using every possible algebraic transformation: power, square root, exponential circular functions, logarithmic functions,*etc.*, as soon as the resulting regression formula remains interpretable. This technique is adapted to residual plots of the first kind ("banana") and can improve the adequacy of the model or reduce its number of terms if we then use a model selection procedure.

Relationship	Domain for Y	Transformation
$\sigma = (cste)Y^k, k \neq 1$	$\mathbb{R}_+^*$	$Y \mapsto Y^{1-k}$
$\sigma = (cste)\sqrt{Y}$	$\mathbb{R}_+^*$	$Y \mapsto \sqrt{Y}$
$\sigma = (cste)Y$	$\mathbb{R}_+^*$	$Y \mapsto \log(Y)$
$\sigma = (cste)Y^2$	$\mathbb{R}_+^*$	$Y \mapsto Y^{-1}$
$\sigma = (cste)\sqrt{Y(1-Y)}$	$[0,1]$	$Y \mapsto \arcsin(\sqrt{Y})$
$\sigma = (cste)\sqrt{1-Y}Y^{-1}$	$[0,1]$	$Y \mapsto \sqrt{1-Y} - \frac{1}{3}(1-Y)^{2/3}$
$\sigma = (cste)(1-Y)^{-2}$	$[-1,1]$	$Y \mapsto \log(1+Y) - \log(1-Y)$

Table 9.1: Change of variable for the variable to be explained to destabilize the variance of Y

- On the other hand, we can only consider transforming the response Y if the residual plot shows some evidence of heteroscedasticity. The linear model assumes that the absolute error is constant, *i.e.* independent of the amplitude of the response. In many cases, the error is proportional to the response: the larger the response, the larger the error. In such a case, a logarithmic transform of the response will fix the problem. A list of the transformations to be used is given in Table 9.1, depending on the relation between the mean response and the standard error. A more rigorous but much more complex alternative is to use a generalized linear model with an all-chosen link function; see [20] for example. We will study some elementary generalized models in the third part of this course.
  Note that these transformations are based on Taylor expansion and are valid for rather large data. In the other cases, generalized linear models are necessary.

9.7.3 (A3) Independence

A relevant graph to ensure the independence of the errors between them is the scatterplot of the residuals $\hat{\varepsilon}_i$ as a function of the order of the data (when the latter makes sense, especially if it represents time). An example is given in Figure 9.10. Such a graph is potentially suspicious if the residuals remain in packets on one side or the other of 0. One can confirm these doubts by performing a runs test (cf [18], p. 157). This test is based on the number of runs, *i.e.*, the number of consecutive residue packets of the same sign.

On the other hand, if the errors are correlated under certain conditions, a classical approach is to use an ARMA model. The resulting model of regression with ARMA errors is called ARMAX [19, 21–23]. Last, there are also correction methods such as generalized or pseudo-generalized least squares estimates, see [19] or others.

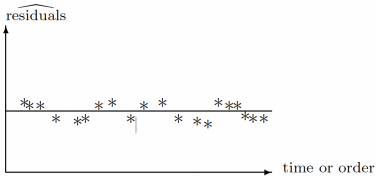


Figure 9.10: Residuals vs Time or order

9.7.4 (H4) Gaussianity

For the Fisher and Student tests to be significant, checking whether the Gaussianity hypothesis is acceptable may be interesting. For this, we strongly advise against the classical non-parametric tests of Kolmogorov-Smirnov and Shapiro-Wilk seen in Section 3.4 because they would be applied on residuals that are (almost) never independent. We prefer to

"settle" for a graphical verification based on a QQ-plot (see the graphs in the upper middle of Figure 9.8 and Figure 9.9).

For the record, this graph connects the points of $\mathbb{R}^2$ formed by the empirical quantiles of the studentized residuals (*i.e.* the $\hat{\varepsilon}_i$ divided by their empirical standard deviation) as a function of the theoretical quantiles (for probabilities $k/n+1$ where $k \in [\![1, n]\!]$, n being the number of data) of a centered reduced normal distribution. Since Student's law strongly looks like a Gaussian distribution as soon as the parameter exceeds ten, if the errors (ε_i) are Gaussian, *i.e.* under (H4), then this QQ-plot is a bisector of the plane.

This type of graph mainly allows us to see if a "heavy tail" distribution would not be more appropriate (in this case, the points move away from Henri's line at its extremities).

9.7.5 Outlier Detection

Finally, we will describe two methods to detect "outlier" data.

9.7.5.1 Hat Matrix and Leverage

Let the hat matrix $H = X({}^tXX)^{-1}\,{}^tX$. Then,

$$\hat{Y}_i = (X\hat{\theta})_i = (HY)_i = H_{ii}Y_i + \sum_{j \neq i} H_{ij}Y_j$$

gives the prediction for the i-th individual. In particular, if $H_{ii} = 1$, then $\hat{Y}_i$ is entirely determined by the i-th observation. On the contrary, if $H_{ii} = 0$, the i-th observation does not influence $\hat{Y}_i$.

Thus, to measure the influence of an observation on its own estimate, one can examine the bar chart of the diagonal terms of H (see Figure 9.11). In other words, the hat matrix H provides a measure of leverage. In practice, one declares the i-th observation to be *leveraged* if H_{ii} exceeds $2k/n$ or $3k/n$.

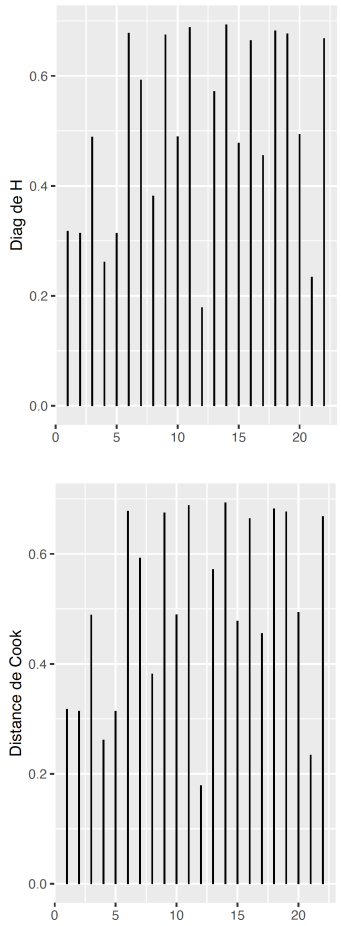


Figure 9.11: *Outlier detection*: Bar graph of the diagonal terms of the hat matrix H (top) and Cook's distances (bottom). Note the similarity between these two graphs.

9.7.5.2 Cook's Distances

The influential points are those points that, if removed from the study, will significantly alter the estimate of the model coefficients. The most classical measure of influence is the Cook's distance. It is a distance between the coefficient estimated with all observations and the one estimated by removing one observation. The Cook's distance for the i-th observation is defined by

$$d_C^i = {}^t(\hat{\theta} - \hat{\theta}^{(-i)})\,{}^tTT(\hat{\theta} - \hat{\theta}^{(-i)})\,,$$

where T is the vector of studentized residuals, and $\hat{\theta}^{(-i)}$ is the estimator of the maximum likelihood without observation i. Here again, the bar graph of the d_C^i can be drawn (see Figure 9.11). A point will be considered

influential if its distance is larger than the others. We must then try to understand why it is influential: Is it a lever, an outlier, or both, . . . ?

High-Dimensional Regression 10

10.1 Curse of Dimensionality

The sustained development of technology, data storage, and computing resources results in the production, storage, and processing of an exponentially growing volume of data. Data is ubiquitous and hugely impacts almost every branch of human endeavor, including science, medicine, business, finance, and government. For example, large-scale data allows us to understand the regulatory mechanisms of living organisms better, create new therapies, monitor climate and biodiversity changes, optimize resources in the health sector, optimize resources in industry and government, and customize the marketing for each consumer, *etc.*

A major characteristic of modern data is that it often simultaneously records thousands, or even millions, of *features* on each *object* or *individual*. Such data are said to be *high-dimensional*.

Simultaneously detecting thousands of variables on each "individual" seems good news: Potentially, we could analyze all the variables likely to influence the studied phenomenon. Unfortunately, statistical reality clashes with this optimistic statement: Separating the signal from the noise is usually almost impossible in high-dimensional data. This phenomenon is often called the "*curse of dimensionality*".

10.1.1 High-Dimensional Geometry

The impact of high dimensionality on statistics is multiple:

1. High-dimensional spaces are broad, and data points are isolated in their vastness;
2. The accumulation of small fluctuations in many different directions can produce a huge overall fluctuation;
3. An event that is an accumulation of rare events may not be rare;
4. Finally, numerical computations and optimizations in high-dimensional spaces can be excessively intensive.

In particular, as the dimension increases, the notion of "nearest points" vanishes. To illustrate this phenomenon, we plot in Figure 10.1 the histograms and boxplots of the distribution of the pairwise-distances $\{\|x^{(i)} - x^{(j)}\|_2 : 1 \leqslant i < j \leqslant n\}$ for $n = 100$ and dimensions $d = 2$, 10, 100, 1000. When the dimension increases, we observe that

- the minimal distance between two points increases,
- all the points are at a similar distance from the others, so the notion of "nearest points" vanishes.

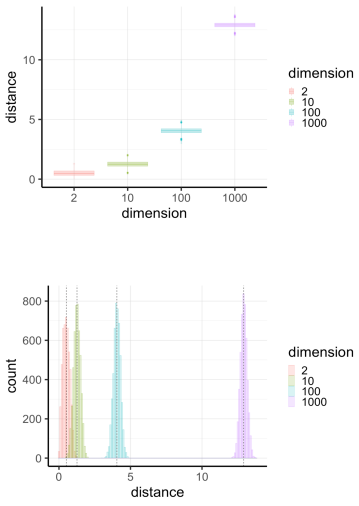


Figure 10.1: *Curse of dimensionality*: Pairwise-distances between $n = 100$ points sampled uniformly in the hypercube $[0, 1]^d$, for $d = 2$, 10, 100, and 1000.

In particular, any estimator based on a local averaging will fail with such data.

More generally, our intuition about space is based on two and three dimensions and can often be misleading in high dimensions. To illustrate this phenomenon, we will look at the hypersphere for the norm 2 in any dimension

The volume $V_d(r)$ of a d-dimensional ball of radius $r \in \mathbb{R}^+$ is equal to

$$V_d(r) = \frac{\sqrt{\pi}^d}{\Gamma\left(\frac{d}{2}+1\right)} r^d \underset{d\to+\infty}{\sim} \left(\frac{2\pi\,\mathrm{e}\,r^2}{d}\right)^{d/2} \frac{1}{\sqrt{d\pi}},$$

where Γ represents the Gamma function $\Gamma(x) = \int_0^{+\infty} t^{x-1}\,\mathrm{e}^{-t}\,\mathrm{d}t$ for $x \in \mathbb{R}^+$.

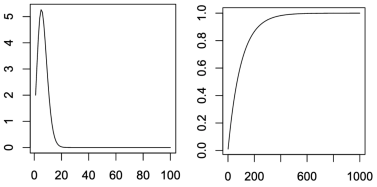


Figure 10.2: *Curse of dimensionality*: Volume (left) and fraction in the crust (right) of a unit sphere according to the dimension of the ambient space.
We observe that for $d = 20$, the volume of the unit ball is already almost 0.

As a consequence,

$$\forall r \in \mathbb{R}^+, \quad \lim_{d\to+\infty} V_d(r) = 0\,.$$

In words, the volume of the d-dimensional sphere with radius goes (very quickly) to 0 as the dimension d increases to infinity, see Figure 10.2 (left). That means a (unit) sphere in high dimensions has almost no volume (compare this to the volume of the unit cube, which is always 1).

Let us consider the volume of the "crust" $C_d(r)$ obtained by removing from the d-dimensional ball with radius r the sub-ball of radius $0.99r$. Hence,

$$\forall r \in \mathbb{R}^+, \quad \frac{C_d(r)}{V_d(r)} = 1 - .99^d\,,$$

which goes exponentially fast to 1. That is, "most" of the volume of the d-dimensional sphere is contained in its "crust". More visually: almost nothing will be left while peeling a high-dimensional orange since almost all of its mass is in its peel. We plot in Figure 10.2 (right) the proportion of points of the ball located in the crust.

The moral of this example is that we have to be careful with our geometric intuitions in high-dimensional spaces: These spaces have some counterintuitive geometric properties.

10.2 Regularized Linear Regression

When we end up with a singular model, $rk(X) < k$, the Gram matrix ${}^tXX$ is no longer invertible and even ill-conditioned (most of the eigenvalues are 0). This case arises when

- our model has more explanatory variables than observations: $p > n$;
- $n > p$ but some variables are linearly redundant, *i.e.* the family $\{X^{(1)}, \ldots, X^{(p)}\}$ is linearly dependent.

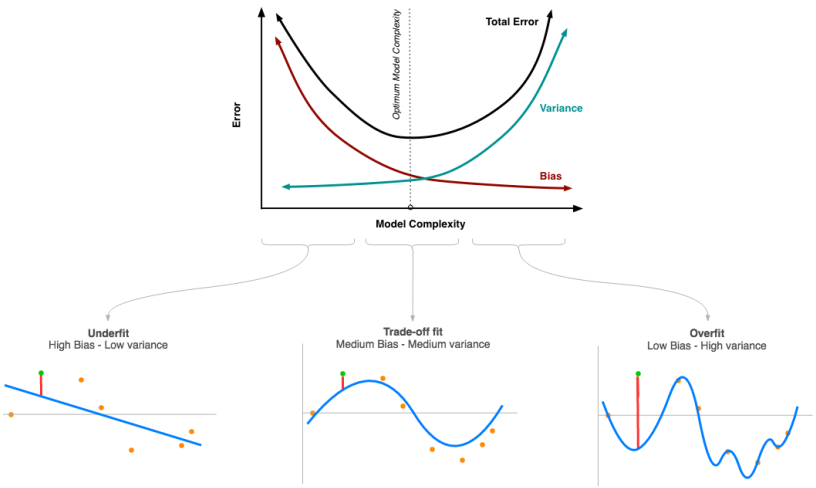


Figure 10.3: *Bias-variance trade-off & Occam's razor*: A high bias or underfitting means that the model cannot capture the trend or pattern in the data. It is usually caused when the hypothesis function is too simple or lacks features. On the contrary, high variance or overfitting means that the model fits the available data but does not generalize well to predict new data. This typically occurs when the hypothesis function is too complex and tries to perfectly fit every data point on the training data set, resulting in unnecessary curves and angles unrelated to data. Increasing the bias can decrease the variance, whereas increasing the variance can decrease the bias. How can we reach the perfect or optimal point for a good model? **Credit**(Picture): Eduard Bonada

In this situation, we have seen previously that the least-squares estimator $\hat{\theta}$ does not exist. The projection $\hat{Y} = P_{[X]}Y$ of the response Y onto $\mathcal{I}m(X) = [X]$ does not have a unique decomposition on the columns of X (the model is unidentifiable). Moreover, since the variance-covariance matrix of $\hat{\theta}$ is $\sigma^2({}^tXX)^{-1}$, the precision of the $\hat{\theta}$ estimator decreases when ${}^tXX$ approaches a non-invertible matrix. In other words, the standard linear model completely fails, and a new high-dimensional specific linear regression framework must be developed.

Example 10.1 DNA microarrays measure the transcription level of tens of thousands of genes simultaneously. We are typically in a situation where p (the number of genes) will be significantly greater than the number of samples n (the number of genomes studied).

10.2.1 Important Balance : Bias-Variance Trade-Off

From a prediction perspective, if x^* is a new vector of values of the explanatory variables, we know that the quality (in the sense of squared deviations) of the prediction $\hat{Y}^*$ of the true response Y^* is decomposed into the squared bias + the variance. Or said more precisely, for any estimator $\hat{\theta}$ of θ,

$$\mathbb{E}\left[(Y - X\hat{\theta})^2\right] = \mathcal{B}ias\left[X\hat{\theta}\right] + \mathcal{V}ar\left[X\hat{\theta}\right] + \sigma^2,$$

where

$$\mathcal{B}ias\left[X\hat{\theta}\right] = \mathbb{E}\left[X\hat{\theta}\right] - X\theta \quad \& \quad \mathcal{V}ar\left[X\hat{\theta}\right] = \mathbb{E}\left[\left(X\hat{\theta} - \mathbb{E}[X\hat{\theta}]\right)^2\right].$$

Thus, to improve the prediction, one may prefer a slight increase in the bias to induce a decrease in the variance. Figure 10.3 illustrates this need to make a trade-off between bias and variance. Figure 10.4 gives an intuition on the influence of bias and variance on the quality of the estimate.

In this context, we will try to use so-called *regularized*, or penalized, regression methods to overcome these difficulties. Their common for-

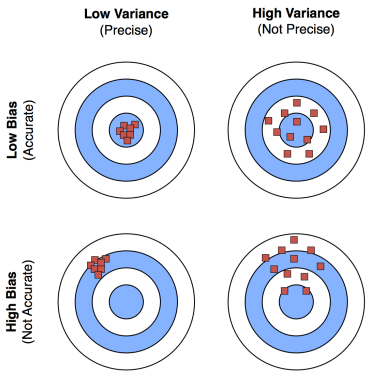


Figure 10.4: *Graphical representation of the quality of an estimator*: predicted values represented on a target centered on the value to be estimated for different bias/-variance situations.
Credit(Picture): Sebastian Raschka

malism is the optimization of a criterion of the form

$$\underset{\theta\in\mathbb{R}^k}{\operatorname{argmin}} \|Y - X\theta\|^2 + \lambda \operatorname{pen}(\theta)\,,$$

where $\lambda \in \mathbb{R}^+$ is a tuning parameter. They differ in the form of the penalty function $\operatorname{pen}(\theta)$, which will involve monitoring a norm of θ.

In practice, we start by centering and reducing the explanatory variables $x^{(j)}$ not to penalize or favor a θ coefficient. Indeed, as mentioned before, the penalties that we will consider rely on using a judicious norm of θ. Therefore, we want to affect each coefficient in a "similar" way. We denote $\tilde{X}$ the matrix of centered-reduced explanatory variables. Moreover, since the intercept θ_0 has a particular role in positioning the model around the mean behavior of Y, it does not have to be involved in the constraint on the norm of θ. Hence, we center the response vector Y, $\tilde{Y} = Y - \bar{Y}\mathbb{1}_n$, and we can potentially reduce it. Note that the model is then of the form

$$\tilde{Y} = \tilde{X}\theta\,, \quad \text{where} \quad \theta = {}^t(\theta_1, \ldots, \theta_p)\,,$$

i.e. $k = p$, and without intercept.

Therefore, after the initial data transformation, we focus on regularized regression methods that seek to minimize the regularized empirical risk (for squared loss):

$$\underset{\theta\in\mathbb{R}^k}{\operatorname{argmin}} \|\tilde{Y} - \tilde{X}\theta\|^2 + \lambda \|\theta\|_q^q\,, \quad \text{where} \quad \|\theta\|_q^q = \sum_{j=1}^{p} (\theta_j)^q\,,$$

We speak of ridge regression when $q = 2$ and Lasso regression when $q = 1$. We will detail these two methods and the Elasticnet regression that combines the first two. To illustrate this section, we use the dataset introduced in the previous chapter, `mensurations.txt`, to which we have added 10 noise variables simulated according to a $\mathcal{N}(0,1)$ distribution. The resulting dataset, available on the moodle page of the course, is called `mensurations_extended.txt`.

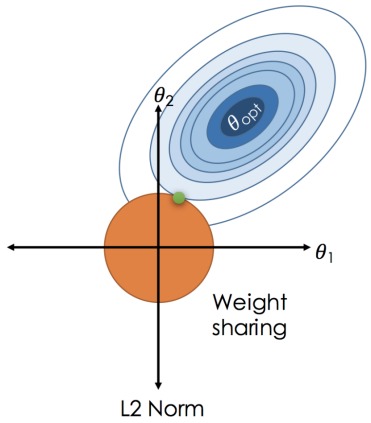


Figure 10.5: *Geometrical interpretation*: Contours of the error and constraint function for the ridge regression. The solid orange areas are the constraint regressions, while the blue ellipses are the contours of the residual sum of squares (RSS). The Ridge estimates can be viewed as the point where the linear regression coefficient contours intersect the circle defined by $\theta_1^2 + \theta_2^2 \leqslant r(\lambda)$.

10.2.2 Ridge Regression

In the context presented above, the difficulty comes from the non-invertibility of ${}^tXX$. This matrix being positive semi-definite, all its eigenvalues are positive. If ${}^tXX$ is not invertible, then at least one of its eigenvalues is zero. Let $\lambda_1 > \lambda_2 > \ldots > \lambda_p$ be its ordered eigenvalues.

Proposition 10.1 *Let $\lambda \in \mathbb{R}^+$. The matrices ${}^t\tilde{X}\tilde{X}$ and ${}^t\tilde{X}\tilde{X} + \lambda I_p$ have the same eigenvectors, but their eigenvalues are $\{\lambda_j\}_{j\in[\![1,p]\!]}$ and $\{\lambda_j + \lambda\}_{j\in[\![1,p]\!]}$ respectively.*

In particular, if $\lambda > 0$, $\mathcal{D}et({}^t\tilde{X}\tilde{X} + \lambda I_p) > \mathcal{D}et({}^t\tilde{X}\tilde{X})$ and ${}^t\tilde{X}\tilde{X} + \lambda I_p$ has "more chance" of being invertible than ${}^t\tilde{X}\tilde{X}$. Therefore, the idea is to replace $\left({}^t\tilde{X}\tilde{X}\right)^{-1}$ in the expression for the least-squares estimator $\hat{\theta}$ with $\left({}^t\tilde{X}\tilde{X} + \lambda I_p\right)^{-1}$. Hence, the ridge estimator is given by

$$\hat{\theta}_{\text{ridge}} = \left({}^t\tilde{X}\tilde{X} + \lambda I_p\right)^{-1} {}^t\tilde{X}\tilde{Y}\,.$$

In particular, this ridge estimator is a solution to the optimization problem

$$\hat{\theta}_{\text{ridge}} \in \underset{\theta\in\mathbb{R}^p}{\operatorname{argmin}} \left\|\tilde{Y} - \tilde{X}\theta\right\|_2^2 + \lambda\|\theta\|_2^2\,,$$

or also a solution to the constrained minimization problem :

$$\hat{\theta}_{\text{ridge}} \in \underset{\theta\in\mathbb{R}^p}{\operatorname{argmin}} \left\|\tilde{Y} - \tilde{X}\theta\right\|_2^2 \quad \text{under the constraint} \quad \|\theta\|_2^2 \leqslant r(\lambda)\,,$$

where r is a bijective function. The ridge regression keeps all the variables, but the constraint $\|\theta\|_2^2 \leqslant r(\lambda)$ prevents the estimators from taking too large values and thus limits the variance of the predictions. The estimator is said to *shrink* the estimated coefficients towards zero: the range of possible estimated parameter values is reduced.

Proposition 10.2 *The ridge estimator* $\hat{\theta}_{ridge}(\lambda) = \left({}^t\tilde{X}\tilde{X} + \lambda I_p\right)^{-1} {}^t\tilde{X}\tilde{Y}$

1. *is biased*
$$\mathbb{E}\left[\hat{\theta}_{ridge}\right] = \theta - \lambda\left({}^t\tilde{X}\tilde{X} + \lambda I_p\right)^{-1}\theta\,;$$
2. *has a smaller variance than the vanilla estimator* $\hat{\theta}$*:*
$$\begin{aligned}\mathcal{V}ar\left(\hat{\theta}_{ridge}\right) &= \sigma^2\left({}^t\tilde{X}\tilde{X} + \lambda I_p\right)^{-1}({}^t\tilde{X}\tilde{X})\left({}^t\tilde{X}\tilde{X} + \lambda I_p\right)^{-1} \\ &\leqslant \sigma^2\left({}^t\tilde{X}\tilde{X}\right)^{-1} = \mathcal{V}ar\left(\hat{\theta}\right).\end{aligned}$$

We define the fitted values for Y by

$$\hat{Y}_{\text{ridge}} = \tilde{X}\hat{\theta}_{\text{ridge}}(\lambda) + \bar{Y}\mathbb{1}_n\,.$$

The tuning parameter $\lambda \in \mathbb{R}^+$ controls the strength of the penalty term. Note that:

- When $\lambda = 0$, we get the linear regression estimate $\hat{\theta}$,
- When $\lambda \to +\infty$, then $\hat{\theta}_{\text{ridge}} \to 0$,
- For λ in between, we are balancing two ideas: fitting a linear model of $\tilde{Y}$ on $\tilde{X}$, and shrinking the coefficients.

The quality of the estimator $\hat{\theta}_{\text{ridge}}$ depends on the choice of λ, which, according to Proposition 10.2, behaves as follows:

- The bias increases as the amount of shrinkage λ increases,
- The variance decreases as the amount of shrinkage λ increases.

Therefore, we need to find the λ that makes the best compromise between bias and variance. This choice is a delicate point; worse, it is almost impossible to make this choice *a priori*. If we plot the *regularization path* of the ridge regression (Figure 10.6), we see that it is continuous, which does not allow an easy adjustment of λ. A first solution is to

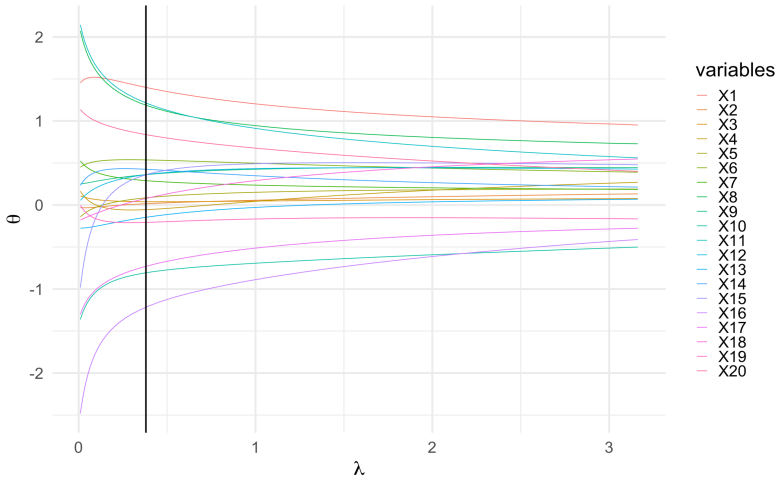


Figure 10.6: *Regularization paths for the ridge regression*: Functions $\lambda \mapsto \left(\hat{\theta}_{\text{ridge}}(\lambda)\right)_j$ for all variable $j \in [\![1, 20]\!]$

follow the recommendations proposed in the literature, see [8, 14–16] for example.

Practical Choice of the Regularization Parameter

In practice, there are two ways to approach this question. A more traditional approach consists of choosing λ so that a specific information criterion, AIC or BIC most of the time, is the smallest. Warning: The number of degrees of freedom in the ridge regression differs from that of the ordinary least squares approach!

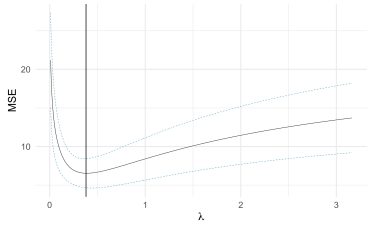


Figure 10.7: *λ selection by cross-validation for ridge regression*: Mean square error as a function of λ. By minimizing this function, we obtain the optimal value of λ, in this case $\lambda^* = 0.38$.

A more machine learning-like approach is to perform *cross-validation* and select the value of λ that minimizes the cross-validated sum of squared residual or some other measure (Figure 10.7). We first partition the data into a training set $(\tilde{Y}^{\text{train}}, \tilde{X}^{\text{train}})$ and a test set $(\tilde{Y}^{\text{test}}, \tilde{X}^{\text{test}})$. We then estimate the ridge regression on the training set for each value of λ in a chosen grid and predict the response on the test set. We then measure the quality of the induced model by comparing the predicted values $\hat{Y}^{\text{test}}_{\text{ridge}}(\lambda) = \tilde{X}^{\text{test}}\hat{\theta}_{\text{ridge}}(\lambda)$ and the real data $\tilde{Y}^{\text{test}}$. We can, for example, use the predicted residual sum of squares (PRESS) criterion

$$PRESS(\lambda) = \|\tilde{Y}^{\text{test}} - \hat{Y}^{\text{test}}_{\text{ridge}}(\lambda)\|^2 .$$

Finally, we choose the value λ so that it minimizes this criterion. In our case, we find for example that $\lambda^* = 0.38$.

The principle of cross-validation is to repeat this split between test and training several times and consider the average of each obtained criterion.

Limit of the Ridge Regression

One may question the validity and usefulness of ridge regression when none of the true coefficients are small, that is, when all the true coefficients are medium or large. Perhaps surprisingly, ridge regression is still valid.

Nevertheless, its advantage is less spectacular here. Moreover, the corresponding range of good lambda values is narrower.

On the contrary, suppose there is a subgroup of real coefficients identically zero. In other words, assume that the average result does not depend on a subset of predictors. We would like to detect these extraneous predictors and remove them (at least virtually) from our predictor set. More generally, we refer to variable selection when our goal is to select relevant variables from a broader set. In addition to predictive accuracy, this can be very important for model interpretation.

So how does ridge regression behave if a group of real coefficients is exactly zero? The answer depends on whether one is interested in prediction or interpretation. Ridge regression will reduce the components of its estimate to zero but will never set those components to zero (unless $\lambda = 0$, but in that case, all features are zero). Thus, the answer will be none other than the one corresponding to this subgroup of small but non-zero coefficients. In terms of prediction, this does not pose much of a problem. However, ridge regression is not as informative as we would like for interpretive purposes.

Strictly speaking, ridge regression does not perform variable selection.

10.2.3 Sparsity: The Lasso Regression

The idea of the LASSO (Least Absolute Selection and Shrinkage Operator) regression proposed by Tibshirani [17] is to cancel some coefficients of the vector θ to have a sparse estimator. This leads to the selection of variables leading to a more interpretable model and a matrix of explanatory variables with better properties than ${}^tXX$.

Example 10.2 In many applications, $p >> n$ but many extracted features in X are irrelevant. Suppose we want to study the size of a tumor Y. It seems reasonable to assume that it can be expressed as a linear combination of the genetic information of the genome described in X. However, most components of X will be zero; most genes will be irrelevant to predict Y.

To force the cancellation of theta coordinates, we constrain its ℓ_1 norm: $\|\theta\|_1 = \sum_{j=1}^{p} |\theta_j|$. As in ridge regression, the first step is to center-reduce the explanatory variables ($X \to \tilde{X}$) and at least center the response vector ($Y \to \tilde{Y}$). Therefore, we define the LASSO estimator: For all $\lambda \in \mathbb{R}_+^*$,

$$\hat{\theta}_{\text{lasso}} \in \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \left\| \tilde{Y} - \tilde{X}\theta \right\|_2^2 + \lambda \|\theta\|_1 . \tag{10.1}$$

This minimization problem is equivalent to minimize $\|\tilde{Y} - \tilde{X}\theta\|_2^2$ under the constraint $\|\theta\|_1 \leqslant r(\lambda)$, where r is a bijective function. The solution of problem (10.1) may not be unique since the above criterion is not strongly convex. However, the resulting vector of fitted values $\tilde{X}\hat{\theta}_{\text{lasso}}(\lambda)$ is always unique.

We have similar behavior to the ridge regression as a function of the penalty parameter λ:

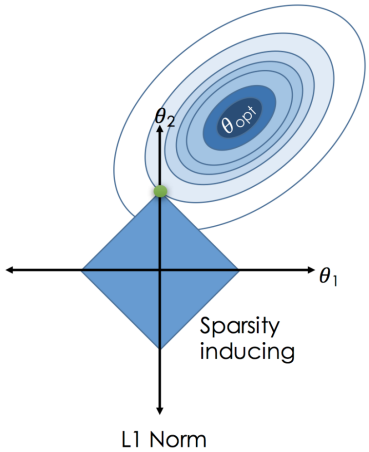


Figure 10.8: *Geometrical interpretation*: Contours of the error and constraint function for the lasso regression. The point where the ellipses intersect the bounding box give us the lasso estimates. Note that the intersection is at a corner, so the coefficient θ_1 in this case is set to zero.

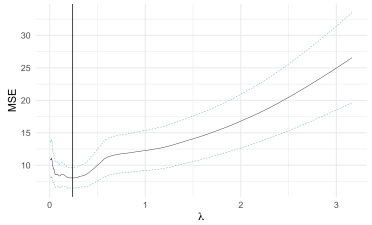


Figure 10.9: *λ selection by cross-validation for lasso regression*: Mean square error as a function of λ. By minimizing this function, we obtain the optimal value of λ, in this case $\lambda^* = 0.24$.

- When $\lambda = 0$, we get the linear regression estimate $\hat{\theta}$,
- A large λ leads to a very sparse solution: $\lim_{\lambda \to +\infty} \hat{\theta}_{\text{lasso}}(\lambda) = 0$,
- The parameter λ must be chosen as judiciously as this choice is tricky and impossible to realize *a priori*.

As with the ridge regression, we can plot the regularization path of the Lasso regression, see Figure 10.10. And, likewise, we go through a *cross-validation* procedure to stabilize the choice of λ, see Figure 10.9.

Ridge vs. Lasso Regression

Neither method is unconditionally better than the other. Lasso tends to do well if there are a small number of significant parameters and the others are close to zero, *i.e.* when only a few predictors influence the response. On the contrary, ridge regression works well if many significant parameters share approximately the same value, *i.e.* when most predictors impact the response. However, we do not know the true parameter values in practice. So, the previous two points are somewhat theoretical. Just run cross-validation to select the more suited model for a specific case or... try to combine the two!

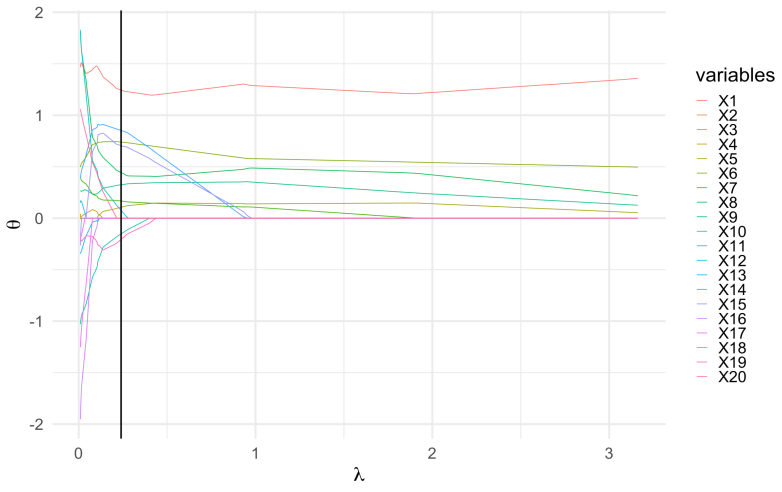


Figure 10.10: *Regularization paths for the lasso regression*: Functions $\lambda \mapsto \left(\hat{\theta}_{\text{lasso}}(\lambda)\right)_j$ for all variable $j \in [\![1, 20]\!]$

10.2.4 Elastic-Net Regression

Elastic Net first emerged as a result of critique on lasso, whose variable selection can be too dependent on data and thus unstable. The solution is to combine the penalties of ridge regression and lasso to get the best of both worlds. The Elastic-Net estimator [24] is defined for $\lambda > 0$ ans $\alpha > 0$ by:

$$\hat{\theta}_{\text{net}} \in \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \left\| \tilde{Y} - \tilde{X}\theta \right\|_2^2 + \lambda \left(\alpha \|\theta\|_1 + (1-\alpha)\|\theta\|_2^2 \right) ,$$

where α is the mixing parameter between ridge ($\alpha = 0$) and lasso ($\alpha = 1$). This minimization problem is equivalent to the minimization of $\left\| \tilde{Y} - \tilde{X}\theta \right\|_2^2$ under the constraint $\alpha\|\theta\|_1 + (1-\alpha)\|\theta\|_2^2 \leqslant r(\lambda)$.

Now, there are two parameters to tune: λ and α. In practice, this calibration is often performed by cross-validation.

Figure Figure 10.13 shows the differences in the regularization paths of the three methods.

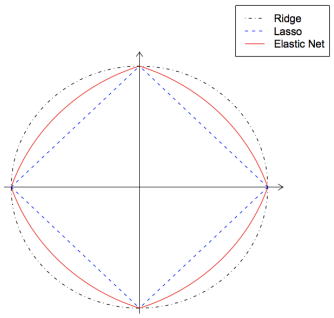


Figure 10.11: *Elastic-Net Regression*. Two-dimensional illustration, $\alpha = 0.5$.

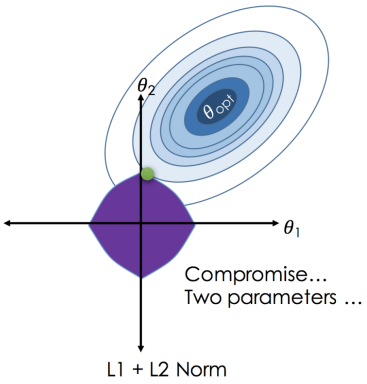


Figure 10.12: *Geometrical interpretation*: Contours of the error and constraint function for the elastic-net regression.

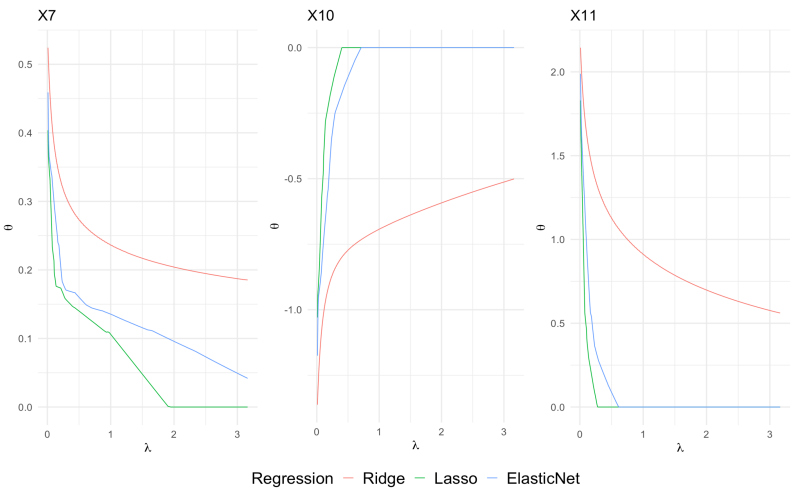


Figure 10.13: *Regularization path* for 3 variables of the dataset for Ridge($\alpha = 0$, red), Lasso ($\alpha = 1$, green) and Elastic-Net regression (here $\alpha = 0.5$, blue).

One-Way Analysis of Variance 11

Analysis of variance (ANOVA) covers a set of tests and estimation techniques intended to assess the effect of one or more qualitative variables on a quantitative variable. In the simplest case, it compares several Gaussian sample means: In other words, we generalize the classical test of equality of two means seen in 3MIC to the test of equality of $I \geqslant 2$ means.

As in the test of equality of two means, normality and independence of the quantitative variable are assumed, as well as equality of variances. The particularity of ANOVA is that the p means are supposed to come from p samples, each corresponding to a modality of a qualitative character used to stratify the population.

In ANOVA, we use a particular vocabulary introduced by agronomists who first addressed this type of problem: the qualitative variable likely to influence the distribution of the observed quantitative response Y is called a *factor*, and its modalities are called *levels*. A factor is said to be *controlled* if its values are not observed but fixed by the experimenter.

he modalities of the corresponding qualitative explanatory variable.

11.1 Experimental Design

An experimental design lists all the combinations of the different factors considered by the experimenter. Here we give some basic definitions of experimental design, which will be helpful for the following. We will not discuss the theory of experimental design in this course.

Definition 11.1 (Design of Experiments)
- *A cell or* treatment combinations *is the the combination of the settings of several factors in a given experimental trial. In other words, it is a cell of the table associated with a combination of the controlled factors;*
- *An experimental design is said to be* complete *if we observe at least one value in each cell;*
- *An experimental design is said to be* repeated *if we observe more than one data per cell;*
- *A* balanced *design is an experimental design where all cells have the same number of observations;*
- *A balanced and repeated design is said to be* equirepeated.

11.2 One-Way Analysis of Variance

In this part, we observe a quantitative variable Y, which we try to explain using a *single* explanatory factor. We note:

- i the index of the level (or "cell") for the explanatory factor,
- I the number of levels: $i \in [\![1, I]\!]$,
- n_i the number of experiments in the level i,
- $j \in [\![1, n_i]\!]$ the index of the experiment in the i-th level,
- $n = \sum_{i=1}^{I} n_i$ the total number of experiments.

An experiment or "individual" is identified by two indices: i, the number of the cell or level, and j which indexes the observation for this level. Thus, we note Y_{ij}, the theoretical value of the quantitative response for the experiment j in the level i.

Table 11.1: Data to illustrate the one-factor anova.

Examiner i	A	B	C
Marks y_{ij}	10	8	10
	11	11	13
	11	11	14
	12	13	14
	13	14	15
	15	15	16
		16	16
		16	
Size n_i	6	8	7
Mean $\bar{y}_{i\cdot}$	12	13	14

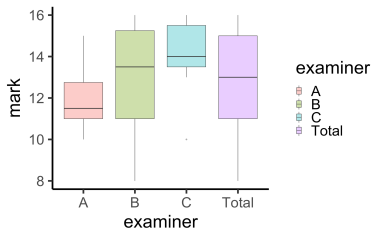


Figure 11.1: Boxplot of marks overall (right) and by examiner.

In this chapter, we will illustrate the concepts discussed with the example presented in Table 11.1 and Figure 11.1. We are interested in the grades obtained by students in an oral exam. Specifically, we ask about a potential effect of the examiner on the obtained scores. Indeed, we observe a difference of 2 points between the best average, which is 14, and the worst, which is 12.

This data can be approached in two ways:

- We dispose of 3 independent samples and we want to compare their averages: this is the "comparison of averages" approach.
- We observe a single sample of length 18 and one factor (the examiner), and we study the effect of this factor on the mean: this is the "analysis of variance" approach.

11.3 One-Way ANOVA Model

We model a quantitative variable as a function of a factor at I levels. For each level $i \in [\![1, I]\!]$ of the factor, we observe n_i repeated measurements of Y, denoted y_{ij}, where $j \in [\![1, n_i]\!]$. We make the following assumptions of normality and independence:

1. For all $i \in [\![1, I]\!]$ and $j \in [\![1, n_i]\!]$, y_{ij} is a realization of a random variable Y_{ij} of law $\mathcal{N}(m_i, \sigma^2)$;
2. The random variables Y_{ij} are globally independent.

These assumptions can be summarized by writing the model:

$$\begin{aligned} \forall i \in [\![1, I]\!]\,,\ \forall j \in [\![1, n_i]\!], \quad Y_{ij} &= m_i + \varepsilon_{ij}\,, \\ \text{where} \quad \varepsilon_{ij} &\overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)\,. \end{aligned} \tag{$\mathcal{M}_I^\star$}$$

In other words, we describe the effect of the factor by assuming:

- a specific expectation m_i for each group or level of the factor,
- and an intra-group variance σ^2 common to all groups.

The purpose of this study will be to know whether, given the data (from Table 11.1, for example), the means of the I samples are equal or different. In other words, we want to know whether the observed empirical average $\bar{y}_{i\cdot}$ differ because of actual differences between the means m_i or whether these differences can reasonably be attributed to sampling fluctuations alone.

Note that we can rewrite this model in matrix form by setting:

$$\begin{pmatrix} Y_{1,1} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{In_I} \end{pmatrix} = \underbrace{\begin{pmatrix} \mathbb{1}_{n_1} & 0_{n_1} & \dots & 0_{n_1} \\ 0_{n_2} & \mathbb{1}_{n_2} & \dots & 0_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{n_I} & 0_{n_I} & \cdots & \mathbb{1}_{n_I} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_I \end{pmatrix}}_{\theta} + \varepsilon\,, \quad \text{where} \quad \varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n)\,.$$

With the R software, we use the command `lm(mark~exam-1)`.
We can also visualize the design matrix X with the command `model.matrix(mark~exam-1)`.

```
> anov_reg = lm(mark~exam-1, data=note)
> summary(anov_reg)

Call:
lm(formula = mark ~ exam - 1, data = note)

Residuals:
   Min     1Q Median     3Q    Max
    -5     -1      0      2      3

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
examA  12.0000     0.9526   12.60 2.30e-10 ***
examB  13.0000     0.8250   15.76 5.63e-12 ***
examC  14.0000     0.8819   15.88 4.98e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.333 on 18 degrees of freedom
Multiple R-squared:  0.9734,    Adjusted R-squared:  0.969
F-statistic: 219.7 on 3 and 18 DF,  p-value: 2.311e-14
```

11.3.1 Decomposition of Effects

For interpretation purposes, we may be interested in a change of parameterization. This is a change of variables in the function to be minimized and whose variables are the model's parameters. Note that the new equations we will define below always correspond to those of a one-factor model.

In particular, to compare the effects of the factor levels, it may be more appropriate to take an average effect as a reference and to examine the deviations of the effects of the different levels from this average effect. Hence, the initial model ($\mathcal{M}_I^\star$) writes

$$Y_{ij} = \underbrace{\mu + \alpha_i}_{m_i} + \varepsilon_{ij}\,, \quad \text{where} \quad \varepsilon_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)\,. \tag{$\mathcal{M}_I$}$$

With this writing, μ denotes the *average* effect and $\alpha_i = m_i - \mu$ the *differential* (centered) effect of level i.

While this model is more easily interpretable, it is also over-parameterized (see Chapter 8). Therefore, we must constrain its parameters to make it regular. Generally, we consider the model ($\mathcal{M}_I$) under the so-called "natural" constraint $\sum_{i=1}^I n_i\alpha_i = 0$. This constraint has the good taste to make the model orthogonal. Another commonly used constraint is to impose $\alpha_1 = 0$. R uses it by default while executing the command `lm(mark~exam)`.

```
> anov_sing = lm(mark~exam, data=note)
> summary(anov_sing)

Call:
lm(formula = mark ~ exam, data = note)

Residuals:
   Min     1Q Median     3Q    Max
    -5     -1      0      2      3

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.0000     0.9526  12.597  2.3e-10 ***
examB         1.0000     1.2601   0.794    0.438
examC         2.0000     1.2981   1.541    0.141
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.333 on 18 degrees of freedom
Multiple R-squared:  0.1167,    Adjusted R-squared:  0.0186
F-statistic:  1.19 on 2 and 18 DF,  p-value: 0.3272
```

Exercise 11.1 *What is the relationship between the parameters of the three models introduced above, namely*

- *the regular model* ($\mathcal{M}_I^\star$),
- *the singular model* ($\mathcal{M}_I$) *under the natural constraint* $\displaystyle\sum_{i=1}^I n_i\alpha_i = 0$,
- *the singular model* ($\mathcal{M}_I$) *under the constraint* $\alpha_1 = 0$?

Definition 11.2 *In the context of ANOVA, the dimension of the space in which the expectation of the random variables Y_{ij} lives is called the* dimension *of the model. This dimension is equal to the number of expectation parameters considered in the modeling minus the number of identifiability constraints necessary (independent) to estimate the said parameters.*

The one-way ANOVA model is of dimension I, hence the notation $(\mathcal{M}_I^\star)$. Indeed, we have:

- Either I parameters, the m_i, and no constraints in the regular model $(\mathcal{M}_I^\star)$;
- Or $I+1$ parameters, μ and the α_i, and a constrain $\sum_{i=1}^I n i_i \alpha_i = 0$, in the singular model $(\mathcal{M}_I)$.

11.3.2 Model Without Treatment Effect

We want to know if the factor really influences the variable of interest Y. To test the absence of effect of the factor, we will test the null hypothesis

$$\mathcal{H}_0 : \text{“} m_1 = m_2 = \ldots = m_I''$$

against the alternative

$$\mathcal{H}_1 : \text{“} \exists (i,j) \quad \text{such that} \quad m_i \neq m_j'' \,.$$

The equality “$m_1 = m2 = \ldots = m_I$” allows us to define a sub-model of the complete one-way ANOVA model. By noting m this common mean, this sub-model writes

$$Y_{ij} = m + \varepsilon_{ij} \,, \quad \text{where} \quad \varepsilon_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2) \,. \qquad (\mathcal{M}_1^\star)$$

This sub-model is of dimension 1: only one parameter and no constraints.

11.4 Estimation and Forecasting

We can now focus on estimating the parameters of the different models. In the case of the regular model $(\mathcal{M}_I^\star)$, these estimates directly follow from the results we obtained in Chapter 6. For the singular parameterizations $(\mathcal{M}_I)$, we have to be more careful.

Thereafter, to refer to the complete model without taking into account the parametrization, we will note $(\mathcal{M}_I^\sqcup)$.

Let $\bar{Y}_{i\cdot}$ and $\bar{Y}_{\cdot\cdot}$ be the averages defined by:

- $\bar{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$ the mean of level i, and
- $\bar{Y}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij}$ the overall average.

11.4.1 Estimation in the Complete Model ($\mathcal{M}_I^{\sqcup}$)

Proposition 11.2 (Least squares estimation) *In the model ($\mathcal{M}_I^{\star}$), the m_i are estimated by*

$$\forall i \in [\![1, I]\!], \quad \hat{m}_i = \bar{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}\,.$$

They are called main effects *of the factors. They are normally distributed and their variance is σ^2/n_i.*

Proposition 11.3 (Least squares estimation) *In the model ($\mathcal{M}_I$),*

1. *Under the "natural" constraint $\sum_{i=1}^{I} n_i\alpha_i = 0$, μ and the α_i are estimated by*

$$\hat{\mu} = \bar{Y}_{\cdot\cdot} \qquad \text{and} \qquad \forall i \in [\![1, I]\!], \quad \hat{\alpha}_i = \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\,.$$

2. *Under the constraint $\alpha_1 = 0$, μ and the α_i are estimated by*

$$\hat{\mu} = \bar{Y}_{1\cdot} \qquad \text{and} \qquad \forall i \in [\![\mathbf{2}, I]\!], \quad \hat{\alpha}_i = \bar{Y}_{i\cdot} - \bar{Y}_{1\cdot}\,.$$

The estimation of the variance does not depend on the parameterization, nor does the definition of the fitted values and the residuals.

Proposition 11.4 (Variance) *The estimator of the variance σ^2 is given by*

$$\hat{\sigma}^2 = \frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2\,.$$

For all $i \in [\![1, I]\!]$ and $j \in [\![1, n_i]\!]$, we predict Y_{ij} by:

- In ($\mathcal{M}_I^{\star}$): $\hat{Y}_{ij} = \hat{m}_i = \bar{Y}_{i\cdot}$;
- In ($\mathcal{M}_I$): $\hat{Y}_{ij} = \hat{\mu} + \hat{\alpha}_i = \bar{Y}_{\cdot\cdot} + \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot} = \bar{Y}_{i\cdot}$.

For each parameterization, we deduce the residuals

$$\hat{\varepsilon}_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_{i\cdot}\,.$$

Exercise 11.5 *Proof of Proposition 11.2. The model $(\mathcal{M}_I^\star)$ being regular, we can use the general formula $\hat{\theta} = ({}^tXX)^{-1}\,{}^tXY$ or minimize the least square function $h\colon (m_1, \dots, m_I) \mapsto \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - m_i\right)^2$.*

Exercise 11.6 *Proof of Proposition 11.3. Consider the parametrization $(\mathcal{M}_I)$.*

1. *Under the "natural" constraint, the formula seen in Chapter 6 is not valid anymore. We must therefore minimize the least squares function.*

$$h\colon (\mu, \alpha_1, \dots, \alpha_I) \mapsto \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - \mu - \alpha_i\right)^2 ,$$

   under the constraint $\sum_{i=1}^{I} n_i\alpha_i = 0$.
2. *Under the constraint $\alpha_1 = 0$, we precede as before but adapting the constraint.*

11.4.2 Estimation in the Sub-Model $(\mathcal{M}_1^\star)$

In this model, we need to estimate m and σ^2:

1. We estimate m by $\hat{m} = \dfrac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij} = \bar{Y}_{\cdot\cdot}$;
2. For all $i \in [\![1, I]\!]$ and $j \in [\![1, n_i]\!]$, we predict Y_{ij} by $\hat{Y}_{ij} = \hat{m} = \bar{Y}_{\cdot\cdot}$;
3. The residuals are then given by $\hat{\varepsilon}_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_{\cdot\cdot}$.

11.4.3 Properties

Go back to the complete model $(\mathcal{M}_I^\sqcup)$. We have the following properties, analogous to those of linear regression, which does not depend on the chosen parameterization.

Proposition 11.7 1. *The average of the residuals per level is zero:*

$$\forall i \in [\![1, I]\!], \quad \frac{1}{n_i} \sum_{j=1}^{n_i} \hat{\varepsilon}_{ij} = 0 .$$

2. *The overall mean of the residuals is zero:*

$$\frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} \hat{\varepsilon}_{ij} = 0 .$$

3. *The average of the adjusted values is equal to the average of the observed values:*

$$\frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} \hat{Y}_{ij} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} Y_{ij} .$$

4. *Residuals and fitted values are not correlated:* $\widehat{Cov}(\hat{\varepsilon}, \hat{Y}) = 0$.

5. *We can decompose the variance into* $\widehat{Var}(Y) = \widehat{Var}(\hat{Y}) + \widehat{Var}(\hat{\varepsilon})$.

Exercise 11.8 *Prove Proposition 11.7. You can use the proof of Proposition 9.2.*

The last property leads us to define the notion of inter- and intra-group variance.

Definition 11.3 (Variance decomposition)

1. *We call* inter-group variance *the variance of the means by level, weighted by the weights of these levels, i.e.*

$$\widehat{Var}(\hat{Y}) = \frac{1}{n} \sum_{i=1}^{I} n_i \left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot} \right)^2 .$$

2. *We call* intra-group variance, *or residual variance, the average of the empirical variances of the observations in the levels, i.e.*

$$\widehat{Var}(\hat{\varepsilon}) = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i\cdot} \right)^2 = \frac{1}{n} \sum_{i=1}^{I} n_i \widehat{Var}_i(Y) ,$$

where $\widehat{\mathcal{V}ar}_i(Y)$ is the empirical variance in the level i:

$$\widehat{\mathcal{V}ar}_i(Y) = \frac{1}{n_i}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2 .$$

Thus the relation [Proposition 11.7.5] "$\widehat{\mathcal{V}ar}(Y) = \widehat{\mathcal{V}ar}(\hat{Y}) + \widehat{\mathcal{V}ar}(\hat{\varepsilon})$" writes in this context[1]

" Total variance = Inter variance + Intra variance " .

The SSR and SSE quantities provide a good definition of what is meant by inter and intra group variance.

- $SSR = \sum_{i=1}^{I} n_i \left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2$ measures the deviation of group means from the overall mean: it is a measure of variability between groups.
- On the other hand, $SSE = \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2$ measures the deviation of each individual from the mean of the group to which he belongs: it is a measure of variability within each group.

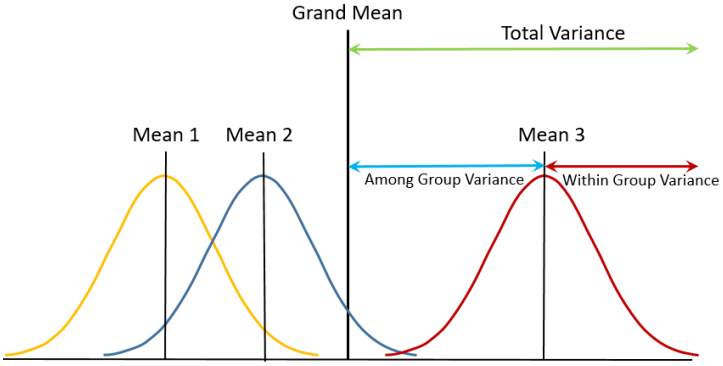


Figure 11.2: Decomposition of the variance in the ANOVA context

Finally, we define the coefficient R^2 as the ratio of the inter-group variance to the total variance:[2]

$$R^2 = \frac{\widehat{\mathcal{V}ar}(\hat{Y})}{\widehat{\mathcal{V}ar}(Y)} = 1 - \frac{\widehat{\mathcal{V}ar}(\hat{\varepsilon})}{\widehat{\mathcal{V}ar}(Y)} .$$

In the context of ANOVA methods, it is often referred to as the *empirical correlation ratio* between the quantitative variable Y and the factor considered. It measures the relationship between a quantitative variable and a qualitative variable.

Let us mention the following two particular cases:

$R^2 = 1$. Hence, $\hat{\varepsilon} = 0_n$, that is $\forall j \in [\![1, n_i]\!], Y_{ij} = \bar{Y}_{i\cdot}$.
i.e. Y is constant in each level.

$R^2 = 0$. Hence, $\widehat{\mathcal{V}ar}(\hat{Y}) = 0$, that is $\forall i \in [\![1, I]\!], \bar{Y}_{i\cdot} = \bar{Y}_{\cdot\cdot}$.
i.e. the mean of Y is the same in each cell.

1: Recall that, in $(\mathcal{M}_I^{\sqcup})$

- $SST = n\,\widehat{\mathcal{V}ar}(Y)$,
- $SSR = n\,\widehat{\mathcal{V}ar}(\hat{Y})$, *inter-group sums of squares,*
- $SSE = n\,\widehat{\mathcal{V}ar}(\hat{\varepsilon})$, *intra-group sums of squares,*

and note that this proposition is only a rewriting of the general result

$$SST = SSE + SSR .$$

2: Also, this formula is just a rephrasing in the anova vocabulary of

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} .$$

11.4.4 Confidence in the Estimate

In the general framework of the Gaussian model, it has been shown that the estimators of the model parameters are Gaussian distributed. This property can be applied to the one-factor ANOVA model as long as normality and independence of errors are assumed.

To construct a confidence interval for the m_i, it is therefore sufficient to construct a Student confidence interval using that, in the complete regular model $(\mathcal{M}_I^\star)$,

$$\hat{m}_i \sim \mathcal{N}(m_i, \frac{\sigma^2}{n_i}) \qquad \text{and} \qquad (n-I)\hat{\sigma}^2 \sim \sigma^2\chi^2(n-I)\,.$$

So we get

$$CI_{1-\delta}(m_i) = \left[\hat{m}_i \pm t_{n-I,1-\delta/2}\frac{\hat{\sigma}}{\sqrt{n_i}}\right].$$

Under R, we execute the `confint` command, see Listing 11.1.

Exercise 11.9 *Construct the confidence intervals given in Listing 11.2.*

Listing 11.1: Confidence interval for the regular model $(\mathcal{M}_I^\star)$.

```
> anov_reg = lm(mark~exam-1,
                data=note)
> confint(anov_reg)

          2.5 %   97.5 %
examA  9.998705 14.00129
examB 11.266828 14.73317
examC 12.147161 15.85284
```

Listing 11.2: Confidence interval for the singular model $(\mathcal{M}_I)$.

```
> anov_sing = lm(mark~exam,
                data=note)
> confint(anov_sing)

              2.5 %   97.5 %
(Intercept)  9.99871 14.00129
examB       -1.64746  3.64746
examC       -0.72731  4.72730
```

11.5 Factor Effect Test

As said in Section 11.3.2, we can study the effect of the factor on the variable Y by assuming equality of all the parameters of the model:

$$\mathcal{H}_0: \begin{cases} \forall i, i' \in [\![1, I]\!], & m_i = m_{i'} := m & \text{in} & (\mathcal{M}_I^\star) \\ \forall i \in [\![1, I]\!], & \alpha_i = 0 & \text{in} & (\mathcal{M}_I) \end{cases}$$

versus

$$\mathcal{H}_1: \begin{cases} \exists i, i' \in [\![1, I]\!],\ i \neq i', & \text{such that} \quad m_i \neq m_{i'} & \text{in} & (\mathcal{M}_I^\star) \\ \exists i \in [\![1, I]\!], & \text{such that} \quad \alpha_i \neq 0 & \text{in} & (\mathcal{M}_I)\,. \end{cases}$$

Under $\mathcal{H}_0$, all parameters m_i are equal and the model writes

$$(\mathcal{M}_1^\star): \quad Y_{ij} = m + \varepsilon_{ij}\,, \qquad \text{where} \qquad \hat{m} = \bar{Y}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij}\,.$$

In other words, we seek to compare the sub-model $(\mathcal{M}_1^\star)$ to the complete model $(\mathcal{M}_I^\star)$. We will therefore perform a Fisher sub-model test.

The equality of the parameters induces that the error sum of squares SSE_1 in the model $(\mathcal{M}_1^\star)$ is equal to the total sum of squares SST in the complete model. Hence, $SSE_1 - SSE = SST - SSE = SSR$.

The Fisher test statistic thus writes:

$$F = \frac{\frac{SSR}{(I-1)}}{\frac{SSE}{(n-I)}} = \frac{\frac{1}{I-1}\sum_{i=1}^{I} n_i \left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2}{\frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2} \underset{\sim}{\mathcal{H}_0} \mathcal{F}(I-1, n-I).$$

We reject $\mathcal{H}_0$ at the level δ if $F > f_{1-\delta,I-1,n-I}$.

In R, we obtain the following output.

```
> anov_cst = lm(mark~1, data=note)
> anova(anov_cst, anov_reg)

Analysis of Variance Table

Model 1: mark ~ 1
Model 2: mark ~ exam - 1
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     20 110.95
2     18  98.00  2    12.952 1.1895 0.3272
```

Exercise 11.10 *In the following two outputs, which hypotheses are being tested? Construct the associated Fisher test. Note the difference between the two procedures.*

```
> anova(anov_reg)

Analysis of Variance Table

Response: mark
          Df Sum Sq Mean Sq F value    Pr(>F)
exam       3   3588 1196.00  219.67 2.311e-14 ***
Residuals 18     98    5.44
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> anova(anov_sing)

Analysis of Variance Table

Response: mark
          Df Sum Sq Mean Sq F value Pr(>F)
exam       2 12.952  6.4762  1.1895 0.3272
Residuals 18 98.000  5.4444
```

11.5.1 Interpretations of the ANOVA Test

The test statistic F defined above can be seen as the ratio of two estimators of σ^2: one that is always good, and one that is only good under $\mathcal{H}_0$. Indeed, we can show that:

$$\mathbb{E}\left[\frac{SSR}{I-1}\right] = \sigma^2 + \frac{1}{I-1}\sum_{i=1}^{I} n_i(m_i-\mu)^2\,, \quad \text{where} \quad \mu = \frac{1}{n}\sum_{i=1}^{I} n_i m_i\,.$$

The quantity $\sum_{i=1}^{I} n_i(m_i-\mu)^2$ is null if and only if for all $i \in [\![1, I]\!]$, $m_i = \mu$, *i.e.* when all averages are equal, *i.e.* when $\mathcal{H}_0$ is true. We can therefore deduce that under $\mathcal{H}_0$, $\frac{SSR}{I-1}$ is an unbiased estimator of σ^2.

Thus, testing the absence of effect of the factor is to compare two estimators of σ^2 :

- one which is only good under $\mathcal{H}_0$, the one given by $\frac{1}{I-1}SSR$, and
- one which is always good, the one obtained in the model ($\mathcal{M}_I^\star$) and given by $\frac{1}{n-I}SSE$.

Remark 11.1 Under $\mathcal{H}_0$, the inter-group variability SSR is comparable to the intra-group variability SSE, since all individual means are confounded.

11.6 One-Factor Analysis of Variance Table

All these estimates can be displayed in a one-factor analysis of variance table:

Source of variation	Degree of freedom	Sum of Squares	Mean sum of Squares	Test statistics	$f_{1-\delta}$
Factor	$I-1$	$SSR = \sum_{i=1}^{I} n_i\left(\bar{Y}_i. - \bar{Y}..\right)^2$	$MSR = \frac{SSR}{I-1}$	$\frac{MSR}{\hat{\sigma}^2}$	$f_{1-\delta,I-1,n-I}$
Residual	$n-I$	$SSE = \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i.\right)^2$	$\hat{\sigma}^2 = \frac{SSE}{n-I}$		
Total	$n-1$	$SST = \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}..\right)^2$			

11.7 Robustness to Assumptions

The one-way ANOVA methodology is more or less robust to the failure of modeling assumptions, namely normality, homoscedasticity, and error independence.

Specifically:

1. The methodology is robust to sample non-normality;
2. The non-homogeneity of variances can be circumvented;
3. The most severe problem is the non-respect of the independence of the errors, in which case it is necessary to use other models than the one-factor ANOVA.

An analysis of the residuals should be carried out before using the model to try to validate it by verifying the model's hypotheses in a descriptive way or by adequate tests.

For example, one can test the homogeneity of the variances, *i.e.* test $\mathcal{H}_0$: "$\sigma_1^2 = \ldots = \sigma_I^2$" against $\mathcal{H}_1$: "$\exists (i,j)$ such that $\sigma_i^2 \neq \sigma_j^2$", where σ_i^2 denotes the variance of the i-th sample. This can be done using Bartlett's test, which is sensitive to non-normality, or Cochran's test, which is robust to non-normality but only applies when the I samples have the same size.

When the equality of variances is not satisfied, one can, for instance, use the non-parametric Kruskal-Wallis test to determine whether the distributions of the I samples are identical.

11.8 Test of Comparison of Variances

The homogeneity of variances between groups is crucial in ANOVA methods, but one rarely checks it. However, it can be tested in different ways. The simplest solution would be to carry out $I(I-1)/2$ comparisons of the variances of all the groups using the classical test of equality of variances of two Gaussian samples. In other words, the simplest solution is to test for any pair (i,j) the hypothesis $\mathcal{H}_0$: "$\sigma_i^2 = \sigma_j^2$" against the alternative $\mathcal{H}_1$: "$\sigma_i^2 \neq \sigma_j^2$". However, we then face the problem of multiple tests: if we choose to perform each test at a level of 5%, we cannot guarantee anything about the global level after having performed the $I(I-1)/2$ tests. Other test procedures (more or less robust to the underlying modeling assumptions) allow for globally testing of the equality of variances, such as the Bartlett test (sensitive to non-normality), the Levene test, or the Cochran test. In the following, we present the *Bartlett test*.

We make the assumptions of normality and independence of the I samples, *i.e.* we suppose that the data y_{ij} are the realizations of random variables Y_{ij} of law $\mathcal{N}(m_i, \sigma_i^2)$, the variables Y_{ij} being globally independent. We pose the null hypothesis

$$\mathcal{H}_0 : \text{“}\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_I^2\text{”}$$

that we want to test against the alternative hypothesis

$$\mathcal{H}_1 : \text{“}\exists i, j \in [\![1, I]\!],\ \sigma_i^2 \neq \sigma_j^2\text{”}.$$

Let us denote S_i^2 the unbiased estimators of the variances σ_i^2 of the i-th sample. Recall that

$$\forall i \in [\![1, I]\!], \quad S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2 ,$$

and therefore, we can rewrite SSE as $SSE = \sum_{i=1}^{I} (n_i - 1) S_i^2$.

Under the assumption $\mathcal{H}_0$, we show that the statistic

$$\frac{2.3026}{C} \left[(n - I) \ln \left(\frac{SSE}{n - I} \right) - \sum_{i=1}^{I} (n_i - 1) \ln(S_i^2) \right] ,$$

where $C = 1 + \frac{1}{3(I-1)} \left[\sum_{i=1}^{I} \frac{1}{n_i - 1} - \frac{1}{I - 1} \right]$, follows approximately a χ^2 with $(I - 1)$ degrees of freedom. Of course, under $\mathcal{H}_1$, this random variable no longer follows a $\chi^2(I - 1)$. This result is therefore sufficient to construct a test of $\mathcal{H}_0$ against $\mathcal{H}_1$.

Two-Way Analysis of Variance 12

In this chapter, we generalize the framework of the analysis of variance from one factor to two factors. In other words, we aim to study the influence of two qualitative variables on a quantitative variable.

As in the one-way ANOVA framework, we assume that :

- the factors influence only the mean of the quantitative variable Y and not its variance;
- the effects of the factors are additive;
- the variations other than those caused by the factors are Gaussian and independent.

Suppose some biologists want to study the wheat yield, but with three different species of wheat, while testing the two different fertilizer levels. The biologists need to investigate not only the average growth between the three species (main effect A) and the average growth for the two fertilizer levels (main effect B), but also the interaction or relationship between the two factors of species and fertilizer. Two-way analysis of variance allows biologists to answer the question about yield affected by species *and* fertilizer levels, and to account for the variation due to both factors simultaneously.

12.1 Two-Way Analysis of Variance

Let Y be the quantitative response variable we want to explain here using two qualitative variables or factors.

- The first factor, called the *row factor* or A, admits I levels;
- the second, called the *column factor* or B, admits J levels.

We assume that the response variable observations are independent and normally distributed with a mean that may depend on the levels of factors A *and* B, but with a constant variance.

A particular combination of levels is called a *treatment* or a cell. There are IJ treatments. In the following, we note:

- $i \in [\![1, I]\!]$: indices of the levels of the line factor A;
- $j \in [\![1, J]\!]$: indices of the levels of the column factor B;
- n_{ij}: number of observations for the level i of the factor A, and for the level j of the factor B, *i.e.* the number of observations in the cell (i, j);
- $\ell \in [\![1, n_{ij}]\!]$: indices of the observations of the cell (i, j);
- $Y_{ij\ell}$: the ℓ-th observation in cell (i, j);
- $\bar{Y}_{ij\cdot}$: average of the observations in the cell (i, j).

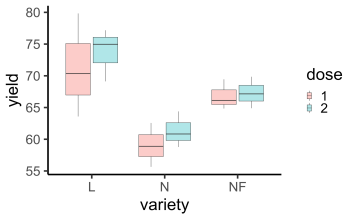


Figure 12.1: Boxplot of wheat yield according to dose and variety.

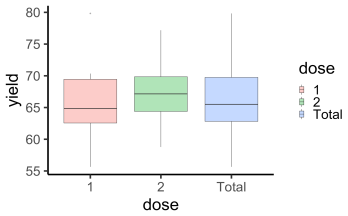


Figure 12.2: Boxplot of wheat yield overall (right) and by dose.

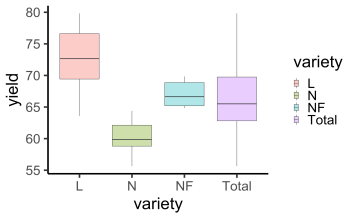


Figure 12.3: Boxplot of wheat yield overall (right) and by variety.

Listing 12.1: Dataset for two-way analysis of variance

```
> summary(wheat)

dose  variety     yield
1:9   L :6    Min.   :55.65
2:9   N :6    1st Qu.:62.82
      NF:6    Median :65.50
              Mean   :66.58
              3rd Qu.:69.75
              Max.   :79.83
```

In the rest of this chapter, we will consider the example described in the preamble and taken from the book of Husson and Pagès [1]. During a study on factors influencing wheat yield, biologists compared

- three wheat varieties: factor B with modalities L, N and NF,
- and two nitrogen applications: factor A of modality "normal application" or dose 1, and "intensive application" or dose 2.

The observation of the couple (variety, dose) is repeated three times. In other words, $n_{ij} = 3$ for all treatment (i, j); the experimental design is said to be balanced. For each replication, we measured Y_{ij}, the yield in q/ha. We are investigating whether there are any differences between varieties or interactions between varieties and nitrogen inputs.

Figure 12.2 displays wheat yield as a function of nitrogen dose; Figure 12.3 displays the same yield but as a function of wheat variety. Finally, Figure 12.1, more classical for two-way ANOVA, represents the yield as a function of the nitrogen dose and the variety.

Definition 12.1 *In the context of ANOVA, we call* dimension *of the model the dimension of the space in which the expectation of the random variables $Y_{ij\ell}$ lives. This dimension is equal to the number of expectation parameters considered in the modeling minus the number of identifiability constraints necessary to estimate the parameters.*

Hereafter, we will denote $\mathcal{M}_d$ each model, where d denotes the dimension of the model then considered.

In the same way as for the one-way ANOVA, we assume that for all (i, j, ℓ), the data $y_{ij\ell}$ is a realization of a random variable $Y_{ij\ell}$ of law $\mathcal{N}(m_{ij}, \sigma^2)$, and that the $Y_{ij\ell}$ are globally independent.

In other words, we assume the following general model with two crossed factors:

$$\forall i \in [\![1, I]\!], \ \forall j \in [\![1, J]\!], \ \forall \ell \in [\![1, n_{ij}]\!], \quad Y_{ij\ell} = m_{ij} + \varepsilon_{ij\ell}, \quad \text{where} \quad \varepsilon_{ij\ell} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2), \tag{$\mathcal{M}^\star_{IJ}$}$$

where m_{ij} is the theoretical mean or expected value of all observations in cell (i, j). We refer to this model as the "*cell means model*"

12.1.1 Decomposition of Effects

This treatment modeling describes the joint effect of the two factors by assuming a specific expectation m_{ij} for each treatment (i, j) and an intra-treatment variance σ^2 common to all treatments. In particular, it does not allow for distinguishing the effects of each factor nor their interaction. Hence, as with the one-factor ANOVA, we often prefer a singular parametrization called the centered parametrization, or "'*factor effects model*". It can decompose m_{ij} into a general mean effect, separate

main effects of the factors, and interaction or joint effects. The complete model then writes, for all $i \in [\![1, I]\!]$, $j \in [\![1, J]\!]$ and $\ell \in [\![1, n_{ij}]\!]$, as:

$$Y_{ij\ell} = \underbrace{\mu + \alpha_i + \beta_j + \gamma_{ij}}_{m_{ij}} + \varepsilon_{ij\ell}\,, \quad \text{where} \quad \varepsilon_{ij\ell} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)\,. \qquad (\mathcal{M}_{IJ})$$

The IJ parameters m_{ij} are thus redefined as a function of:

- μ: the overall effect, *i.e.* the common effect of the two factors regardless of their modalities,
- α_i: $I - 1$ parameters that characterize the main effects of factor A,
- β_j: $J - 1$ parameters that characterize the main effects of factor B,
- γ_{ij}: $(I-1)(J-1)$ interaction terms. They allow to take into account the specific effect of the treatments beyond the sum of their main effects (non-additivity of the main effects).

We then have a model defined via $1 + I + J + IJ$ parameters. Thus, as said, this model is singular. In particular, we must introduce $1 + I + J$ constraints to estimate its parameters. In Chapter 8, we have seen the interest in considering constraints in the framework of an orthogonal system. In the case of the analysis of variance with two crossed factors, the following property characterizes this orthogonality.

Proposition 12.1 *In the two-way analysis of variance model, there exist constraints that make the partition μ, α, β, γ orthogonal if and only if*

$$\forall i \in [\![1, I]\!], \quad \forall j \in [\![1, J]\!], \quad n_{ij} = \frac{n_{i+}\, n_{+j}}{n}\,.$$

In this case, the constraints, called type I constraints, *are*

$$\begin{aligned} &\sum_{i=1}^{I} n_{i+}\,\alpha_i = 0\,, && \sum_{j=1}^{J} n_{+j}\,\beta_j = 0\,, \\ \forall i,\ &\sum_{j=1}^{J} n_{ij}\,\gamma_{ij} = 0\,, && \forall j,\ \sum_{i=1}^{I} n_{ij}\,\gamma_{ij} = 0\,. \end{aligned} \qquad (\mathcal{C}^{\perp})$$

Exercise 12.2 *Show that a complete and balanced experimental design, i.e. a design such that for all treatment (i, j) $n_{ij} = cste > 0$, is orthogonal.*

The converse is not true.

In practice, the constraints used are often those of *type III*:

$$\sum_i \alpha_i = 0\,, \quad \sum_j \beta_j = 0\,, \quad \forall i,\ \sum_j \gamma_{ij} = 0 \quad \text{and} \quad \forall j,\ \sum_j \gamma_{ij} = 0\,.$$

With this constraint system, orthogonality is only possible if the model is balanced, i.e., if n_{ij} is constant according to Proposition 12.1.

Remark 12.1 In these identifiability constraints,

$$\forall i,\ \sum_j \gamma_{ij} = 0 \quad \text{and} \quad \forall j,\ \sum_j \gamma_{ij} = 0$$

are not independent. Indeed, we can show that the $(I + J)$ equations defining the constraints can actually be reduced to $(I + J - 1)$ equations.

Exercise 12.3 *Demonstrate this for $I = 2$ and $J = 3$. More precisely, show that the linear system made of the 5 constraints actually reduces to a linear system with four equations.*

Beware! The $(C^\perp)$ constraints are not the default constraints under `R` (*cf.* `model.matrix(yield~dose*variety)`). Indeed, under `R`, in a similar way to the one-way ANOVA, the default constraints are

$$\alpha_1 = \beta_1 = \gamma_{1j} = \gamma_{i1} = 0\,. \qquad (C^{\mathsf{R}})$$

But, they can be easily modified.

In the following, we consider an orthogonal experimental design.

12.1.2 Two-Way Additive ANOVA

However, the presence of the interaction effect is not systematic. For example, when $n_{ij} = 1$ for all (i, j) (absence of repetitions), we cannot take this term into account in the modeling because we do not observe enough data to estimate it, which does not mean that the interaction does not exist.

In the additive two-way ANOVA model, we assume that there is no interaction effect between the two factors. Then, for all $i \in [\![1, I]\!]$, $j \in [\![1, J]\!]$ and $\ell \in [\![1, n_{ij}]\!]$, the model writes

$$Y_{ij\ell} = \mu + \alpha_i + \beta_j + \varepsilon_{ij\ell}\,, \quad \text{where} \quad \varepsilon_{ij\ell} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)\,. \qquad (\mathcal{M}_{I+J-1})$$

Note that the additive model is a sub-model of the complete model with interaction.

Exercise 12.4 *Prove that this model is orthogonal under the constraints $\sum_{i=1}^I \alpha_i = 0$ and $\sum_{j=1}^J \beta_j = 0$.*

12.1.3 Model Without Effect of Factor A

The model where a possible effect of factor A is not taken into account is defined by, for all $i \in [\![1, I]\!]$, $j \in [\![1, J]\!]$ and $\ell \in [\![1, n_{ij}]\!]$,

$$Y_{ij\ell} = \mu + \beta_j + \varepsilon_{ij\ell}\,, \quad \text{where} \quad \varepsilon_{ij\ell} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)\,, \qquad (\mathcal{M}_J)$$

and the identifiability constraint $\sum_{j=1}^{J} \beta_j = 0$.

12.1.4 Model Without Effect of Factor B

Likewise, the model where a possible effect of factor B is not taken into account is defined by, for all $i \in [\![1, I]\!]$, $j \in [\![1, J]\!]$ and $\ell \in [\![1, n_{ij}]\!]$,

$$Y_{ij\ell} = \mu + \alpha_i + \varepsilon_{ij\ell}\,, \quad \text{where} \quad \varepsilon_{ij\ell} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)\,, \qquad (\mathcal{M}_I)$$

and the identifiability constraint $\sum_{i=1}^{I} \alpha_i = 0$.

12.1.5 Model Without Treatment Effect

The model in which neither of the two factors A and B are taken into account is

$$Y_{ij\ell} = \mu + \varepsilon_{ij\ell}\,, \quad \text{where} \quad \varepsilon_{ij\ell} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)\,. \qquad (\mathcal{M}_1^\star)$$

This model is regular and does not require identifiability constraints.

12.2 Estimation and Forecasting

Let the following notations: For all $i \in [\![1, I]\!]$ and $j \in [\![1, J]\!]$,

$$\bar{Y}_{i\cdot\cdot} = \frac{1}{n_{i+}} \sum_{j=1}^{J} \sum_{\ell=1}^{n_{ij}} Y_{ij\ell}\,, \quad \text{where} \quad n_{i+} = \sum_{j=1}^{J} n_{ij}\,,$$

$$\bar{Y}_{\cdot j\cdot} = \frac{1}{n_{+j}} \sum_{i=1}^{I} \sum_{\ell=1}^{n_{ij}} Y_{ij\ell}\,, \quad \text{where} \quad n_{+j} = \sum_{i=1}^{I} n_{ij}\,,$$

$$\bar{Y}_{\cdots} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{\ell=1}^{n_{ij}} Y_{ij\ell}\,, \quad \text{where} \quad n = \sum_{i=1}^{I} n_{i+} = \sum_{j=1}^{J} n_{+j}\,.$$

12.2.1 Estimation in the Cell Means Model

Proposition 12.5 *Let $i \in [\![1, I]\!]$, $j \in [\![1, J]\!]$ and $\ell \in [\![1, n_{ij}]\!]$.*
In the regular *parametrization $(\mathcal{M}_{IJ}^\star)$: $Y_{ij\ell} = m_{ij} + \varepsilon_{ij\ell}$,*

- m_{ij} *is estimated by*

$$\hat{m}_{ij} = \frac{1}{n_{ij}} \sum_{\ell=1}^{n_{ij}} Y_{ij\ell} = \bar{Y}_{ij\cdot} \sim \mathcal{N}\left(m_{ij}, \frac{\sigma^2}{n_{ij}}\right);$$

- *the variance is estimated by*

$$\hat{\sigma}^2 = \frac{1}{n - IJ} \sum_{ij\ell} \left(\hat{\varepsilon}_{ij\ell}\right)^2 = \frac{1}{n - IJ} \sum_{ij\ell} \left(Y_{ij\ell} - \bar{Y}_{ij\cdot}\right)^2 .$$

Exercise 12.6 *Prove this proposition by using the fact that we are dealing with a regular linear model.*

The forecast of a $Y_{ij\ell}$ in this model is therefore given by:

- Adjusted values: $\hat{Y}_{ij\ell} = \hat{m}_{ij} = \bar{Y}_{ij\cdot}$;
- Residuals: $\hat{\varepsilon}_{ij\ell} = Y_{ij\ell} - \bar{Y}_{ij\cdot}$.

12.2.2 Estimation in the Factor Effects Model

Proposition 12.7 *Let $i \in [\![1, I]\!]$, $j \in [\![1, J]\!]$, $\ell \in [\![1, n_{ij}]\!]$ and the complete* singular *model $(\mathcal{M}_{IJ})$: $Y_{ij\ell} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij\ell}$.*
Then, under the type I constraints $(C^\perp)$, we have the following estimates:

$$\begin{cases} \hat{\mu} = \bar{Y}_{\cdots}, \\ \hat{\alpha}_i = \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots}, \\ \hat{\beta}_j = \bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdots}, \\ \hat{\gamma}_{ij} = \bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdots}. \end{cases}$$

Exercise 12.8 *Prove this proposition. To this aim, you can minimize the least squares function under the constraints $(C^\perp)$.*

The forecast of a $Y_{ij\ell}$ in this model is therefore given by

$$\hat{Y}_{ij\ell} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij} = \bar{Y}_{ij\cdot}.$$

For our example, the results obtained with R are reported below. The first output is related to the default constraints used by R (C^{R}), while the second implements the orthogonality constraints ($C^{\perp}$).

```
> summary(lm(yield ~ dose * variety, data=wheat))

Call:
lm(formula = yield ~ dose * variety, data = wheat)

Residuals:
   Min     1Q Median     3Q    Max
-7.667 -2.296 -0.325  2.623  8.573

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      71.257      2.536  28.101 2.55e-12 ***
dose2             2.500      3.586   0.697  0.49899
varietyN        -12.223      3.586  -3.409  0.00519 **
varietyNF        -4.453      3.586  -1.242  0.23801
dose2:varietyN   -0.200      5.071  -0.039  0.96919
dose2:varietyNF  -2.007      5.071  -0.396  0.69928
---
Signif. codes:  0 ‘***‘ 0.001 ‘**‘ 0.01 ‘*‘ 0.05 ‘.‘ 0.1 ‘ ‘ 1

Residual standard error: 4.392 on 12 degrees of freedom
Multiple R-squared:  0.6725,    Adjusted R-squared:  0.536
F-statistic: 4.928 on 5 and 12 DF,  p-value: 0.01105
```

Listing 12.2: Default constraints (C^{R})

```
> summary(lm(yield ~ C(dose,sum) + C(variety,sum)
                   + C(dose,sum):C(variety,sum), data=wheat))

Call:
lm(formula = yield ~ C(dose, sum) + C(variety, sum)
                   + C(dose, sum):C(variety, sum), data = wheat
    )

Residuals:
   Min     1Q Median     3Q    Max
-7.667 -2.296 -0.325  2.623  8.573

Coefficients:
                          Estimate Std.Err t value Pr(>|t|)
(Intercept)                 66.580   1.035  64.316  < 2e-16 ***
C(dose, sum)1               -0.882   1.035  -0.852 0.410775
C(variety, sum)1             5.927   1.464   4.048 0.001615 **
C(variety, sum)2            -6.397   1.464  -4.369 0.000913 ***
C(dose,sum)1:C(variety,sum)1 -0.368   1.464  -0.251 0.805897
C(dose,sum)1:C(variety,sum)2 -0.268   1.464  -0.183 0.857923
---
Signif. codes:  0 ‘***‘ 0.001 ‘**‘ 0.01 ‘*‘ 0.05 ‘.‘ 0.1 ‘ ‘ 1

Residual standard error: 4.392 on 12 degrees of freedom
Multiple R-squared:  0.6725,    Adjusted R-squared:  0.536
F-statistic: 4.928 on 5 and 12 DF,  p-value: 0.01105
```

Listing 12.3: Orthogonality constraints ($C^{\perp}$)

Exercise 12.9 *Check that $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\beta}_j$ and $\hat{\gamma}_{ij}$ are indeed unbiased estimators*

Note that we have not yet estimated the residual variance σ^2. For this, we need an additional assumption (see Section 12.3.2).

12.2.3 Estimation in Sub-Models

Since the experimental design is orthogonal, for all submodels of the full model ($\mathcal{M}_{IJ}$), the estimate of the parameters μ, α_i and β_j is unchanged. Regardless of the submodel considered, we have

$$\begin{cases} \hat{\mu} = \bar{Y}_{\cdots}, \\ \hat{\alpha}_i = \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots}, \\ \hat{\beta}_j = \bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdots}. \end{cases}$$

Using the estimators of the different parameters, we obtain, for each sub-model, the following predictions:

- Additive model ($\mathcal{M}_{I+J-1}$): $\hat{Y}_{ij\ell} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j = \bar{Y}_{i\cdot\cdot} + \bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdots}$;
- No factor A model ($\mathcal{M}_J$): $\hat{Y}_{ij\ell} = \hat{\mu} + \hat{\beta}_j = \bar{Y}_{\cdot j\cdot}$;
- No factor B model ($\mathcal{M}_I$): $\hat{Y}_{ij\ell} = \hat{\mu} + \hat{\alpha}_i = \bar{Y}_{i\cdot\cdot}$;
- No treatment model ($\mathcal{M}_1^\star$): $\hat{Y}_{ij\ell} = \hat{\mu} = \bar{Y}_{\cdots}$.

12.3 Variance Analysis

In the previous paragraphs, we were only interested in the least squares estimation of the parameters modeling the average of the observations, *i.e.* μ, α, β and γ. Our goal is now to study the intra-class variance σ^2.

12.3.1 Variance Decomposition

Let assume an orthogonal design. Then, as in the one-way analysis of variance, the total variability of Y is decomposed into inter- and intra-group variability:

- SSR, the inter-variance, explained by the model,
- and SSE the intra-variance, not-explained by the model.

Consider the complete model $(\mathcal{M}_{IJ})$. We find the same decomposition as before:

$$\underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{\ell=1}^{n_{ij}}\left(Y_{ij\ell}-\bar{Y}_{\dots}\right)^2}_{SST} = \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\left(\bar{Y}_{ij\cdot}-\bar{Y}_{\dots}\right)^2}_{SSR} + \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\widehat{\mathcal{V}ar}_{ij}(Y)}_{SSE},$$

where $\widehat{\mathcal{V}ar}_{ij}(Y) = \dfrac{1}{n_{ij}}\displaystyle\sum_{\ell=1}^{n_{ij}}\left(Y_{ij\ell}-\bar{Y}_{ij\cdot}\right)^2$.

In the case of the two-way crossed model, the inter-cell variance SSR can be decomposed into a variance explained by the first factor A, a variance explained by the second factor B, and a variance explained by the interactions between the two factors. Hence, for two-way orthogonal design, we define the following quantities:

- SSA, the factor A main effect sums of squares:

$$SSA = \sum_{i=1}^{I} n_{i+}(\bar{Y}_{i\cdot\cdot}-\bar{Y}_{\dots})^2 = \sum_{i=1}^{I} n_{i+}\,(\hat{\alpha}_i)^2\,;$$

- SSB, the factor B main effect sums of squares:

$$SSB = \sum_{j=1}^{J} n_{+j}(\bar{Y}_{\cdot j\cdot}-\bar{Y}_{\dots})^2 = \sum_{j=1}^{J} n_{+j}\,(\hat{\beta}_j)^2\,;$$

- $SSAB$, the interaction sum of squares:

$$SSAB = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}(\bar{Y}_{ij\cdot}-\bar{Y}_{i\cdot\cdot}-\bar{Y}_{\cdot j\cdot}+\bar{Y}_{\dots})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\,(\hat{\gamma}_{ij})^2\,.$$

Hence, we can prove that $SSR = SSA + SSB + SSAB$, leading to

$$SST = SSA + SSB + SSAB + SSE\,.$$

Let $d \in \{1, I, J, I+J-1, IJ\}$. Given the model $(\mathcal{M}_d)$, we will from now denote SSR_d and SSE_d as the corresponding sums of squares[1] in case of ambiguity. When the dimension is not specified, we place ourselves in the complete model with interaction.

1: Note that the definition of SST does not depend on the considered model, as well as SSA, SSB and $SSAB$, once they are defined.

12.3.2 Variance Estimation

Let us now assume a *complete and balanced* experimental design, *i.e.* that for each treatment (i, j), we have a constant and strictly positive number of measures of Y: $n_{ij} = L$. Note that in this case $n = \sum_i^I \sum_{j=1}^J n_{ij} = IJL$. This design is in particular orthogonal.

Proposition 12.10 *Under the assumptions of the full balanced model with interaction,* $\hat{\sigma}^2 = \dfrac{SSE_d}{n-d}$ *is an unbiased estimator of* σ^2 *in the model* $(\mathcal{M}_d)$. *Moreover,*

$$(n-d)\,\hat{\sigma}^2 \sim \sigma^2\,\chi^2(n-d)\,.$$

At Table 12.1, you can find for each model $(\mathcal{M}_d)$ the value of their error sum of squares. In this table, the notation $\sum_{i,j,\ell}$ refers to the triple sum on $i \in [\![1, I]\!]$, $j \in [\![, J]\!]$ and $\ell \in [\![1, L]\!]$.

Table 12.1: Error sum of squares, or intra-class variance, depending on the considered sub-model of the complete model.

SSE_{IJ}	$\sum_{i,j,\ell} \left(Y_{ij\ell} - \bar{Y}_{ij\cdot}\right)^2$
SSE_{I+J-1}	$\sum_{i,j,\ell} \left(Y_{ij\ell} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdot\cdot\cdot}\right)^2$
SSE_I	$\sum_{i,j,\ell} \left(Y_{ij\ell} - \bar{Y}_{i\cdot\cdot}\right)^2$
SSE_J	$\sum_{i,j,\ell} \left(Y_{ij\ell} - \bar{Y}_{\cdot j\cdot}\right)^2$
SSE_1	$\sum_{i,j,\ell} \left(Y_{ij\ell} - \bar{Y}_{\cdot\cdot\cdot}\right)^2$

We demonstrate this result in the $(\mathcal{M}_{IJ})$ model for more readability. However, the proof below can be adapted to any of its sub-models.

Proof. First, note that $\sum_{\ell=1}^{L} \left(Y_{ij\ell} - \bar{Y}_{ij\cdot}\right) = 0$. Hence,

$$\begin{aligned}
\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{\ell=1}^{L} \varepsilon_{ij\ell}^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{\ell=1}^{L} \left(Y_{ij\ell} - \bar{Y}_{ij\cdot} + \bar{Y}_{ij\cdot} - \mu - \alpha_i - \beta_j - \gamma_{ij}\right)^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{\ell=1}^{L} \left(\bar{Y}_{ij\cdot} - \mu - \alpha_i - \beta_j - \gamma_{ij}\right)^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{\ell=1}^{L} \left(Y_{ij\ell} - \bar{Y}_{ij\cdot}\right)^2 \\
&\quad + 2\sum_{i=1}^{I}\sum_{j=1}^{J} \underbrace{\sum_{\ell=1}^{L} \left(\bar{Y}_{ij\cdot} - \mu - \alpha_i - \beta_j - \gamma_{ij}\right)\left(Y_{ij\ell} - \bar{Y}_{ij\cdot}\right)}_{= \left(\bar{Y}_{ij\cdot} - \mu - \alpha_i - \beta_j - \gamma_{ij}\right) \sum_{\ell=1}^{L} \left(Y_{ij\ell} - \bar{Y}_{ij\cdot}\right) = 0} \\
&= L\sum_{i=1}^{I}\sum_{j=1}^{J} \left(\bar{Y}_{ij\cdot} - \mu - \alpha_i - \beta_j - \gamma_{ij}\right)^2 + SSE_{IJ}\,.
\end{aligned}$$

Then, since the random variables $\varepsilon_{ij\ell}$ are i.i.d of law $\mathcal{N}(0, \sigma^2)$, we deduce that

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{\ell=1}^{L} \left(\frac{\varepsilon_{ij\ell}}{\sigma}\right)^2 \sim \chi^2(n)\,, \quad \text{where} \quad n = IJL\,.$$

In the same way, since the random variables $\bar{Y}_{ij\cdot}$ are independent of law $\mathcal{N}\left(\mu + \alpha_i + \beta_j + \gamma_{ij}, \frac{\sigma^2}{r}\right)$, we deduce that the variables $\frac{\sqrt{L}}{\sigma}\left(\bar{Y}_{ij\cdot} - (\mu + \alpha_i + \beta_j + \gamma_{ij})\right)$ are i.i.d of law $\mathcal{N}(0, 1)$. Consequently,

$$L\sum_{i=1}^{I}\sum_{j=1}^{J} \left(\frac{\bar{Y}_{ij\cdot} - \mu - \alpha_i - \beta_j - \gamma_{ij}}{\sigma}\right)^2 \sim \chi^2(IJ)$$

and we conclude using Cochran's theorem. □

12.4 Factor Effect Test

In two-factor ANOVA, three assumptions are commonly considered:

1. The assumption of *non-interaction between the two factors* or additivity of the two factors: Within $(\mathcal{M}_{IJ})$,

$$\mathcal{H}_0^{AB}\colon \forall (i, j) \in [\![1, I]\!] \times [\![1, J]\!]\,, \quad \gamma_{ij} = 0\,.$$

   This assumption imposes $(I - 1)(J - 1)$ constraints.

2. The assumption of *no effect of the factor A*: Within $(\mathcal{M}_{I+J-1})$,

$$\mathcal{H}_0^{A}\colon \forall i \in [\![1, I]\!]\,, \quad \alpha_i = 0\,.$$

This assumption imposes $(I-1)$ constraints.

3. The assumption of *no effect of the factor B*: Within $(\mathcal{M}_{I+J-1})$,

$$\mathcal{H}_0^B : \forall j \in [\![1, J]\!], \quad \beta_j = 0 .$$

This assumption imposes $(J-1)$ constraints.

Remark 12.2 A crucial point about the approach to these hypothesis tests: *If there are interactions between the two factors, both factors that make up the interaction must be introduced into the model.* In this case, there is no need to test the effect of each of the two factors. Indeed, the presence of interactions between the two factors means that there is a combined effect and, thus, in particular, an effect of each factor.

Regarding the presence or not of interaction, we can start by considering a graphical response.

12.4.1 Interaction Plot

The interaction diagram allows to visualize graphically the presence or absence of interactions.

For each fixed j, we represent in an orthogonal reference frame the $M_{(i,j)}$ of coordinates $(i,\ \hat{m}_{ij} = \bar{Y}_{ij\cdot})$. Then, we draw the segments joining the pairs of points $M_{(i-1,j)}$ and $M_{(i,j)}$. Thus, for each j fixed, we obtain a broken line.

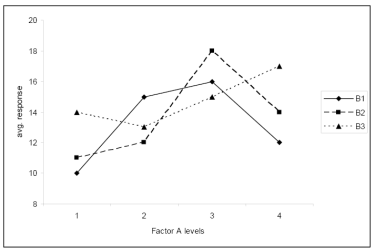
Proposition 12.11 *If the non-interaction hypothesis is true, then the broken lines in the interaction diagram are parallel.*

Proof. The broken line associated with level j joins the points $(1, \hat{m}_{1j})$, $(2, \hat{m}_{2j}), \ldots,$ and $(I, \hat{m}_{Ij})$.

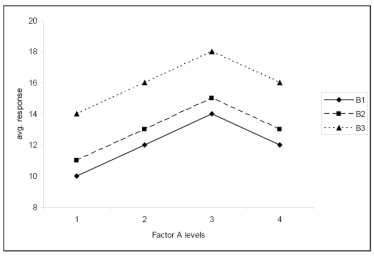
If there is no interaction, then these points have coordinates $(1, \hat{\alpha}_1 + \hat{\beta}_j)$, $(2, \hat{\alpha}_2 + \hat{\beta}_2), \ldots,$ and $(I, \hat{\alpha}_I + \hat{\beta}_j)$ respectively. Therefore, the broken lines associated with levels j and j' correspond by a vertical translation of amplitude $\hat{\beta}_j - \hat{\beta}_{j'}$. □

On this graph, we can read the main effect of the modalities j (the average level of a broken line) and the main effect of the modalities i (the average of the ordinates of the points with fixed abscissa). As far as interactions are concerned, we will rarely obtain strictly parallel broken lines. The problem will then be whether their non-parallelism reflects a significant interaction. Therefore, a test is necessary.

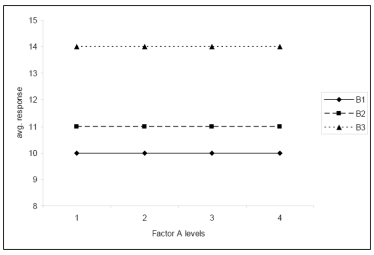
Figure 12.4 illustrates the behavior of the model cell means for different situations. Each line is called a profile, and the crossing of these profiles characterizes the presence of interactions, while parallelism indicates the absence of interactions. It is also possible to detect the presence of an effect of a factor, or to question the relevance of the slicing into the given levels of a factor.



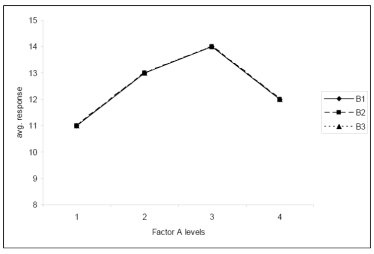
(a) Significant interaction.



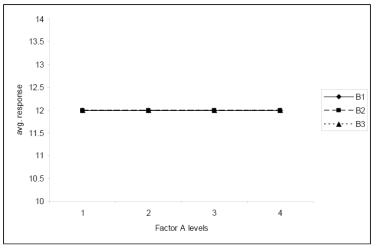
(b) No significant interaction.



(c) No effect of the factor A.



(d) No effect of the factor B.



(e) None of the factors have an effect.

Figure 12.4: *Interaction plot*: Means of the variable Y, for each level of one factor, as a function of the levels of the other.

- Figure 12.4a clearly shows a *significant interaction* between the factors: The change in response when level B changes, depends on level A;
- Figure 12.4b shows *no significant interaction*: The change in response for the level of factor A is the same for each level of factor B;
- Figure 12.4c shows no significant interaction, and that the average response does *not depend on the level of factor A*;
- Figure 12.4d shows no significant interaction, and that the average response does *not depend on the level of factor B*;
- Figure 12.4e illustrates no interaction and neither factor has any effect on the response.

The question is obviously to test whether observed crossings, or groupings, are considered significant.

Beware! A lack of parallelism may also be due to a non-linear relationship between the variable Y and one of the factors.

12.4.2 Fisher Sub-Model Tests

In the following, for the sake of brevity, we will not specify the sets of definitions of the indices i, j and ℓ while setting up the different models. For the record, $i \in [\![1, I]\!]$, $j \in [\![1, J]\!]$ and $\ell \in [\![1, n_{ij}]\!]$. In the following, unless otherwise stated, $\varepsilon_{ij\ell}$ is assumed to be i.i.d. of law $\mathcal{N}(0, \sigma^2)$.

Moreover, *we assume a complete and balanced experimental design*. As a consequence, ℓ is valued in $[\![1, L]\!]$.

12.4.2.1 Non-Interaction Between the two Factors

We want to test

$$\mathcal{H}_0^{AB} : \forall (i, j) \in [\![1, I]\!] \times [\![1, J]\!] , \quad \gamma_{ij} = 0$$

against the alternative

$$\mathcal{H}_1^{AB} : \exists (i, j) \in [\![1, I]\!] \times [\![1, J]\!] , \quad \gamma_{ij} \neq 0 .$$

This amounts to establish whether the additive model

$$(\mathcal{M}_{I+J-1}) : \; Y_{ij\ell} = \mu + \alpha_i + \beta_j + \varepsilon_{ij\ell}$$

is an acceptable sub-model of the complete model with interaction

$$(\mathcal{M}_{IJ}) : \; Y_{ij\ell} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij\ell} .$$

The Fisher statistic for this test is:

$$F_{AB} = \frac{\dfrac{SSAB}{(I-1)(J-1)}}{\dfrac{SSE_{IJ}}{n-IJ}} \overset{H_0}{\sim} \mathcal{F}\big((I-1)(J-1), n-IJ\big) ,$$

and the rejection area is $\mathcal{R}_\delta = \{F_{AB} > f_{1-\delta}\}$, where $f_{1-\delta}$ is the $(1-\delta)$ quantile of the law $\mathcal{F}\big((I-1)(J-1), n-IJ\big)$.

Exercise 12.12 *Prove this result.*

12.4.2.2 No Effect of Factor A

As explained previously, this test is only interesting if the previous test has shown the absence of interaction. We can present this test in two ways, which leads to two different writing of the Fisher test statistic:

1. Consider the complete model $(\mathcal{M}_{IJ})$ and simultaneously test the nullity of the α_i and the γ_{ij};
2. Start by testing the absence of interaction effects. Then, in the additive model $(\mathcal{M}_{I+J-1})$, test only the nullity of the α_i.

These two tests are actually equivalent.

Method #1: Let the complete model $(\mathcal{M}_{IJ})$. We test the hypothesis

$$\mathcal{H}_0^A : \forall i \in [\![1, I]\!],\ \alpha_i = 0 \quad \text{and} \quad \forall (i,j) \in [\![1, I]\!] \times [\![1, J]\!],\ \gamma_{ij} = 0$$

against

$$\mathcal{H}_1^A : \exists i \in [\![1, I]\!],\ \alpha_i \neq 0 \quad \text{or} \quad \exists (i,j) \in [\![1, I]\!] \times [\![1, J]\!],\ \gamma_{ij} \neq 0\,.$$

In other words, we compare the B-factor ANOVA model

$$(\mathcal{M}_J):\ Y_{ij\ell} = \mu + \beta_j + \varepsilon_{ij\ell}$$

to the complete model

$$(\mathcal{M}_{IJ}):\ Y_{ij\ell} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij\ell}\,.$$

The Fisher statistic for this test is:

$$F_A^{\#1} = \frac{\dfrac{SSA + SSAB}{IJ - J}}{\dfrac{SSE_{IJ}}{n - IJ}} \overset{H_0}{\sim} \mathcal{F}\big(IJ - J, n - IJ\big)\,,$$

and the rejection area is $\mathcal{R}_\delta = \{F_A^{\#1} > f_{1-\delta}\}$, where $f_{1-\delta}$ is the $(1-\delta)$ quantile of the law $\mathcal{F}\big(IJ - J, n - IJ\big)$.

Method #2: Let assume that we already established that $(\mathcal{M}_{I+J-1})$ is an acceptable sub-model of $(\mathcal{M}_{IJ})$. By working within $(\mathcal{M}_{I+J-1})$, we now want to test

$$\mathcal{H}_0^A : \forall i \in [\![1, I]\!] \,,\ \alpha_i = 0 \quad \text{against} \quad \mathcal{H}_1^A : \exists i \in [\![1, I]\!] \,,\ \alpha_i \neq 0 \,.$$

In other words, we compare the B-factor ANOVA model

$$(\mathcal{M}_J) \colon\ Y_{ij\ell} = \mu + \beta_j + \varepsilon_{ij\ell}$$

to the additive model

$$(\mathcal{M}_{I+J-1}) \colon\ Y_{ij\ell} = \mu + \alpha_i + \beta_j + \varepsilon_{ij\ell} \,.$$

The Fisher statistic for this test is:

$$F_A^{\#2} \;=\; \frac{\dfrac{SSA}{I-1}}{\dfrac{SSE_{I+J-1}}{n-(I+J-1)}} \;\overset{H_0}{\sim}\; \mathcal{F}\big(I-1, n-(I+J-1)\big) \,,$$

and the rejection area is $\mathcal{R}_\delta = \{F_A^{\#2} > f_{1-\delta}\}$, where $f_{1-\delta}$ is the $(1-\delta)$ quantile of the law $\mathcal{F}\big(I-1, n-(I+J-1)\big)$.

Exercise 12.13 *Prove the previous results.*

12.4.2.3 No Effect of Factor B

The roles played by factors A and B are symmetrical. Hence, the implementation of the test of the effect of factor B on Y is identical to that for factor A. In other words, we can:

- either compare the B-factor ANOVA model

  $$(\mathcal{M}_I) \colon\ Y_{ij\ell} = \mu + \alpha_i + \varepsilon_{ij\ell}$$

  to the complete model

  $$(\mathcal{M}_{IJ}) \colon\ Y_{ij\ell} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij\ell} \,,$$

  which leads to the test statistic

  $$F_B^{\#1} \;=\; \frac{\dfrac{SSB + SSAB}{IJ-I}}{\dfrac{SSE_{IJ}}{n-IJ}} \;\overset{H_0}{\sim}\; \mathcal{F}\big(IJ-I, n-IJ\big) \,,$$

of rejection area is $\mathcal{R}_\delta = \{F_B^{\#1} > f_{1-\delta}\}$, where $f_{1-\delta}$ is the $(1-\delta)$ quantile of the law $\mathcal{F}\big(IJ-I, n-IJ\big)$;

- or compare the B-factor ANOVA model

$$(\mathcal{M}_I)\colon\ Y_{ij\ell} = \mu + \alpha_i + \varepsilon_{ij\ell}$$

to the additive model

$$(\mathcal{M}_{I+J-1})\colon\ Y_{ij\ell} = \mu + \alpha_i + \beta_j + \varepsilon_{ij\ell}\,,$$

which leads to the test statistic

$$F_B^{\#2} = \frac{\dfrac{SSB}{J-1}}{\dfrac{SSE_{IJ}}{n-IJ}} \overset{H_0}{\sim} \mathcal{F}\big(J-1, n-IJ\big)\,,$$

of rejection area is $\mathcal{R}_\delta = \{F_B^{\#2} > f_{1-\delta}\}$, where $f_{1-\delta}$ is the $(1-\delta)$ quantile of the law $\mathcal{F}\big(J-1, n-IJ\big)$.

Exercise 12.14 *Prove the previous results.*

12.4.2.4 No Treatment Effect

The model $(\mathcal{M}_1^\star)$ at the same time sub-model of $(\mathcal{M}_I)$, $(\mathcal{M}_J)$, $(\mathcal{M}_{I+J-1})$ and $(\mathcal{M}_{IJ})$, and $(\mathcal{M}_I)$, $(\mathcal{M}_J)$, $(\mathcal{M}_{I+J-1})$ being themselves sub-models of $(\mathcal{M}_{IJ})$, *etc.* there are various ways to perform this test procedure.

One way is to compare the model $(\mathcal{M}_1^\star)$ with the complete model $(\mathcal{M}_{IJ})$. In this case, we obtain for test statistic

$$F = \frac{\dfrac{SSR_{IJ}}{IJ-1}}{\dfrac{SSE_{IJ}}{n-IJ}} \overset{H_0}{\sim} \mathcal{F}\big(IJ-1, n-IJ\big)\,,$$

and for rejection zone $\mathcal{R}_\delta = \{F > f_{1-\delta}\}$, where $f_{1-\delta}$ is the $(1-\delta)$ quantile of the law $\mathcal{F}\big(IJ-1, n-IJ\big)$.

Remark 12.3 Notice that in F_{AB}, $F_A^{\#1}$, $F_B^{\#1}$ and F, the denominator is always SSE with its degree of freedom $n - IJ$; the numerators change depending on the test. This is true as long as the effects are fixed. That is to say that the levels of our variables are of intrinsic interest in themselves - they are fixed by the experimenter.

Exercise 12.15 *Consider the following R outputs associated with our example. Which model should be used to study these data?*

```
> anov_sing = lm(yield ~ dose * variety, data=wheat)
> anov_add = lm(yield ~ dose + variety, data=wheat)
> anova(anov_add,anov_sing)

Analysis of Variance Table

Model 1: yield ~ dose + variety
Model 2: yield ~ dose * variety
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     14 235.14
2     12 231.47  2    3.6654 0.095   0.91
```

```
> anov_dose = lm(yield ~ dose, data=wheat)
> anova(anov_dose,anov_add)

Analysis of Variance Table

Model 1: yield ~ dose
Model 2: yield ~ dose + variety
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     16 692.72
2     14 235.14  2    457.58 13.622 0.0005192 ***
---
Signif. codes:  0 ‘***‘ 0.001 ‘**‘ 0.01 ‘*‘ 0.05 ‘.‘ 0.1 ‘ ‘ 1
```

```
> anov_variety = lm(yield ~ variety, data=wheat)
> anova(anov_variety,anov_add)

Analysis of Variance Table

Model 1: yield ~ variety
Model 2: yield ~ dose + variety
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     15 249.15
2     14 235.14  1     14.01 0.8341 0.3765
```

12.5 Analysis of Variance Table with two Crossed Factors

In the case of the two-way crossed ANOVA with an orthogonal design, we recall that we can decompose the total variability into

$$SST = SSR + SSE = SSA + SSB + SSAB + SSE\,.$$

The analysis of variance table for an orthogonal design with two crossed factors can thus be drawn up.

Source of variation	Degree of freedom	Sum of Squares	Mean sum of Squares	Test statistics	$f_{1-\delta}$
"Line" A	$I-1$	SSA	$MSA = \dfrac{SSA}{I-1}$	$F_A = \dfrac{MSA}{\hat{\sigma}^2}$	$f_{1-\delta,I-1,n-IJ}$
"Column" B	$J-1$	SSB	$MSB = \dfrac{SSB}{J-1}$	$F_B = \dfrac{MSB}{\hat{\sigma}^2}$	$f_{1-\delta,J-1,n-IJ}$
Interactions	$(I-1)(J-1)$	$SSAB$	$MSAB = \dfrac{SSAB}{(I-1)(J-1)}$	$F_{AB} = \dfrac{MSAB}{\hat{\sigma}^2}$	$f_{1-\delta,(I-1)(J-1),n-IJ}$
Residual	$n-IJ$	SSE	$\hat{\sigma}^2 = \dfrac{SSE}{n-IJ}$		
Total	$n-1$	SST			

Recall that:

$$\begin{aligned}
SST &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{\ell=1}^{n_{ij}}\left(Y_{ij\ell} - \bar{Y}_{\cdots}\right)^2;\\
SSA &= \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\left(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots}\right)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\left(\hat{\alpha}_i\right)^2;\\
SSB &= \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\left(\bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdots}\right)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\left(\hat{\beta}_j\right)^2;\\
SSAB &= \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\left(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdots}\right)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\left(\hat{\gamma}_{ij}\right)^2.
\end{aligned}$$

Analysis of Covariance 13

In this chapter, we will present the analysis of covariance (ANCOVA) model only in the simple framework where we seek to explain a quantitative variable Y as a function of another quantitative variable x, called covariate, and a qualitative variable or factor A. The notions we will study here can be generalized to the case of several covariates operating in a linear or polynomial way, as well as to the case of several factors, possibly with a crossed or hierarchical structure.

In particular, the analysis of covariance falls within the general framework of the linear model. It can be seen as a mixture of analysis of variance (Chapter 11) and linear regression (Chapter 9).

Assume that the factor A has I levels. Each individual in the sample is marked by a double index (i, j):

- The index $i \in [\![1, I]\!]$ represents the level of the factor A to which the individual belongs,
- and j corresponds to the index of the individual in the level i.

For each individual (i, j), we observe a value x_{ij} of the variable x and a value Y_{ij} of the variable Y. For each level $i \in [\![1, I]\!]$ of A, we observe

- n_i values $(x_{i1}, \ldots, x_{in_i})$ of x
- and a n_i-sample $(Y_{i1}, \ldots, Y_{in_i})$ of the random variable Y

Finally, we note $n = \sum_{i=1}^{I} n_i$the total number of observations.

In this chapter, we will illustrate the discussed concepts with the help of a data set listing the weight of oysters as a function of the temperature and oxygenation of their culture medium. More precisely, we have $n = 20$ bags of 10 oysters, and we place, during one month, these 20 bags in a random way in $I = 5$ different locations of a cooling channel of a power plant at a rate of $n_i = 4$ bags per location. These locations differ in temperature and oxygenation. For each bag, we observe:

- `initial_weight`: its weight before the experiment,
- `final_weight`: its weight after the experiment,
- `treatment`: its location encoded from 1 to 5.

The purpose of this study is to know if temperature and oxygenation conditions influence the evolution of oyster weight.

Figure 13.1 displays the data. In order to jointly visualize the effect of the treatment factor and the possible (linear) relationship between the final weight Y and the initial weight x of the oysters, we plot the cloud of points with coordinates (x_{ij}, Y_{ij}), where the same symbol represents all the points of level i. We can also plot a boxplot of each location's initial and final weights (Figure 13.2).

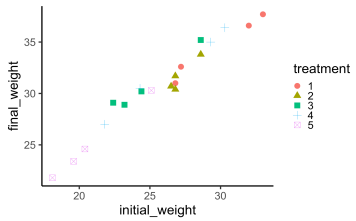


Figure 13.1: Final versus initial weights by location..

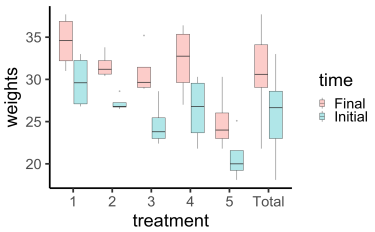


Figure 13.2: Evolution of initial and final weights for each treatment.

We can view ANCOVA from two perspectives.

- The first is a simple linear regression by subgroup. ANCOVA can then allow us to answer the question: "Is the linear relationship between the response Y and the quantitative variable x different between the subgroups, *i.e.*, according to the modalities of the qualitative variable or levels of the factor A? For example, " Does the linear relationship between the weight of oysters after and before the experiment depend on their culture medium?
  When ANCOVA is considered from this linear regression point of view, rather intuitively, one represents the data as straight lines, as in Figure 13.3.
- The second way of conceiving ANCOVA is to compare the predicted means of the response Y, among the subgroups. ANCOVA allows us to compare the adjusted means of each of the groups induced by the levels of the factor A, taking into account/correcting for the variability of the covariate x. In this situation, the ANCOVA enables us to answer the question: "Does the oysters' weight at the end of the experiment depend on their location in the channel once we take their initial weight into account?"
  In this situation, the graphical representation in Figure 13.9 is more appropriate.

13.1 Analysis of Covariance

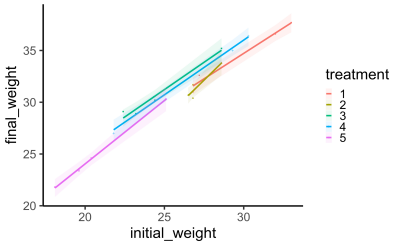


Figure 13.3: *Complete ANCOVA model*. Fitting the regression lines to the data, according to the treatment.

In the simple ANCOVA framework, the regular model writes as:

$$\begin{aligned} \forall i \in [\![1, I]\!]\,, \quad \forall j \in [\![1, n_i]\!], \quad Y_{ij} &= a_i + b_i x_{ij} + \varepsilon_{ij}\,, \\ \text{where} \quad \varepsilon_{ij} &\overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)\,. \end{aligned} \tag{$\mathcal{M}_{2I}^{\star}$}$$

In other words, for each level i of factor A, we estimate a linear regression line of Y on x. At a fixed level i, this line is parameterized by its intercept a_i and by its slope b_i.'

We can write the model in matrix form:

$$\underbrace{\begin{pmatrix} Y_{(1)} \\ \vdots \\ \vdots \\ Y_{(I)} \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} X_{(1)} & & & \\ & X_{(2)} & & \\ & & \ddots & \\ & & & X_{(I)} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} a_1 \\ b_1 \\ \vdots \\ a_I \\ b_I \end{pmatrix}}_{\theta} + \underbrace{\begin{pmatrix} \varepsilon_{(1)} \\ \vdots \\ \vdots \\ \varepsilon_{(I)} \end{pmatrix}}_{\varepsilon},$$

where for all level $i \in [\![1, I]\!]$,

$$Y_{(i)} = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{in_i} \end{pmatrix}, \qquad X_{(i)} = \begin{pmatrix} 1 & x_{i1} \\ \vdots & \vdots \\ 1 & x_{in_i} \end{pmatrix} \qquad \text{and} \qquad \varepsilon_{(i)} = \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{in_i} \end{pmatrix}.$$

In graphic terms, we observe the adjustment of the regression lines in Figure 13.3.

13.1.1 Decomposition of Effects

As with the factorial models, there is a reparameterization that reveals differential effects with respect to a reference level. The model associated with this new parametrization then writes

$$Y_{ij} = \underbrace{\mu + \alpha_i}_{a_i} + \underbrace{(\beta + \gamma_i)}_{b_i} x_{ij} + \varepsilon_{ij}\,, \quad \text{where} \quad \varepsilon_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)\,. \qquad (\mathcal{M}_{2I})$$

This parametrization reveals :

- γ_i: interaction effect between the covariate x and the factor A;
- α_i: differential effect of the factor A on the variable Y;
- β: differential effect of the covariate x on the variable Y.

The model is then over-parameterized and we have to add identification constraints. As for the ANOVA, several choices are possible. The R software imposes to see the first level as reference, *i.e.* it is imposed in R that $\alpha_1 = \gamma_1 = 0$. In practice, we often use the so-called natural constraints: $\sum_{i=1}^{I} n_i \alpha_i = \sum_{i=1}^{I} n_i \beta_i = 0$.

13.1.2 Model without Interaction

Still following the idea of ANOVA, we can define an additive model, i.e. an ANCOVA model without interaction between the covariate x and the factor A. We then obtain the model of equation:

$$Y_{ij} = \mu + \alpha_i + \beta\, x_{ij} + \varepsilon_{ij}\,, \quad \text{where} \quad \varepsilon_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)\,. \qquad (\mathcal{M}_{I+1})$$

As before, this model becomes identifiable under the constraint $\alpha_1 = 0$ (R) $\sum_{i=1}^{I} n_i \alpha_i = 0$ (natural constraint). Hence, its dimension is $I + 2 - 1 = I + 1$.

Graphically, this model consists in fitting parallel lines, with slope β, for each treatment. See Figure 13.4 for an example.

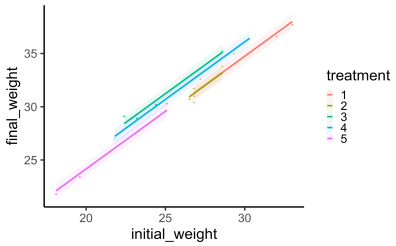


Figure 13.4: Fitting the model without interaction ($\mathcal{M}_{I+1}$) to the data.

Often, one prefers the parametrization,

$$Y_{ij} = \mu + \alpha_i + \beta\,(x_{ij} - \bar{x}_{\cdot\cdot}) + \varepsilon_{ij}\,, \quad \text{where} \quad \varepsilon_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)\,,$$

which makes visible the mean value of the covariate $\bar{x}_{\cdot\cdot} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{J} x_{ij}$.

13.1.3 Model without Effect of the Factor

If the factor A is not involved in the modeling, then the different regression lines will be identical for each level of the factor, *i.e.*

$$a_1 = a_2 = \ldots = a_I \quad \Longleftrightarrow \quad \alpha_1 = \alpha_2 = \ldots = \alpha_I = 0\,,$$

and the model writes:

$$Y_{ij} = \mu + \beta\, x_{ij} + \varepsilon_{ij}\,, \quad \text{where} \quad \varepsilon_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)\,. \qquad (\mathcal{M}_2^\star)$$

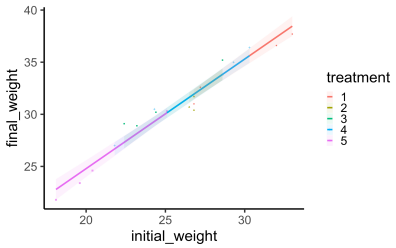


Figure 13.5: Fitting the simple linear regression model ($\mathcal{M}_2^\star$) to the data.

Actually, we set up a simple linear regression model. This situation is illustrated in Figure 13.5

13.1.4 Model without Effect of the Covariate

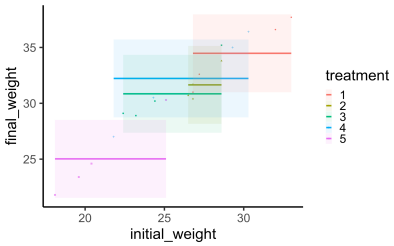


Figure 13.6: Fitting the one-way analysis of variance model ($\mathcal{M}_I$) to the data.

Alternatively, if we want to neglect the covariate x then we write

$$b_1 = b_2 = \ldots = b_I = \beta = 0\,,$$

leading to

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}\,, \quad \text{where} \quad \varepsilon_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)\,. \qquad (\mathcal{M}_I)$$

We actually find a one-factor ANOVA model, whose graphical representation is given in Figure 13.6. A constraint is required to make it identifiable, see Chapter 11 for examples of constraints.

13.1.5 Absence of any Effect

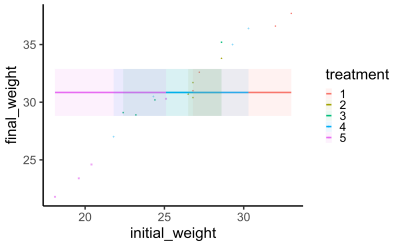


Figure 13.7: Fitting the blank model ($\mathcal{M}_1^\star$) to the data.

Finally, we can consider the sub-model in which neither the factor nor the covariate affects our observations. This amounts to studying the following blank or constant model:

$$Y_{ij} = \mu + \varepsilon_{ij}\,, \quad \text{where} \quad \varepsilon_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)\,. \qquad (\mathcal{M}_1^\star)$$

Graphically, we try to fit a horizontal line to our data, as shown in Figure 13.7.

13.2 Estimation and Forecasting

Let $x_{(i)} = {}^t\left(x_{i1} \ \ldots \ x_{in_i}\right)$, and $\bar{Y}_{i\cdot}$ and $\bar{x}_{i\cdot}$ be the averages defined by:

- $\bar{Y}_{i\cdot} = \dfrac{1}{n_i}\displaystyle\sum_{j=1}^{n_i} Y_{ij}$ the observed mean of level i, and
- $\bar{x}_{i\cdot} = \dfrac{1}{n_i}\displaystyle\sum_{j=1}^{n_i} x_{ij}$ the covariate mean of level i.

13.2.1 Estimation in the Complete Model

Proposition 13.1 (Least squares estimation) *In the model $(\mathcal{M}_{2I}^\star)$, the a_i and b_i are estimated by*

$$\forall i \in [\![1, I]\!]\,, \quad \begin{cases} \hat{a}_i = \bar{Y}_{i\cdot} - \bar{x}_{i\cdot}\hat{b}_i\,, \\ \hat{b}_i = \dfrac{\widehat{Cov}\left(Y_{(i)}, x_{(i)}\right)}{\widehat{Var}\left(x_{(i)}\right)}\,. \end{cases}$$

They are normally distributed.

Exercise 13.2 (Estimation for the regular model) *In the case of the regular model* ($\mathcal{M}_{2I}^{\star}$)*, we can use the general formula* $\hat{\theta} = ({}^tXX)^{-1}\,{}^tXY$*. Using the fact that the matrix* X *is block diagonal,* $X = \mathcal{D}iag\left(X_{(1)}, \ldots, X_{(I)}\right)$*, show that:*

$$\hat{\theta} = \begin{pmatrix} \left({}^tX_{(1)}X_{(1)}\right)^{-1}\,{}^tX_{(1)}Y_{(1)} \\ \vdots \\ \left({}^tX_{(I)}X_{(I)}\right)^{-1}\,{}^tX_{(I)}Y_{(I)} \end{pmatrix}$$

Deduce the estimators of $\hat{a}_i$ *and* $\hat{b}_i$*.*

In R, we get:

```
> ancova.reg = lm(lm(final_weight ~ initial_weight * treatment
    -1, data=oyster))
> summary(ancova.reg)

Call:
lm(formula = lm(final_weight ~ initial_weight * treatment - 1,
data = oyster))

Residuals:
     Min       1Q   Median       3Q      Max
-0.68699 -0.28193  0.02184  0.10425  0.63075

Coefficients:
Estimate Std. Error t value Pr(>|t|)
initial_weight             0.98265    0.09588  10.249 1.27e-06 *
    **
treatment1                 5.24126    2.86473   1.830   0.0972 .
treatment2                -9.14932    8.70021  -1.052   0.3177
treatment3                 4.81796    2.75927   1.746   0.1114
treatment4                 4.29576    2.02339   2.123   0.0597 .
treatment5                -0.43183    2.13283  -0.202   0.8436
initial_weight:treatment2  0.51871    0.33406   1.553   0.1515
initial_weight:treatment3  0.07342    0.14699   0.499   0.6282
initial_weight:treatment4  0.07428    0.12229   0.607   0.5571
initial_weight:treatment5  0.24124    0.13980   1.726   0.1151
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5324 on 10 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9997
F-statistic:  6840 on 10 and 10 DF,  p-value: < 2.2e-16
```

Listing 13.1: Estimation in the regular model ($\mathcal{M}_{2I}^{\star}$)

In the case of the regular model, we can directly build confidence intervals, perform tests, *etc.* In the case of the singular parameterization, one must be more careful.

Assume the default in R constraint $\alpha_1 = \gamma_1 = 0$. Using the link between the parameters in regular model ($\mathcal{M}_{2I}^{\star}$) and the singular one ($\mathcal{M}_{2I}$), we can easily deduce that for all $i \in [\![2, I]\!]$,

$$\begin{cases} \hat{\mu} &= \hat{a}_1 = \bar{Y}_1. - \bar{x}_1. \hat{\beta}\,, \\ \hat{\alpha}_i &= \hat{a}_i - \hat{a}_1 = \bar{Y}_i. - \hat{\mu} - \left(\hat{\gamma}_i + \hat{\beta}\right) \bar{x}_i.\,, \\ \hat{\beta} &= \hat{b}_1 = \dfrac{\widehat{Cov}\left(Y_{(1)}, x_{(1)}\right)}{\widehat{Var}\left(x_{(1)}\right)}\,, \\ \hat{\gamma}_i &= \hat{b}_i - \hat{b}_1 = \dfrac{\widehat{Cov}\left(Y_{(i)}, x_{(i)}\right)}{\widehat{Var}\left(x_{(i)}\right)} - \dfrac{\widehat{Cov}\left(Y_{(1)}, x_{(1)}\right)}{\widehat{Var}\left(x_{((1)}\right)}\,. \end{cases}$$

```
> ancova.sing = lm(lm(final_weight ~ initial_weight * treatment
    -1, data=oyster))
> summary(ancova.sing)

Call:
lm(formula = lm(final_weight ~ initial_weight * treatment, data
    = oyster))

Residuals:
     Min       1Q   Median       3Q      Max
-0.68699 -0.28193  0.02184  0.10425  0.63075

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)                 5.24126    2.86473   1.830   0.0972
    .
initial_weight              0.98265    0.09588  10.249 1.27e-06
    ***
treatment2                -14.39058    9.15971  -1.571   0.1472
treatment3                 -0.42330    3.97747  -0.106   0.9174
treatment4                 -0.94550    3.50725  -0.270   0.7930
treatment5                 -5.67309    3.57150  -1.588   0.1433
initial_weight:treatment2   0.51871    0.33406   1.553   0.1515
initial_weight:treatment3   0.07342    0.14699   0.499   0.6282
initial_weight:treatment4   0.07428    0.12229   0.607   0.5571
initial_weight:treatment5   0.24124    0.13980   1.726   0.1151
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5324 on 10 degrees of freedom
Multiple R-squared:  0.9921,    Adjusted R-squared:  0.985
F-statistic: 139.5 on 9 and 10 DF,  p-value: 2.572e-09
```

Listing 13.2: Estimation in the singular model ($\mathcal{M}_{2I}$)

We predict the value of Y_{ij} using

$$\hat{Y}_{ij} = \hat{a}_i + \hat{b}_i x_{ij} = \hat{\mu} + \hat{\alpha}_i + (\hat{\beta} + \hat{\gamma}_i) x_{ij}\,,$$

which leads to the residuals

$$\varepsilon_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_i. - \hat{b}_i \left(x_{ij} - \bar{x}_i.\right).$$

Proposition 13.3 *The residual variance σ^2 is estimated by*

$$\hat{\sigma}^2 = \frac{\|Y - X\hat{\theta}\|^2}{n - 2I} = \frac{SSE_{2I}}{n - 2I}.$$

Moreover, $\hat{\sigma}^2$ is independent of $\hat{\theta}$ and $(n-2I)\hat{\sigma}^2 \sim \sigma\chi^2(n-2I)$.

In particular, we can define the error sum of squares SSE_{2I} in the full model by

$$\begin{aligned} SSE_{2I} &= \sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - \bar{Y}_{i\cdot} - \hat{b}_i(x_{ij} - \bar{x}_{i\cdot})\right)^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2 - \sum_{i=1}^{I}\sum_{j=1}^{J}\hat{b}_i^2\left(x_{ij} - \bar{x}_{i\cdot}\right)^2. \end{aligned}$$

13.2.2 Estimation in the Sub-Models

The only model we have not yet studied is the $(\mathcal{M}_{I+1})$ ancova model with effect of covariate, factor, but no intercation between the two.

13.2.2.1 Model without Interaction $(\mathcal{M}_{I+1})$

Recall that for all $i \in [\![1, I]\!]$ and $j \in [\![1, n_i]\!]$,

$$(\mathcal{M}_{I+1}):\ Y_{ij} = \mu + \alpha_i + \beta\, x_{ij} + \varepsilon_{ij}.$$

Proceeding in the same way as before, we can show that

$$\forall i \in [\![1, I]\!],\quad \begin{cases} \hat{a}_i = \hat{\mu} + \hat{\alpha}_i = \bar{Y}_{i\cdot} - \bar{x}_{i\cdot}\hat{b}, \\ \hat{b} = \hat{\beta} = \dfrac{\sum_{i=1}^{I} n_i \widehat{Cov}\left(Y_{(i)}, x_{(i)}\right)}{\sum_{i=1}^{n_i} n_i \widehat{Var}\left(x_{(i)}\right)}. \end{cases}$$

In particular, the residual variance is now estimated by

$$\hat{\sigma}^2 = \frac{SSE_{I+1}}{n - I - 1} \sim \frac{\sigma}{n - I + 1}\chi^2(n - I - 1),$$

where

$$SSE_{I+1} = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2 - \hat{\beta}^2\sum_{i=1}^{I}\sum_{j=1}^{J}\left(x_{ij} - \bar{x}_{i\cdot}\right)^2.$$

13.2.2.2 Model without Effect of the Factor

The model with no effect of the covariate $(\mathcal{M}_2^\star)$ amounts to a linear regression: For all $i \in [\![1, I]\!]$ and $j \in [\![1, n_i]\!]$,

$$(\mathcal{M}_2^\star):\ Y_{ij} = \mu + \beta\, x_{ij} + \varepsilon_{ij} + \varepsilon_{ij}.$$

We thus obtain:

$$\hat{\mu} = \bar{Y}_{\cdot\cdot} - \hat{\beta}\,\bar{x}_{\cdot\cdot} \qquad \text{and} \qquad \hat{\beta} = \frac{\widehat{Cov}(Y,x)}{\widehat{Var}(x)}\,.$$

The residual variance is estimated by

$$\hat{\sigma}^2 \;=\; \frac{SSE_2}{n-2} \;\sim\; \frac{\sigma}{n-2}\,\chi^2(n-2)\,,$$

where

$$SSE_2 \;=\; \sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - \bar{Y}_{\cdot\cdot}\right)^2 - \hat{\beta}^2 \sum_{i=1}^{I}\sum_{j=1}^{J}\left(x_{ij} - \bar{x}_{\cdot\cdot}\right)^2\,.$$

13.2.2.3 Model without Effect of the Covariate

The model with no effect of the covariate ($\mathcal{M}_I$) amounts to a one-way ANOVA model: For all $i \in [\![1, I]\!]$ and $j \in [\![1, n_i]\!]$,

$$(\mathcal{M}_I)\colon\; Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}\,.$$

Hence, for all $i \in [\![2, I]\!]$

$$\hat{\mu} \;=\; \bar{Y}_{1\cdot} \qquad \text{and} \qquad \hat{\alpha}_i \;=\; \bar{Y}_{i\cdot} - \bar{Y}_{1\cdot}\,,$$

under the R constraint $\alpha_1 = 0$.

The estimator of the variance is given by

$$\hat{\sigma}^2 \;=\; \frac{SSE_I}{n-I} \;=\; \frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2 \;\sim\; \frac{\sigma}{n-I}\,\chi^2(n-I)\,.$$

13.2.2.4 Absence of any Effect

The model writes

$$(\mathcal{M}_1^\star)\colon\; Y_{ij} = \mu + \varepsilon_{ij}\,,$$

and we have the following estimations $\hat{\mu} = \bar{Y}_{\cdot\cdot}$ and

$$\hat{\sigma}^2 = \frac{SSE_1}{n-1} = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - \bar{Y}_{\cdot\cdot}\right)^2 \;\sim\; \frac{\sigma}{n-1}\,\chi^2(n-1)\,.$$

13.3 Effect Test

We now have all the ingredients to test the relevance of the different variables in the proposed ANCOVA model. For each of the tests, we report the result we obtained using the R software.

Remark 13.1 As with the ANOVA model, if there is an interaction effect between the factor A and the covariate x, then their individual effects must be included in the model.

13.3.1 Non-Interaction Between the Covariate and the Factor

We want to establish whether the additive model without interaction

$$(\mathcal{M}_{I+1}):\ Y_{ij} = \mu + \alpha_i + \beta\, x_{ij} + \varepsilon_{ij}$$

is an acceptable sub-model of the complete model with interaction

$$(\mathcal{M}_{2I}):\ Y_{ij} = \mu + \alpha_i + (\beta + \gamma_i)\, x_{ij} + \varepsilon_{ij}\,.$$

The Fisher statistic for this test is:

$$F = \frac{\dfrac{SSE_{I+1} - SSE_{2I}}{I+1}}{\dfrac{SSE_{2I}}{n-2I}} \overset{H_0}{\sim} \mathcal{F}(I+1, n-2I)\,,$$

and the rejection area is $\mathcal{R}_\delta = \{F > f_{1-\delta}\}$, where $f_{1-\delta}$ is the $(1-\delta)$ quantile of the law $\mathcal{F}(I+1, n-2I)$.

Due to the graphical representation (Figure 13.4), we sometime refers to this test as the test of homogeneity of the regression slopes.

```
> ancova.indep = lm(lm(final_weight ~ initial_weight+treatment,
                                                    data=oyster))
> anova(ancova.indep,ancova.sing)

Analysis of Variance Table

Model 1: final_weight ~ initial_weight + treatment
Model 2: final_weight ~ initial_weight * treatment
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     14 4.2223
2     10 2.8340  4    1.3883 1.2247 0.3602
```

13.3.2 No Effect of Factor A

We want to test if the linear regression model

$$(\mathcal{M}_2^\star):\ Y_{ij} = \mu + \beta\, x_{ij} + \varepsilon_{ij} + \varepsilon_{ij}$$

is an acceptable sub-model of the additive model without interaction

$$(\mathcal{M}_{I+1}):\ Y_{ij} = \mu + \alpha_i + \beta\, x_{ij} + \varepsilon_{ij}\,.$$

The Fisher statistic for this test is:

$$F = \frac{\frac{SSE_2 - SSE_{I+1}}{I-1}}{\frac{SSE_{I+1}}{n-I-1}} \overset{H_0}{\sim} \mathcal{F}(I-1, n-I-1),$$

and the rejection area is $\mathcal{R}_\delta = \{F > f_{1-\delta}\}$, where $f_{1-\delta}$ is the $(1-\delta)$ quantile of the law $\mathcal{F}(I-1, n-I-1)$.

```
> reglin = lm(lm(final_weight ~ initial_weight, data=oyster))
> anova(reglin,ancova.indep)

Analysis of Variance Table

Model 1: final_weight ~ initial_weight
Model 2: final_weight ~ initial_weight + treatment
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     18 16.3117
2     14  4.2223  4    12.089 10.021 0.0004819 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

13.3.3 No Effect of Covariate x

We want to test if the one-way anova model

$$(\mathcal{M}_I): \; Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

is an acceptable sub-model of the additive model without interaction

$$(\mathcal{M}_{I+1}): \; Y_{ij} = \mu + \alpha_i + \beta\, x_{ij} + \varepsilon_{ij}\,.$$

The Fisher statistic for this test is:

$$F = \frac{\frac{SSE_I - SSE_{I+1}}{}}{\frac{SSE_{I+1}}{n-I-1}} \overset{H_0}{\sim} \mathcal{F}(1, n-I-1),$$

and the rejection area is $\mathcal{R}_\delta = \{F > f_{1-\delta}\}$, where $f_{1-\delta}$ is the $(1-\delta)$ quantile of the law $\mathcal{F}(1, n-I-1)$.

```
> anova = lm(lm(final_weight ~ treatment, data=oyster))
> anova(anova,ancova.indep)

Analysis of Variance Table

Model 1: final_weight ~ treatment
Model 2: final_weight ~ initial_weight + treatment
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     15 160.263
2     14   4.222  1    156.04 517.38 1.867e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Remark 13.2 We could also perform a Student's nullity test of the β parameter in $(\mathcal{M}_{I+1})$.

13.3.3.1 Absence of any Effect

We want to test if the blank model

$$(\mathcal{M}_1^\star):\ Y_{ij} = \mu + \varepsilon_{ij}$$

is an acceptable sub-model of the complete model[1]

$$(\mathcal{M}_{2I}):\ Y_{ij} = \mu + \alpha_i + (\beta + \gamma_i)\, x_{ij} + \varepsilon_{ij}\,.$$

1: Or of any of the model $(\mathcal{M}_{I+1})$, $(\mathcal{M}_I)$ of $(\mathcal{M}_{2I}^\star)$, but be careful to change the degrees of freedom in the Fisher statistic accordingly.

The Fisher statistic for this test is:

$$F = \frac{\dfrac{SSE_1 - SSE_{2I}}{2I-1}}{\dfrac{SSE_{2I}}{n-2I}} = \frac{\dfrac{SSR_{2I}}{2I-1}}{\dfrac{SSE_{2I}}{n-2I}} \overset{H_0}{\sim} \mathcal{F}(2I-1, n-2I)\,,$$

and the rejection area is $\mathcal{R}_\delta = \{F > f_{1-\delta}\}$, where $f_{1-\delta}$ is the $(1-\delta)$ quantile of the law $\mathcal{F}(2I-1, n-2I)$.

```
> ancova.cst = lm(lm(final_weight ~ 1, data=oyster))
> anova(ancova.cst,ancova.sing)

Analysis of Variance Table

Model 1: final_weight ~ 1
Model 2: final_weight ~ initial_weight * treatment
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     19 358.67
2     10   2.83  9    355.84 139.51 2.572e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exercise 13.4 *Using R output, which submodel is best suited to study oyster culture data.*

Remark 13.3 Like ANOVA, ANCOVA assumes equality of variances for all groups. To conduct a rigorous study, we would have to check this.

13.3.4 Comparison of Groups: Raw vs. Adjusted Means

In the ANCOVA model, the mean of group i is given by $\bar{\mu}_{i\cdot}$:[2]

$$\forall i \in [\![1, I]\!], \quad \bar{\mu}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbb{E}[Y_{ij}] = \mu + \alpha_i + \frac{(\beta+\gamma_i)}{n_i}\sum_{j=1}^{n_i} x_{ij}$$
$$= \mu + \alpha_i + (\beta+\gamma_i)\,\bar{x}_{i\cdot}.$$

We note $\hat{\mu}_{i\cdot}$ its estimate, obtained from the estimates of the different parameters of the model: $\hat{\mu}_{i\cdot} = \hat{\mu} + \hat{\alpha}_i + (\hat{\beta}+\hat{\gamma}_i)\bar{x}_{i\cdot}$. For each group i, it corresponds to the response prediction when the covariate x is equal to its mean $\bar{x}_{i\cdot}$ for this group.

By comparing these different means $\bar{\mu}_{i\cdot}$, we can compare the average behavior of the different groups i. However, with this definition of average behavior by group, a significant difference between these responses may be the consequence of a significant difference in $\bar{x}_{i\cdot}$ abscissae. In particular, if this difference between abscissae is the result of poor sampling, then using these means to compare groups is irrelevant. One way to avoid this is to compare the difference in responses between groups obtained for the same abscissa. A natural choice is the grand mean of the covariate $\bar{x}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} x_{ij}$. This is referred to as the *adjusted mean*, in the sense that it adjusts for, or is corrected by, the possible effect of the covariate x:

$$\tilde{\mu}_{i\cdot} = \mu + \alpha_i + (\beta+\gamma_i)\,\bar{x}_{\cdot\cdot}.$$

Figure 13.8 illustrates the difference between classical and adjusted means.

Finally, in Figure 13.9, we represent the marginal means estimated by the model, namely the quantities $\hat{\mu} + \hat{\alpha}_i + (\hat{\beta}+\hat{\gamma}_i)\,\bar{x}_{\cdot\cdot}$ for each level, and the observations associated with this treatment.

2: Since the noise ε has a null mean, $\bar{\mu}_{i\cdot} = \bar{Y}_{i\cdot}$.

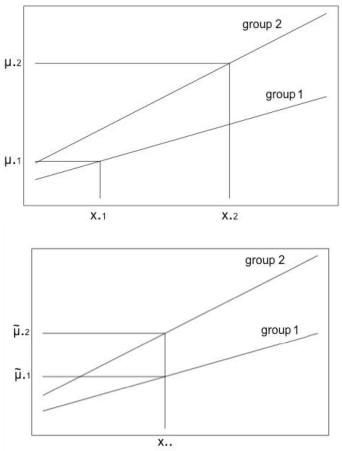


Figure 13.8: Classical (top) and adjusted (bottom) means for $I = 2$ groups.

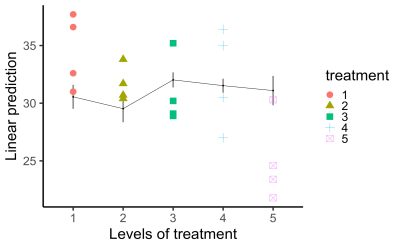


Figure 13.9: Estimated marginal means.

Generalized Linear Model

Appendix

Quantiles Tables and Summary Sheet

A

A.1 Summary Sheet on Non-Parametric Tests

For each test, we specify in brackets the pages where it was defined.

- *Kolmogorov* test (p. 26):

$$D_n = \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F_0(x)| ;$$

- *Kolmogorov-Smirnov* test (p. 29):

$$D_{(n,m)} = \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - \hat{G}_m(x)| ;$$

- *Mann-Whitney* test (p. 31):

$$U^{X<Y}_{(n,m)} = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{1}_{X_i<Y_j} ;$$

- *Wilcoxon* test (p. 33):

$$W^{Y}_{(n,m)} = \sum_{j=1}^{m} R_j ;$$

- *Median* test (p. 35):

$$M_{(n,m)} = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}_{R_j>\frac{N+1}{2}} ;$$

- *Kolmogorov-Smirnov normality* test (p. 39):

$$D_n = \sup_{x\in\mathbb{R}} \left| \hat{F}_n(x) - \Phi\left(x; \bar{X}, S^2_X\right) \right| ;$$

- *Shapiro-Wilk* test (p. 42):

$$W_n = \frac{\hat{\sigma}_n\left({}^t\alpha B^{-1}\alpha\right)^2}{\sum_{i=1}^{n}(x_i-\bar{X})^2\left({}^t\alpha B^{-2}\alpha\right)} ;$$

- *Chi-squared* tests:
  - *Pearson's fit* test (p. 45):

$$T_n = \sum_{k=1}^{K} \frac{(N_k - np^0_k)^2}{np^0_k} ;$$

- *Goodness-of-fit* test (p. 47):

$$\widehat{T}_n = \sum_{k=1}^{K} \frac{\left(N_k - np_k(\hat{\theta})\right)^2}{np_k(\hat{\theta})};$$

- *Independence* test (p. 49):

$$I_n = \sum_{k=1}^{K}\sum_{\ell=1}^{L} \frac{\left(N_{k,\ell} - \frac{N_{k,\cdot}N_{\cdot,\ell}}{n}\right)^2}{\frac{N_{k,\cdot}N_{\cdot,\ell}}{n}};$$

- *Homogeneity* test (p. 51):

$$J_n = \sum_{k=1}^{K}\sum_{\ell=1}^{L} \frac{\left(N_{k,\ell} - \frac{N_{k,\cdot}N_{\cdot,\ell}}{n}\right)^2}{\frac{N_{k,\cdot}N_{\cdot,\ell}}{n}}.$$

A.2 Quantiles Tables

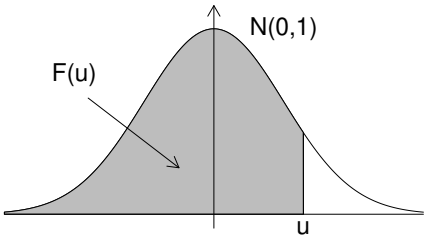
Afterwards several usual quantiles tables are displayed.

Loi normale centrée réduite

Table de la fonction de répartition de $U \sim \mathcal{N}(0,1)$:

$$\mathbb{P}(U \leq u) = F(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx,$$

avec $F(-u) = 1 - F(u)$ et $\mathbb{P}(|U| \leq u) = 2F(u) - 1$.



u	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

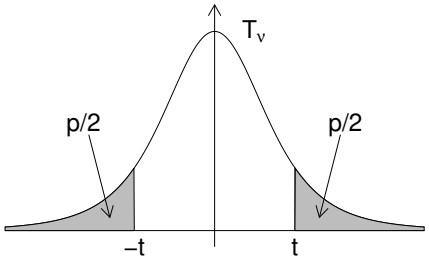
Loi de Student

Table de t en fonction du degré de liberté ν et de la probabilité p tels que :

$$\mathbb{P}\left(|T_\nu| > t\right) = p,$$

avec

$$T_\nu = \frac{U}{\sqrt{Y/\nu}} \quad \text{où} \quad U \sim \mathcal{N}(0,1) \perp\!\!\!\perp Y \sim \chi^2(\nu).$$



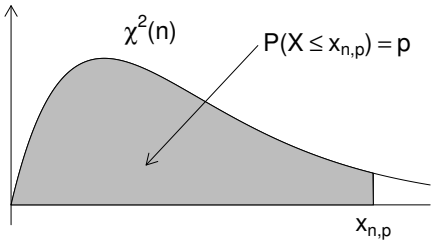
ν \ p	0.9	0.7	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01
1	0.158	0.510	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	0.142	0.445	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	0.137	0.424	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841
4	0.134	0.414	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604
5	0.132	0.408	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032
6	0.131	0.404	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707
7	0.130	0.402	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499
8	0.130	0.399	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355
9	0.129	0.398	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250
10	0.129	0.397	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169
11	0.129	0.396	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106
12	0.128	0.395	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055
13	0.128	0.394	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012
14	0.128	0.393	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977
15	0.128	0.393	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947
16	0.128	0.392	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921
17	0.128	0.392	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898
18	0.127	0.392	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878
19	0.127	0.391	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861
20	0.127	0.391	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845
21	0.127	0.391	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831
22	0.127	0.390	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819
23	0.127	0.390	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807
24	0.127	0.390	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797
25	0.127	0.390	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787
26	0.127	0.390	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779
27	0.127	0.389	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771
28	0.127	0.389	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763
29	0.127	0.389	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756
30	0.127	0.389	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750
$+\infty$	0.12566	0.38532	0.67449	0.84162	1.03643	1.28155	1.64485	1.95996	2.32635	2.57583

La loi limite, lorsque ν tend vers l'infini, est une loi normale centrée réduite.

Loi du khi-deux

Table des quantiles de $X \sim \chi^2(n)$:

$$\mathbb{P}\left(X \leq x_{n,p}\right) = p.$$



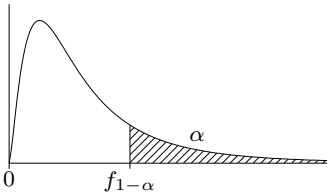
n \ p	0.005	0.01	0.025	0.05	0.1	0.25	0.5	0.75	0.9	0.95	0.975	0.99	0.995
1	0.00	0.00	0.00	0.00	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34	45.62	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	52.29	59.33	66.98	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	61.70	69.33	77.58	85.53	90.53	95.02	100.4	104.2
80	51.17	53.54	57.15	60.39	64.28	71.14	79.33	88.13	96.58	101.9	106.6	112.3	116.3
90	59.20	61.75	65.65	69.13	73.29	80.62	89.33	98.65	107.6	113.1	118.1	124.1	128.3
100	67.33	70.06	74.22	77.93	82.36	90.13	99.33	109.1	118.5	124.3	129.6	135.8	140.1

Loi de Fisher

Si F est une variable aléatoire suivant la loi de Fisher–Snedecor à (ν_1, ν_2) degrés de liberté, la table donne la valeur $f_{1-\alpha}$ telle que

$$\mathbb{P}\{F \geqslant f_{1-\alpha}\} = \alpha = 0{,}05.$$

Ainsi, $f_{1-\alpha}$ est le quantile d'ordre $1 - \alpha = 0{,}95$ de la loi de Fisher–Snedecor à (ν_1, ν_2) degrés de liberté.



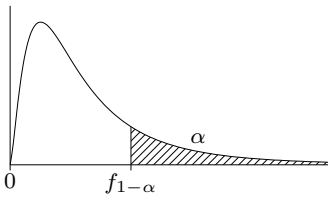
ν_2 \ ν_1	1	2	3	4	5	6	8	10	15	20	30	∞
1	161	200	216	225	230	234	239	242	246	248	250	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,85	8,79	8,70	8,66	8,62	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,96	5,86	5,80	5,75	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,74	4,62	4,56	4,50	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,06	3,94	3,87	3,81	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,64	3,51	3,44	3,38	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,35	3,22	3,15	3,08	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,14	3,01	2,94	2,86	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,98	2,85	2,77	2,70	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,85	2,72	2,65	2,57	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,85	2,75	2,62	2,54	2,47	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,77	2,67	2,53	2,46	2,38	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,60	2,46	2,39	2,31	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,54	2,40	2,33	2,25	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,49	2,35	2,28	2,19	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,45	2,31	2,23	2,15	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,41	2,27	2,19	2,11	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,38	2,23	2,16	2,07	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,35	2,20	2,12	2,04	1,84
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,30	2,15	2,07	1,98	1,78
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,25	2,11	2,03	1,94	1,73
26	4,23	3,37	2,98	2,74	2,59	2,47	2,32	2,22	2,07	1,99	1,90	1,69
28	4,20	3,34	2,95	2,71	2,56	2,45	2,29	2,19	2,04	1,96	1,87	1,65
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,16	2,01	1,93	1,84	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,08	1,92	1,84	1,74	1,51
50	4,03	3,18	2,79	2,56	2,40	2,29	2,13	2,03	1,87	1,78	1,69	1,44
60	4,00	3,15	2,76	2,53	2,37	2,25	2,10	1,99	1,84	1,75	1,65	1,39
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,95	1,79	1,70	1,60	1,32
100	3,94	3,09	2,70	2,46	2,31	2,19	2,03	1,93	1,77	1,68	1,57	1,28
∞	3,84	3,00	2,60	2,37	2,21	2,10	1,94	1,83	1,67	1,57	1,46	1,00

Loi de Fisher (suite)

Si F est une variable aléatoire suivant la loi de Fisher–Snedecor à (ν_1, ν_2) degrés de liberté, la table donne la valeur $f_{1-\alpha}$ telle que

$$\mathbb{P}\{F \geqslant f_{1-\alpha}\} = \alpha = 0,025.$$

Ainsi, $f_{1-\alpha}$ est le quantile d'ordre $1 - \alpha = 0{,}975$ de la loi de Fisher–Snedecor à (ν_1, ν_2) degrés de liberté.



ν_2 \ ν_1	1	2	3	4	5	6	8	10	15	20	30	∞
1	648	800	864	900	922	937	957	969	985	993	1 001	1 018
2	38,5	39,0	39,2	39,2	39,3	39,3	39,4	39,4	39,4	39,4	39,5	39,5
3	17,4	16,0	15,4	15,1	14,9	14,7	14,5	14,4	14,3	14,2	14,1	13,9
4	12,2	10,6	9,98	9,60	9,36	9,20	8,98	8,84	8,66	8,56	8,46	8,26
5	10,0	8,43	7,76	7,39	7,15	6,98	6,76	6,62	6,43	6,33	6,23	6,02
6	8,81	7,26	6,60	6,23	5,99	5,82	5,60	5,46	5,27	5,17	5,07	4,85
7	8,07	6,54	5,89	5,52	5,29	5,12	4,90	4,76	4,57	4,47	4,36	4,14
8	7,57	6,06	5,42	5,05	4,82	4,65	4,43	4,30	4,10	4,00	3,89	3,67
9	7,21	5,71	5,08	4,72	4,48	4,32	4,10	3,96	3,77	3,67	3,56	3,33
10	6,94	5,46	4,83	4,47	4,24	4,07	3,85	3,72	3,52	3,42	3,31	3,08
11	6,72	5,26	4,63	4,28	4,04	3,88	3,66	3,53	3,33	3,23	3,12	2,88
12	6,55	5,10	4,47	4,12	3,89	3,73	3,51	3,37	3,18	3,07	2,96	2,72
13	6,41	4,97	4,35	4,00	3,77	3,60	3,39	3,25	3,05	2,95	2,84	2,60
14	6,30	4,86	4,24	3,89	3,66	3,50	3,29	3,15	2,95	2,84	2,73	2,49
15	6,20	4,76	4,15	3,80	3,58	3,41	3,20	3,06	2,86	2,76	2,64	2,40
16	6,12	4,69	4,08	3,73	3,50	3,34	3,12	2,99	2,79	2,68	2,57	2,32
17	6,04	4,62	4,01	3,66	3,44	3,28	3,06	2,92	2,72	2,62	2,50	2,25
18	5,98	4,56	3,95	3,61	3,38	3,22	3,01	2,87	2,67	2,56	2,44	2,19
19	5,92	4,51	3,90	3,56	3,33	3,17	2,96	2,82	2,62	2,51	2,39	2,13
20	5,87	4,46	3,86	3,51	3,29	3,13	2,91	2,77	2,57	2,46	2,35	2,09
22	5,79	4,38	3,78	3,44	3,22	3,05	2,84	2,70	2,50	2,39	2,27	2,00
24	5,72	4,32	3,72	3,38	3,15	2,99	2,78	2,64	2,44	2,33	2,21	1,94
26	5,66	4,27	3,67	3,33	3,10	2,94	2,73	2,59	2,39	2,28	2,16	1,88
28	5,61	4,22	3,63	3,29	3,06	2,90	2,69	2,55	2,34	2,23	2,11	1,83
30	5,57	4,18	3,59	3,25	3,03	2,87	2,65	2,51	2,31	2,20	2,07	1,79
40	5,42	4,05	3,46	3,13	2,90	2,74	2,53	2,39	2,18	2,07	1,94	1,64
50	5,34	3,98	3,39	3,06	2,83	2,67	2,46	2,32	2,11	1,99	1,87	1,55
60	5,29	3,93	3,34	3,01	2,79	2,63	2,41	2,27	2,06	1,94	1,82	1,48
80	5,22	3,86	3,28	2,95	2,73	2,57	2,36	2,21	2,00	1,88	1,75	1,40
100	5,18	3,83	3,25	2,92	2,70	2,54	2,32	2,18	1,97	1,85	1,71	1,35
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,19	2,05	1,83	1,71	1,57	1,00

References (mostly in French)

[1] François Husson, Sébastien Lê, and Jérôme Pagès. *Analyse de données avec R*. Presses universitaires de Rennes, 2016 (cited on page 142).

[2] Philippe Capéraà and Bernard Van Cutsem. *Méthodes et modèles en statistique non paramétrique: Exposé fondamental*. Vol. 1. Presses Université Laval, 1988 (cited on page 31).

[3] Françoise Couty, Jean Debord, and Daniel Fredon. *Probabilités et Statistiques*. Dunod, 1999.

[4] Jean-Jacques Daudin, Stéphane Robin, and Colette Vuillet. *Statistique inférentielle ; Idées, démarches, exemples*. 2001.

[5] Alain Monfort. *Cours de statistique mathématique*. Économica, 1982.

[6] Tassi Philippe. *Méthodes statistiques*. fre. Économica, 1985.

[7] Christophe Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2021 (cited on page 101).

[8] Colin L Mallows. 'Some comments on C_p'. In: *Technometrics* 42.1 (2000), pp. 87–94 (cited on pages 104, 122).

[9] Gilbert Saporta. 'Probabilités, Statistique et Analyse des données'. In: *Éditions Technip* 488 (1990).

[10] Hirotugu Akaike. 'A Bayesian analysis of the minimum AIC procedure'. In: *Selected Papers of Hirotugu Akaike*. Springer, 1998, pp. 275–280 (cited on page 105).

[11] Hirotogu Akaike. 'Information theory and an extension of the maximum likelihood principle'. In: *Selected papers of hirotugu akaike*. Springer, 1998, pp. 199–213 (cited on page 105).

[12] Gideon Schwarz. 'Estimating the dimension of a model'. In: *The annals of statistics* (1978), pp. 461–464 (cited on page 106).

[13] Jean-Marc Azaïs and Jean-Marc Bardet. *Le modèle linéaire par l'exemple : Régression, Analyse de la variance et Plans d'expériences. Illustrations numériques avec les logiciels R, SAS et Splus*. 2006 (cited on page 105).

[14] Arthur E Hoerl, Robert W Kannard, and Kent F Baldwin. 'Ridge regression: some simulations'. In: *Communications in Statistics-Theory and Methods* 4.2 (1975), pp. 105–123 (cited on page 122).

[15] Arthur E Hoerl and Robert W Kennard. 'Ridge regression iterative estimation of the biasing parameter'. In: *Communications in Statistics-Theory and Methods* 5.1 (1976), pp. 77–88 (cited on page 122).

[16] Gary C McDonald and Diane I Galarneau. 'A Monte Carlo evaluation of some ridge-type estimators'. In: *Journal of the American Statistical Association* 70.350 (1975), pp. 407–416 (cited on page 122).

[17] Robert Tibshirani. 'Regression shrinkage and selection via the lasso'. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288 (cited on page 123).

[18] Norman R Draper and Harry Smith. *Applied regression analysis*. Vol. 326. John Wiley & Sons, 1998 (cited on page 113).

[19] X Guyon. 'Modele linéaire et économétrie'. In: *Ellipse, Paris* (2001) (cited on page 113).

[20] J Koerts. *Generalized linear models: Monographs on Statistics and Applied Probability, Chapman and Hall, London, 1983, xiii+ 261 pages,£ 15.00*. 1984 (cited on page 113).

[21] Amemiya Takeshi. *Advanced econometrics*. Harvard university press, 1985 (cited on page 113).

[22] William H Greene. 'Econometric analysis 4th edition'. In: *International edition, New Jersey: Prentice Hall* (2000), pp. 201–215 (cited on page 113).

[23] J Dave Jobson. *Applied multivariate data analysis: regression and experimental design*. Springer Science & Business Media, 2012 (cited on page 113).