Mathematics for Purchasing
Msc Supply Chain & Purchasing

Guillaume Metzler

Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France

guillaume.metzler@univ-lyon2.fr

Contents

Part I

Introduction

The aim of linear (or non-linear) modeling is to describe phenomena using an equation linking random variables. More precisely, it seeks to predict or explain the values of a random variable Y using several explanatory variables $X_1, X_2, \ldots, X_p$.

To establish this link, we rely on observations to estimate the parameters of the model describing the phenomenon. However, the process of collecting or processing data is subject to error, which can induce a bias in the learned model. This error is often modeled by a random variable ε, the nature of which will depend on the type of model considered.

Finally, modelling will consist of determining the unknown function f which will link the variable to be explained Y to the explanatory variables $X_1, \ldots, X_p$, taking into account any noise (our error) in the data, *i.e.*,

$$Y = f(\mathbf{X}) + \varepsilon,$$

where $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ and $f$is the function we're trying to determine, which will depend on parameters.

This modeling work is often accompanied by a data exploration phase.

Statistics

Exploration + Modeling $\longrightarrow$ *Data Mining*

Some problems

The nature of the modeling changes according to the nature of the Y :

- if Y is qualitative, this is called a **classification** problem.
- if Y is quantitative, it's called a **regression** problem.

These are classic modeling contexts. There is a final case, not dealt with here, in which we have no Y variable, but only explanatory variables, and we want to construct groups. This is known as **clustering**, a method often applied in high-dimensional statistics.

Choose the model

There are several ways of estimating the f function, which can lead to different estimates. The latter may also depend on the amount of information used, *i.e.* the number of explanatory variables employed.
The aim of the regression problems we'll be studying is to strike a balance between

- **a large number of explanatory variables**: this will enable the model to explain the data better, but with weaker predictive power, i.e. a higher risk of poor predictions.
- **few explanatory variables:** the model will have low variance, and therefore potentially weaker predictions. On the other hand, it will have greater difficulty in explaining the data.

This notion will later refer to the **bias - variance trade-off of the model**, a very important notion that you will find in a statistical learning context and which will be linked to the notion of **complexity of the model**. The latter is closely linked to the amount of information, and therefore variables, used.

Model Selection Technique

These are often divided into two categories

Variable Selection

It is based on statistical criteria to measure the quality of a model, taking into account the number of parameters used. Statistical tests are then used to determine whether the difference in results is significant or not.

Regularization or Penalization

A process often used in high-dimensional statistics to automatically select the most relevant variables. This is done by adding so-called **penalities** terms to the problem we are trying to solve.

Studied model.

We will be looking at several models in this course. The simplest is the linear Gaussian model in the form

$$Y = f(\mathbf{X}) + \varepsilon,$$

where we will assume that $\mathbf{X}$ are deterministic and that $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, where σ^2 denotes the variance or noise present in our data.

The function f will then have a very specific form in this context. We will take an affine function of the form

$$f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p = \sum_{j=0}^{p} \beta_j x_j.$$

Using logistic regression, we will also see how to deal with a **classification** problem using a so-called **regression** model. In this context, the variable we are seeking to explain.

In the following class, *i.e.*, in the data science class, we we will go a little bit further and try to study more general models used in Machine Learning for prediction tasks.

Part II

Gaussian Linear Models

We now place ourselves in a well-defined framework, where we seek to explain the values taken by a quantitative random variable Y as a function of the values taken by a set of quantitative or qualitative variables $X_1, X_2, \dots, X_p$ such that

$$Y = f(\mathbf{X}) + \varepsilon,$$

where $\mathbf{X} = (X_1, X_2, \dots, X_p)$ and where, hence the name Gaussian, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ where σ^2 is unknown.
We will also assume that the function f considered is a **linear** function, *i.e.*, our model can be written as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_p X_p + \varepsilon.$$

The values of the vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)$ are then to be determined.

To do this, we have a dataset that will allow us to obtain an estimate of these parameters using a criterion that we will define and seek to minimize. This will take the form of an optimization problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \varphi(\boldsymbol{\beta}).$$

We will seek to solve this problem with a sample of data $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ and $\mathbf{x}_i \in \mathbb{R}^p$ such that

$$\begin{array}{rcl}
y_1 & = & \beta_0 + \beta_1 X_{1,1} + \beta_2 X_{1,2} + \dots + \beta_p X_{1,p} + \varepsilon_1, \\
y_2 & = & \beta_0 + \beta_1 X_{2,1} + \beta_2 X_{2,2} + \dots + \beta_p X_{2,p} + \varepsilon_2, \\
\dots & = & \dots \\
y_{n-1} & = & \beta_0 + \beta_1 X_{n-1,1} + \beta_2 X_{n-1,2} + \dots + \beta_p X_{n-1,p} + \varepsilon_{n-1}, \\
y_n & = & \beta_0 + \beta_1 X_{n,1} + \beta_2 X_{n,2} + \dots + \beta_p X_{n,p} + \varepsilon_n,
\end{array}$$

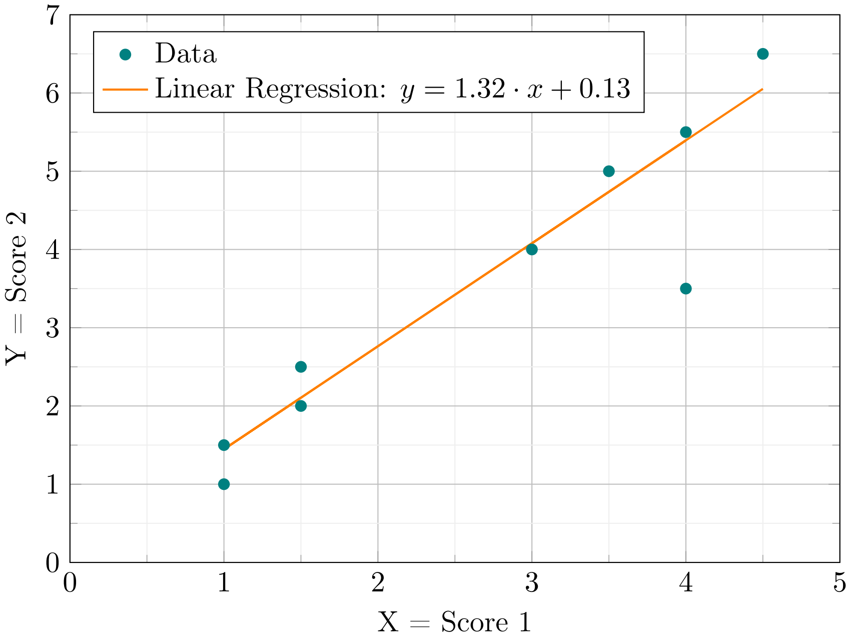
where $X_{i,j}$ indicates the j-th feature of individual i.

Let us look at an example where we want to predict the score obtained on a second exam, based on the score obtained on a first exam for a set of 10 students.

Our data is presented as follows

Y : Score exam 2	3.5	4	5	1	2	1.5	2.5	5.5	6	6.5
X : Score exam 1	4	3	3.5	1	1.5	1	1.5	4	3.5	4.5

The objective here will be to learn the coefficients of the regression line. We can graphically represent the data in the graph below, as well as the obtained line



L'objectif sera d'étudier comment nous pouvons déterminer ces coefficients à l'aide notre jeu de données.
The objective will be to study how we can determine these coefficients using our dataset.

We will also extend our study by examining the statistical properties of the estimators $\hat{\boldsymbol{\beta}}$ obtained from $\boldsymbol{\beta}$, this will be used to build appropriate statistical tests in order to check if the avaiable informations are important for the prediction task.

1 Simple Linear and Gaussian Model

In this first part, we will focus on the **simple** linear model, meaning the model where we aim to predict the values of the variable Y based solely on a single variable X.

Our model is therefore written as

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where β_0 represents the intercept of the model and β_1 represents the slope of our line.
It is this coefficient β_1 that describes the impact of the variable X on the variable Y.

Mathematically speaking, we should more precisely use the term *affine* rather than *linear*, since the learned line does not necessarily pass through the origin, except in the case where $\beta_0 = 0$.

From now on, we will also assume that the explanatory variable X follows a normal distribution.
Let is now try to understand how we can estimate these parameters.

1.1 Assumptions of the Gaussian Linear Model

Our *simple* Gaussian model (since we use only one variable) aims to explain the relationship between two quantitative variables, X and Y, through an affine relation

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where $Y \in \mathbb{R}$ denotes the dependent variable, $X \in \mathbb{R}$ the explanatory variable, $\boldsymbol{\beta} = (\beta_0, \beta_1)$ the model parameters we seek to estimate, and ε an independent random error term.

We stated that to estimate the model parameters, we use a dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, *i.e.*,

$$\forall i \in \llbracket 1, n \rrbracket,\ Y_i = \beta_0 + \beta_1 X_{i,1} + \varepsilon_i.$$

Gaussian Model Assumptions

We formulate the following assumptions for our Gaussian linear model:

1. $(Y_i, X_i)_{i=1}^n$ must be *i.i.d.*, i.e., independently and identically distributed,
2. $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$,
3. $\varepsilon_i \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$: homoscedasticity assumption.

The first and second assumptions specify that the values Y_i are **observed and random**, whereas the values X_i are **observed and non-random** (also referred to as deterministic). The third assumption states that the errors are random and follow a Gaussian distribution that is (i) centered, (ii) of unknown variance σ^2, and (iii) independent. This last point means that the **covariance** between the errors associated with individuals i and j is zero, *i.e.*,

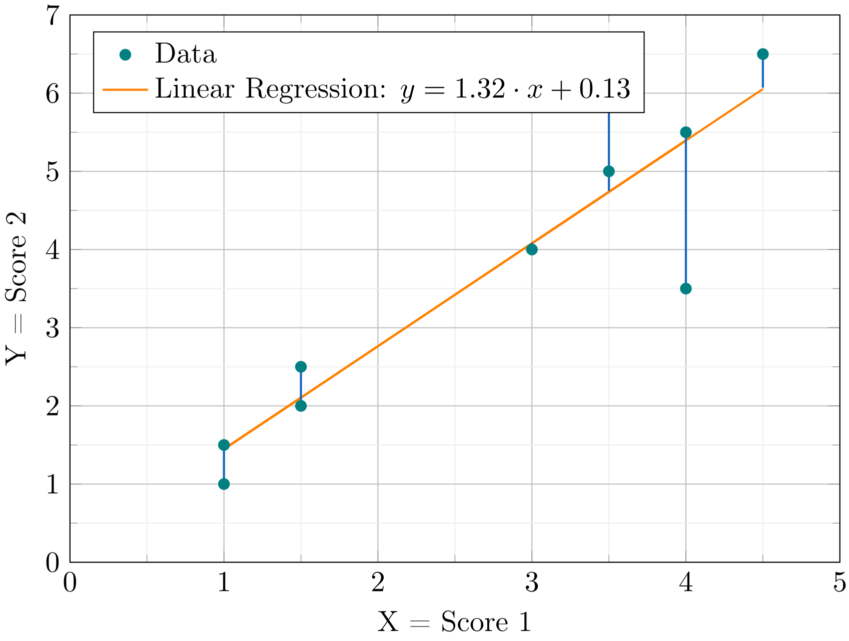
$$Cov(\varepsilon_i, \varepsilon_j) = 0, \quad \forall\, i \neq j.$$

1.2 Optimization

Our objective is to determine the values of the model parameters such that the predicted value

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

is as close as possible to the actual value y_i for the different individuals x_i, with $i = 1, \ldots, n$.



We could be tempted to evaluate this difference $y_i - \hat{y}_i = \varepsilon_i$ over all individuals, *i.e.*, we could try to solve the problem

$$\min_{\beta_0,\beta_1 \in \mathbb{R}} \sum_{i=1}^{n} \varepsilon_i = \min_{\beta_0,\beta_1 \in \mathbb{R}} \sum_{i=1}^{n} y_i - \hat{y}_i = \min_{\beta_0,\beta_1 \in \mathbb{R}} \sum_{i=1}^{n} y_i - (\beta_0 + \beta_1 x_i).$$

However, this would not be a good definition of the model errors, also called **residuals**. In fact, errors should be counted positively, but here $\varepsilon_i = y_i - \hat{y}_i$ can be either positive or negative, leading to compensatory effects. Moreover, by definition, these errors are centered, so the sum of the errors, as defined here, would be equal to 0.

We could take the **absolute value of the difference** between the observation y_i and the prediction $\hat{y}_i$, *i.e.*,

$$\min_{\beta_0,\beta_1 \in \mathbb{R}} \sum_{i=1}^{n} |y_i - (\beta_0 + \beta_1 x_i)|\,,$$

but this problem remains difficult to solve mathematically. Therefore, we prefer to minimize the **squared distance** between y_i and $\hat{y}_i$. This approach is known as the **least squares method**. This method has a considerable advantage, compared to the maximum likelihood method we will see later, because it does not require any assumptions about the distribution of the errors.
We will thus solve the problem

$$\min_{\beta_0,\beta_1\in\mathbb{R}} \sum_{i=1}^{n} \varepsilon_i^2 = \min_{\beta_0,\beta_1\in\mathbb{R}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min_{\beta_0,\beta_1\in\mathbb{R}} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 .$$

1.3 Expression of the Solutions

Before presenting the expression for the solutions to our optimization problem, we recall the following probability result.

Recap of Probability. We recall the following results from probability theory.

Lemma 1.1: Variance of Random Variables

Consider X and Y as random variables with second-order moments, *i.e.*, they have a variance. Then,

(i) For the variance of a random variable, we have the **Koenig-Huygens Formula**:

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 .$$

(ii) The covariance between two random variables is also given by

$$Cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y].$$

Proof. We prove the two points separately.

(i) Koenig-Huygens Formula. Starting from the definition of variance:

$$\begin{aligned}
\mathrm{Var}[X] &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\
&\quad \downarrow \text{ by expanding the expression} \\
&= \mathbb{E}\left[X^2 - 2X\,\mathbb{E}[X] - \mathbb{E}[X]^2\right] , \\
&\quad \downarrow \text{ by the linearity of expectation} \\
&= \mathbb{E}[X^2] - 2\,\mathbb{E}[X]^2 + \mathbb{E}[X]^2, \\
\mathrm{Var}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2.
\end{aligned}$$

(ii) Covariance Equality. Similarly, starting from the definition of covariance:

$$\begin{aligned}
Cov(X,Y) &= \mathbb{E}\left[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\right], \\
&\quad \downarrow \text{ by expanding} \\
&= \mathbb{E}\left[XY-\mathbb{E}[X]Y-Y\,\mathbb{E}[X]+\mathbb{E}[X]\,\mathbb{E}[Y]\right], \\
&\quad \downarrow \text{ by the linearity of expectation} \\
&= \mathbb{E}[XY]-\mathbb{E}[X]\,\mathbb{E}[Y]-\mathbb{E}[Y]\,\mathbb{E}[X]+\mathbb{E}[X]\,\mathbb{E}[Y], \\
Cov(X,Y) &= \mathbb{E}[XY]-\mathbb{E}[X]\,\mathbb{E}[Y].
\end{aligned}$$

□

The previous lemma will allow us to provide a simpler proof for the expression of the optimal parameters of our regression model.

Proposition 1.1: Linear Regression Problem

Consider the Gaussian linear regression problem of the form

$$Y=\beta_0+\beta_1X+\varepsilon.$$

The parameters a and b are solutions to the optimization problem

$$\min_{\beta_0,\beta_1\in\mathbb{R}}\sum_{i=1}^{n}\varepsilon_i^2=\min_{\beta_0,\beta_1\in\mathbb{R}}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\min_{\beta_0,\beta_1\in\mathbb{R}}\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1x_i))^2.$$

The solutions are given by

$$\hat{\beta}_1=\frac{Cov[X,Y]}{\mathrm{Var}[X]}=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\quad\text{and}\quad\hat{\beta}_0=\mathbb{E}[Y]-\mathbb{E}[X]\times\hat{\beta}_1=\bar{y}-\hat{\beta}_1\times\bar{x}.$$

Proof. The function L that we seek to optimize, defined by

$$L(\beta_0,\beta_1)=\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1x_i))^2$$

is a convex function in the variables β_0 and β_1, so it has a unique solution. This solution is obtained by solving the Euler equation, which takes the form of a linear system

$$\frac{\partial L}{\partial \beta_1} = 0 \iff -2\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0)x_i = 0, \tag{1}$$

$$\frac{\partial L}{\partial \beta_0} = 0 \iff -2\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0) = 0. \tag{2}$$

To finish the proof, it remains to solve the linear system.

□

Exercise

Finish the above proof by solving the linear system in order to find the results presented in Proposition 1.1, *i.e.*, find the expression of the estimated $\hat{\beta}_0$ and $\hat{\beta}_1$.

1.4 Estimation of the Variance σ^2

In the case of the simple linear model, we can obtain two estimates of this variance by focusing on the model's residuals.

If we estimate σ^2 using the traditional definition of the variance, we can show that one estimator is given by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2.$$

However, this estimator is biased. A debiased version of this estimator is given by the expression

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2.$$

Thus, we need to divide the sum of the squared residuals ε_i by $n-2$ rather than by n. Although the reasoning behind dividing by $n-2$ is not explained here, we can note that the 2 refers to the number of parameters in our regression model, β_0 and β_1.

Determining the expectation and variance of these estimators will allow us to assess whether the learned model is meaningful, *i.e.* **whether we are able to correctly predict the values of the variable Y using the values of X and this linear relationship**.

However, we postpone this analysis to the Section 2.

1.5 Measuring the Relationship Between the Explanatory and Response Variables

We have already introduced a measure to study the relationship between two random variables, X and Y, called the **covariance**. However, the value of this covariance depends on the scale of the values taken by the different random variables. Thus, to measure the relationship between two random variables, we calculate the **linear correlation coefficient**.

Definition 1.1: Linear Correlation Coefficient

Let X and Y be two random variables with second-order moments. The linear correlation coefficient between the variables X and Y is the quantity ρ defined by

$$\rho = \frac{\operatorname{Cov}[X, Y]}{\sqrt{\operatorname{Var}[X] \operatorname{Var}[Y]}}.$$

If $|\rho|$ is close to 1, we say that the correlation between the two variables is *strong*. Conversely, if it is close to 0, the correlation is *weak*.
Furthermore, a **negative** value of ρ means that, generally, *increasing* values of X lead to *decreasing* values of Y (and vice versa), *i.e.*, the slope of the regression line will be **negative**. Similarly, a positive value of ρ means that *increasing* values of X lead to *increasing* values of Y (and vice versa), *i.e.*, the slope of the regression line will be **positive**.

Exercise

Consider to random variables X and Y and let us denote by ρ their Linear Correlation Coefficient. Draw a of a set of points for which we have:

1. a positive correlation, *i.e.* $\rho > 0$,
2. a negative correlation, *i.e.* $\rho < 0$,
3. a correlation close to 0, *i.e.* $\rho \simeq 0$.

Using the definition of β_1, *i.e.* the slope, and ρ, show that these two quantities are linked.

1.6 Significance of the Model

Given the remark made earlier, we can therefore focus on either of the two quantities.

Significance of the Slope β_1 We perform the test to determine whether the slope is significantly different from 0. The hypotheses are as follows:

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 = \beta_1 \neq 0$$

Our estimator of the slope, $\hat{\beta}_1$, follows a normal distribution, just like the random variable Y. The parameters of this distribution allow us to write that

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).$$

We thus consider the following statistical test t_{test}:

$$t_{\text{test}} \underset{\text{under } H_0}{=} \frac{\hat{\beta}_1}{\sqrt{\dfrac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim T_{n-2}.$$

Recall that we reject H_0 at the significance level $\alpha \in [0, 1]$ if the test statistic t_{test} lies outside the confidence interval at the $1 - \alpha$ level, *i.e.* if

$$t_{\text{test}} \notin [t_{\alpha/2,n-2}, t_{1-\alpha/2,n-2}].$$

Alternatively, we could compare the p-value $= 2\,\mathbb{P}[T \geq |t_{\text{test}}|]$ to the significance level α and reject H_0 if the p-value is smaller than this threshold.

Significance of the Correlation Now, we seek to perform the same analysis but this time study the significance of the correlation coefficient ρ.

The test leads us to pose the following hypotheses:

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

It relies on a test statistic similar to the one for the slope:

$$t_{\text{test}} = \frac{\hat{\rho} - \rho}{\sqrt{\dfrac{1 - \hat{\rho}^2}{n - 2}}} \underset{\text{under } H_0}{=} \frac{\hat{\rho}}{\sqrt{\dfrac{1 - \hat{\rho}^2}{n - 2}}} \sim T_{n-2}.$$

We proceed in the same way as before to conclude on the significance of the slope, given a significance level α.

Exercise

Show that the two statistical test introduced in this section are equal, using the relation between $\hat{\beta}_1$ and $\hat{\rho}$.

1.7 Writing the Model in Matrix Form

Finally, note that we could have written our simple linear regression model in the following matrix form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_{n-1} \\ 1 & x_n \end{pmatrix} \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{n-1} \\ \varepsilon_n \end{pmatrix},$$

We will use this form when presenting the multiple linear model, *i.e.* when the model employed uses multiple descriptors (or covariates) $X_1, X_2, \ldots, X_p$, which is the subject of Section 2.

2 Multiple Linear Regression Model

In this section, we will assume that the number of examples n is always greater than the number of descriptors (or variables) $p+1$[1].

The multiple linear regression model is written in the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n-1,1} & x_{n-1,2} & \dots & x_{n-1,p} \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{n-1} \\ \varepsilon_n \end{pmatrix},$$

and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{p-1}, \beta_p) \in \mathbb{R}^{p+1}$ is our vector of model parameters. The vector $\mathbf{y} \in \mathbb{R}^n$ is the vector whose values we aim to explain, the matrix $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ is the explanatory matrix, also called the *design matrix*, and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is the vector of errors associated with each example, which are assumed to be Gaussian.

Hypotheses . The following hypotheses are typically made for the study of the Gaussian linear model:

1. The model is assumed to be identifiable, *i.e.*, there exists a unique vector $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ such that $\mathbb{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$. This is equivalent to the condition that the columns of the matrix $\mathbf{X}$ are linearly independent, *i.e.*, the rank of the matrix $\mathbf{X}$ is $p+1$.

2. Our data are *i.i.d.*, as in the case of simple linear regression.

3. The errors are assumed to be centered, so $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$.

4. The errors have the same variance and are independent, thus $\text{Var}[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I}_n$, or equivalently $\text{Var}[\mathbf{y}] = \sigma^2\mathbf{I}_n$.

2.1 Estimation by the Least Squares Method

Let us now look at the expression for the least squares estimator.

[1] The case where $p+1$ is much larger than n would lead us to perform high-dimensional statistics, which is not the focus of this course, especially since it would make the study of models more complex.

Proposition 2.1: Solution of Multiple Regression

Consider the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

with the same meaning as before. If the model is identifiable, *i.e.*, if the matrix $\mathbf{X}$ has rank $p+1$, then the matrix $\mathbf{X}^\top\mathbf{X}$ is invertible, and the least squares estimator of $\boldsymbol{\beta}$, the solution to the problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

is given by

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}.$$

Proof. To determine the expression of the estimator $\hat{\boldsymbol{\beta}}$, we will differentiate our problem and look for critical points, then determine their nature by studying the associated Hessian.
The extrema of the function $\boldsymbol{\beta} \mapsto \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2$ are found by looking for the point where the gradient of this function is zero. We will thus seek the values of $\boldsymbol{\beta}$ such that

$$\frac{\partial}{\partial\boldsymbol{\beta}}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = 0 \iff -2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \tag{3}$$

By differentiating the function again, we obtain

$$\frac{\partial^2}{\partial\boldsymbol{\beta}^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = 2\mathbf{X}^\top\mathbf{X} \succ 0,$$

i.e. the Hessian matrix is positive definite, which is the case here since it is the variance-covariance matrix of the data. This convexity allows us to conclude that the vector $\boldsymbol{\beta}$ satisfying equation (3) is indeed the solution to our minimization problem. Now, we have

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\beta}}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = 0 &\iff -2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0,\\
&\downarrow \text{ (we can divide by } -2)\\
&\iff \mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0,\\
&\iff \mathbf{X}^\top\mathbf{y} - \mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = 0,\\
&\iff \mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{y},
\end{aligned}$$

$$\frac{\partial}{\partial \boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = 0 \underset{\downarrow \text{ (since the model is identifiable, the matrix } \mathbf{X}^\top\mathbf{X} \text{ is invertible)}}{\iff} \boldsymbol{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.$$

□

Exercise

Use the expression of the solution presented in Proposition 2.1 in order to find the expression of the estimated slope and intercept presented in Proposition 1.1. Saying differently, find the expression of $\boldsymbol{\beta}$ when $p = 1$.

Tout comme dans le cas du modèle linéaire simple, les prédictions $\hat{\mathbf{y}}$ sur les données $\mathbf{X}$ utilisées pour estimer $\boldsymbol{\beta}$ sont définies par

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.$$

Or, equivalently for a single instance $\mathbf{x} = (x_1, \dots, x_p)$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p.$$

We now study the properties of the solution that are going to be used to study the significance of the variables.

Proposition 2.2: Properties of the estimator $\hat{\beta}$

The ordinary least squares estimator $\hat{\boldsymbol{\beta}}$ has the following properties:

(i) It is an unbiased estimator of the parameter $\boldsymbol{\beta}$, *i.e.* $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$

(ii) Its variance is equal to $\mathrm{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$.

(iii) Moreover, $\hat{\boldsymbol{\beta}}$ is the **unbiased estimator with minimum variance** among all unbiased linear estimators of $\boldsymbol{\beta}$.

We will just use the to first point of the proposition which provide information about the expectation the variance of the estimated parameter $\hat{\boldsymbol{\beta}}$.

2.2 Estimation of the variance σ^2

The parameter σ^2, which is simply the variance of our residuals, is defined by

$$\sigma^2 = \mathrm{Var}[\varepsilon] = \mathrm{Var}[Y] = \mathbb{E}[(Y - \mathbb{E}[Y])^2].$$

Recall that in a linear model, the expectation of the random variable Y is estimated by $X\hat{\boldsymbol{\beta}}$. We will then estimate the value of σ^2 using the residuals $\hat{\varepsilon}_i$ from our model. The estimator $\hat{\sigma}^2$ of the variance of our residuals is defined by

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} \hat{\varepsilon}_i^2.$$

This used to complete the point *(ii)* of Proposition 2.2, where the value σ^2 so that an estimated is required.

2.3 Test of Nullity of a Regression Coefficient

We wish to determine whether the j-th coefficient of the regression is significantly different from 0 or not. In other words, we will try to determine whether the j-th variable helps to explain, in part, the values taken by the random variable Y.
We therefore state the following hypotheses:

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

To construct this test, we need to know the distribution of $\hat{\beta}_j$ in order to deduce the distribution of our test statistic under H_0.

We have previously seen that the estimator $\hat{\boldsymbol{\beta}}$ follows a multivariate Gaussian distribution with mean $\boldsymbol{\beta}$ and variance matrix $(\mathbf{X}^\top \mathbf{X})^{-1}$, *i.e.*,

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}).$$

Without going into deeper details on how to build the statistical test, we will just present it in the following proposition.

Corollary 2.1: Test Statistics

For all $j \in [\![0, p]\!]$, consider β_j as the coefficient of the regression associated with the variable X_j and $\hat{\beta}_j$ as its estimator.
Then

$$t_{\text{test}} = \frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}} = \underset{\text{under } H_0}{=} \frac{\hat{\beta}_j}{\sigma_{\hat{\beta}_j}} \sim T_{n-(p+1)},$$

where $\sigma^2_{\hat{\beta}_j}$ denotes the square root of the value located at position $(j+1, j+1)$ in the matrix $\hat{\sigma}^2(\mathbf{X}^\top \mathbf{X})^{-1}$.

In the context of a two-tailed test, where we aim to check if the coefficient β_j is significantly different from 0, we will reject the null hypothesis, with a risk of error $\alpha \in (0, 1)$, if the test statistic

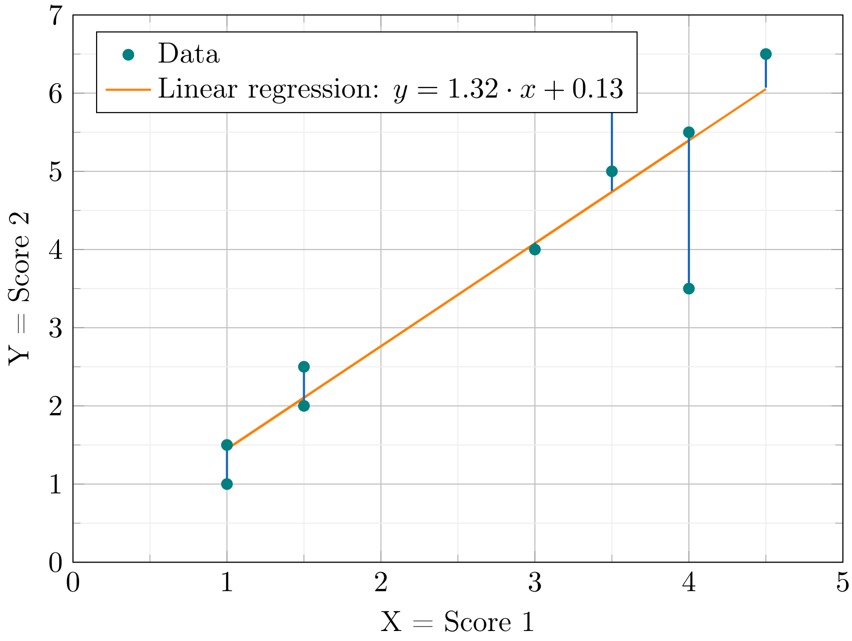
$$|t_{\text{test}}| = \left| \frac{\hat{\beta}_j}{\sigma^2_{\hat{\beta}_j}} \right| \geq t_{1-\alpha/2, n-p-1}.$$

This result also allows for constructing confidence intervals for the estimators of the model's parameters.

2.4 Model Quality

Recall that we aim to estimate the parameters of the model in order to minimize the squared difference between the observed values y_i and the values predicted by the model $\hat{y}_i$, *i.e.* to solve the problem

$$\min_{\beta_0, \beta_1 \in \mathbb{R}} \sum_{i=1}^{n} \varepsilon_i^2 = \min_{\beta_0, \beta_1 \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min_{\beta_0, \beta_1 \in \mathbb{R}} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 .$$



The sum of squared differences between y_i and $\hat{y}_i$ is also called the *Residual Sum of Squares (SSR)*, *i.e.*;

$$\mathrm{SSR} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2.$$

It is closely related to the variance initially present in our data. In fact, we can show that the following relationship holds between the variance of our observations and the SSR:

$$\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{SST} = \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}_{SSE} + \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{SSR},$$

where $\bar{y}$ denotes the mean value of $\mathbf{y} \in \mathbb{R}^n$.

The term **SST** can be seen as the variation (or amount of information) present in the data, ESS represents the variation explained by the model, and RSS represents the variation not explained by the model (or residual variance).

Using these different quantities, we can again construct a statistical test to test the overall significance of the model, *i.e.*, we perform the following test:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p \quad \text{vs.} \quad \exists\, j \in [\![1, p]\!]\ \beta_j \neq 0.$$

In other words, testing the significance of the model involves testing the hypothesis that none of the covariates explain the observed values of y. Let's look at how this test is constructed.

Analysis of Variance and Model Significance Since the previous terms represent variances, we often summarize the information from our model in an analysis of variance table

Analysis of Variance Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares
Model (SSE)	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	p	$\text{MSE} = \frac{\text{SSE}}{p}$
Residual (SSR)	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-p-1$	$\text{MSR} = \frac{\text{SSR}}{n-p-1}$
Total (SST)	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$	$\text{MST} = \frac{\text{SST}}{n-1}$

Note that the different sums of squares can also be written using the norm of two vectors, as was done for the residual part.

We can then define the following test statistic F_{test} to test the overall significance of the model:

$$F_{\text{test}} = \frac{\text{MSA}}{\text{MSW}} \; F_{p,n-p-1}.$$

This test statistic follows a Fisher distribution with p and $n-p-1$ degrees of freedom. We then reject the hypothesis H_0 with a risk of error $\alpha \in (0,1)$ if

$$F_{\text{test}} > f_{p,n-p-1,1-\alpha},$$

i.e. if the test statistic takes a value greater than the $(1-\alpha)$ quantile of a Fisher distribution with p and $n-p-1$ degrees of freedom.
We note that this is a variance ratio test, and it is a *one-tailed upper* test.

Model Quality and Fit In the case of simple linear models, we could assess the quality of the model's fit to the data by evaluating the correlation between the two variables X and Y. However, it becomes more difficult to do this in higher dimensions, but it is still possible to assess the fit using **the coefficient of determination** R^2, which studies the proportion of variance explained by the model (SCE) relative to the total variance (SCT). More precisely:

$$R^2 = \frac{\text{SCE}}{\text{SCT}} = \frac{\|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2}{\|\mathbf{y} - \bar{\mathbf{y}}\|^2}.$$

This coefficient of determination is a value between 0 and 1. The closer this value is to 1, the more the model explains the observed data.
Note: One might think that having a value close to 1 is very interesting in practice, but it is not always a guarantee that the learned model is reliable and performs well. It is possible that the model memorizes the data, which is known as *overfitting*. We will discuss this later in the *Machine Learning* course.

This criterion unfortunately has a disadvantage: its value naturally increases with the number p of explanatory variables in the model. Therefore, it is sometimes common to consider another criterion, called the **adjusted coefficient of determination**, the adjusted R^2, which takes into account the number of variables in the model. This quantity is defined by:

$$R^2\text{-adjusted} = 1 - \frac{n-1}{n-p-1}(1 - R^2).$$

In addition to evaluating the quality of models, these criteria will help us in model selection.

2.5 Model Construction and Selection

Several questions arise when constructing a prediction model using the available information:

1. Is the learned model of good quality? This is a point we have already studied.
2. Do the variables used to build the model provide different information? This is the question of information redundancy.
3. Do all these variables contribute significantly to the model's performance? Here, we are more concerned with the importance of the information.

Information Redundancy The answer to the second question is important to ensure the validity of the model. Indeed, we recall that in order to estimate the parameters $\boldsymbol{\beta}$, the matrix $\mathbf{X}$ must be of full rank so that the matrix $\mathbf{X}^\top\mathbf{X}$ is invertible. If there exists a variable X_j that can be expressed as a linear combination of all the other variables X_k, then our matrix will no longer be invertible, *i.e.* if there exist coefficients α_k such that

$$X_j = \sum_{k=1,\ k\neq j}^{p} \alpha_k X_k.$$

This relation also means that we can predict the value of X_j given the values of X_k using a linear model! From this observation, we could also estimate that variables X_j are only *weakly useful* (or even harmful to the model quality) if there is a strong linear relationship with the other variables X_k.
To evaluate this in practice, we will construct a linear model between X_j and the other variables X_k and assess the *goodness of fit*, the R^2. If this value is too high, we will consider the variable X_j redundant.
This criterion is called the *VIF* or *Variance Inflation Factor* and is defined by:

$$VIF(X_j) = \frac{1}{1 - R_j^2},$$

where R_j^2 is the coefficient of determination associated with the model $X_j = \sum_{k=1,\ k\neq j}^{p} \alpha_k X_k$.
The variable is then excluded from the data if its VIF is greater than 10 (some authors may choose 5).

Thus, before attempting to build a model, we first aim to remove redundant information by applying the following procedure:

1. Calculate the VIFs for all the variables X_j by evaluating the R_j^2 associated with the models

$$X_j = \sum_{k=1,\ k\neq j}^{p} \alpha_k X_k.$$

2. If only one variable has a VIF greater than 10, exclude this variable from the dataset and stop the procedure.

3. If several variables have a VIF greater than 10, remove the variable with the highest VIF and return to step 1, until all variables have a VIF less than 10.

Now, we need to consider which information is essential to build the best model.

Model Selection We will now look at how to select models in general, *i.e.*, what criterion(ia) we can use to compare models. We will then consider the specific case of nested models.

We have already seen the adjusted R^2 criterion earlier, but we can also use criteria based on the likelihood of our data.

(i) Mallows' C_p coefficient [Gilmour, 1996].

For a model Ω_q containing $q < p$ variables, this criterion is defined by:

$$C_p(\Omega_q) = \frac{\|\mathbf{y} - \hat{\mathbf{y}}(\Omega_q)\|^2}{\|\mathbf{y} - \hat{\mathbf{y}}\|^2} 2(q+1) - n,$$

where $\hat{\mathbf{y}}$ represents the predictions made with the full model and $\hat{\mathbf{y}}(\Omega_q)$ represents the predictions made with the reduced model Ω_q.

(ii) *Akaike Information Criterion (AIC)* [Akaike, 1974].

This criterion was primarily motivated by the study of Gaussian models and is defined for a model Ω_q, containing $q < p$ variables:

$$\text{AIC}(\Omega_q) = n\left(\ln(2\pi) + 1\right) + n \ln\left(\frac{\|\mathbf{y} - \hat{\mathbf{y}}(\Omega_q)\|^2}{n}\right) + 2(q+2),$$

where $\hat{\mathbf{y}}(\Omega_q)$ represents the predictions made by the model with q variables.

In the Gaussian case, it can be shown that the AIC and C_p criteria are equivalent. The following criterion is the most commonly used in statistics.

(iii)*Bayesian Information Criterion (BIC)* [Schwarz, 1978].

Using the same notation as before, we have:

$$\text{BIC}(\Omega_q) = n\left(\ln(2\pi) + 1\right) + n \ln\left(\frac{\|\mathbf{y} - \hat{\mathbf{y}}(\Omega_q)\|^2}{n}\right) + \ln(n)(q+2).$$

It can be shown that when $n > 7$, we have $\ln(n) > 2$, so the BIC criterion tends to select smaller models than the AIC criterion. The goal is to select the model Ω_q that **minimizes** one of these three criteria.

There is a procedure to test whether a variable significantly increases or not the performance of a model. We say we are *comparing nested models.*

To do this, consider an integer $q < p$ and consider the models Ω_q and Ω_{q+1} where Ω_q is a model containing q variables from the $q+1$ variables in the model Ω_q. To test whether adding or removing this variable significantly affects the model's performance, we can perform the statistical test with the following hypotheses:

$$H_0 : \text{the model } \Omega_q \text{ is valid} \quad \text{v.s.} \quad \text{the model } \Omega_{q+1} \text{ is valid.}$$

To see if the addition of the new variable is significant or not, we compare the adjusted R^2-adjusted and if there is an improvement, we assume that the new information is important.

2.6 Residual Analysis and Outlier Detection

We now want to check whether the assumptions of the Gaussian linear model are valid or not. This is a step we perform *a posteriori* after selecting the best model according to the statistical criteria defined in the previous section.

We have already checked the assumption of identifiability of the model (for obtaining the solutions) when we introduced the VIF for detecting potential collinearities between the variables. But we must also check the assumptions listed below:

Gaussian Model Assumptions

We formulate the following assumptions for our linear Gaussian model:

1. $(Y_i, X_i)_{i=1}^n$ must be *i.i.d.*, *i.e.* independent and identically distributed,
2. $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i +, \sigma^2)$,
3. $\varepsilon_i \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$: homoscedasticity assumption.

It is essentially a matter of verifying the assumptions related to the **residuals** $\hat{\varepsilon}_i$ of the model.

Residual Analysis Recall that the residuals are defined, for any integer $i \in [\![1, n]\!]$, by

$$\hat{\varepsilon}_i = y_i - \hat{y}_i.$$

The validation of the assumptions will primarily be done using graphs.

(i) Homoscedasticity of the residuals. To verify this first assumption, we will study the so-called *normalized* residuals.

Recall that we have:

$$\mathrm{Var}[\hat{\varepsilon}] = \sigma^2(\mathbf{I} - \mathbf{H}),$$

where $\mathbf{H}$ is the orthogonal projection matrix onto the space spanned by $\mathbf{X}$. In particular, we thus have

$$\mathrm{Var}[\hat{\varepsilon}] = \sigma^2(\mathbf{I} - \mathbf{H}_{i,i}),$$

Again, we need to use an estimate $\hat{\sigma}^2$ of σ^2 to perform our analysis. Let

$$\mathrm{Var}[\hat{\varepsilon}] = \hat{\sigma}^2(\mathbf{I} - \mathbf{H}_{i,i}).$$

The residuals may have different variances (this depends on the value of $H_{i,i}$), so we will normalize them:

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - H_{i,i}}}.$$

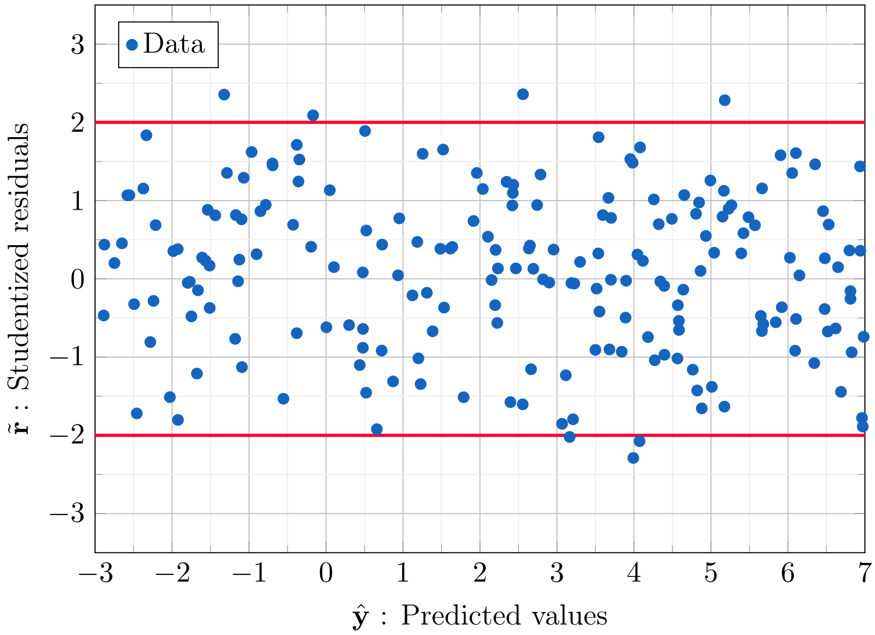
These residuals r_i are called **standardized residuals**, where $\hat{\sigma}^2 = \dfrac{1}{n-(p+1)} \sum_{i=1}^n \hat{\varepsilon}_i^2$.

However, there is an issue with the definition of these residuals: the value $\hat{\varepsilon}_i$ appears both in the numerator and denominator, making these two quantities dependent, which may hinder the analysis of the *homoscedasticity* assumption.

We will therefore consider the so-called **studentized** residuals $\tilde{r}_i = \dfrac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1 - H_{i,i}}}$,

where $\hat{\sigma}_{(i)} = \dfrac{1}{n-p-2} \sum_{j \neq i}^n \hat{\varepsilon}_j^2$.

We will thus create a plot of the **studentized residuals versus the predicted values**. This choice is explained by Cochran's Theorem, which guarantees that the predicted values $\hat{y}_i$ and the associated residuals $\hat{\varepsilon}_i$ are independent. This is preferable compared to the graphs $(\hat{\varepsilon}_i, X_i)$ or $(\hat{\varepsilon}_i, Y_i)$.



For large values of n, we know that 95% of the values of the Student's t distribution should lie within the interval $[-2, 2]$. If too many values lie outside this interval, we cannot say that σ^2 is independent of X_i, thus contradicting the homoscedasticity assumption.

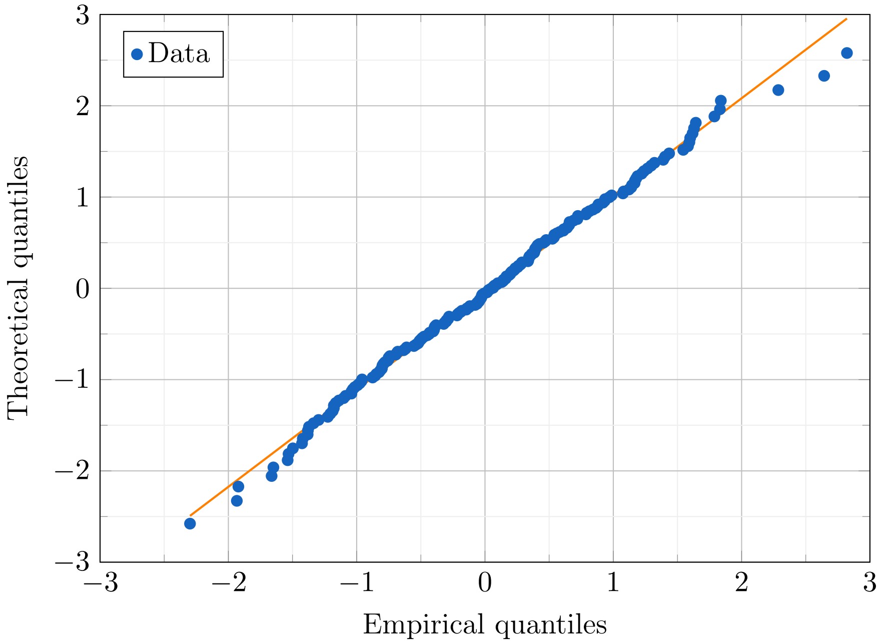
(ii) Normality of the residuals. We will now look at how to test the **normality of the residuals**. We can do this in two ways: *(i)* using a statistical test or *(ii)* using a graphical method.

For the first approach *(i)*, we perform a *Shapiro-Wilk test*. The test takes the following form:

$$H_0 \,:\, \text{the residuals are normally distributed vs. } H_1 \,:\, \text{they are not}$$

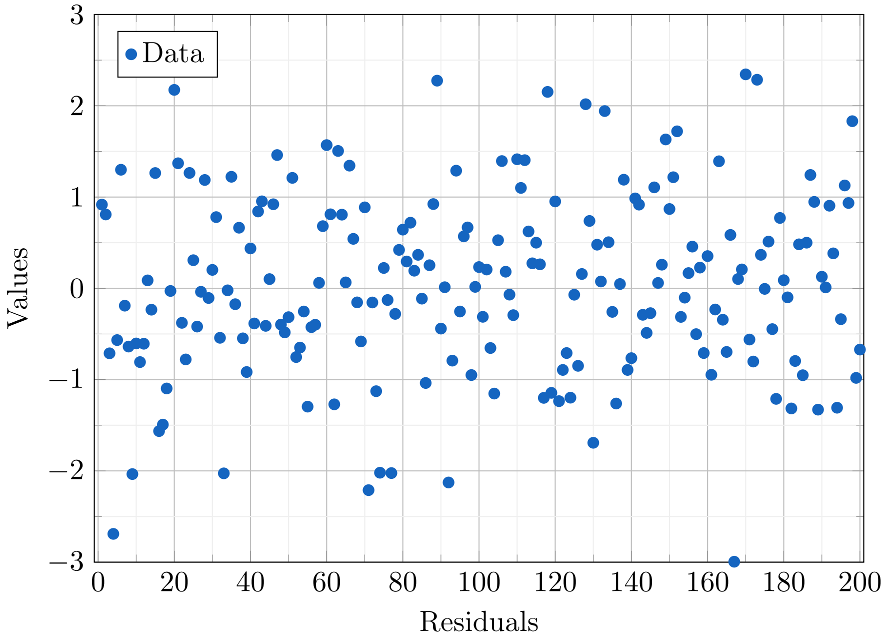
This test is extremely powerful but also extremely rigid, making it less useful in practice because it will tend to reject the normality hypothesis very often.

We therefore prefer to use a graphical method *(ii)* based on comparing the empirical quantiles of the residuals with the theoretical quantiles of the standard normal distribution.

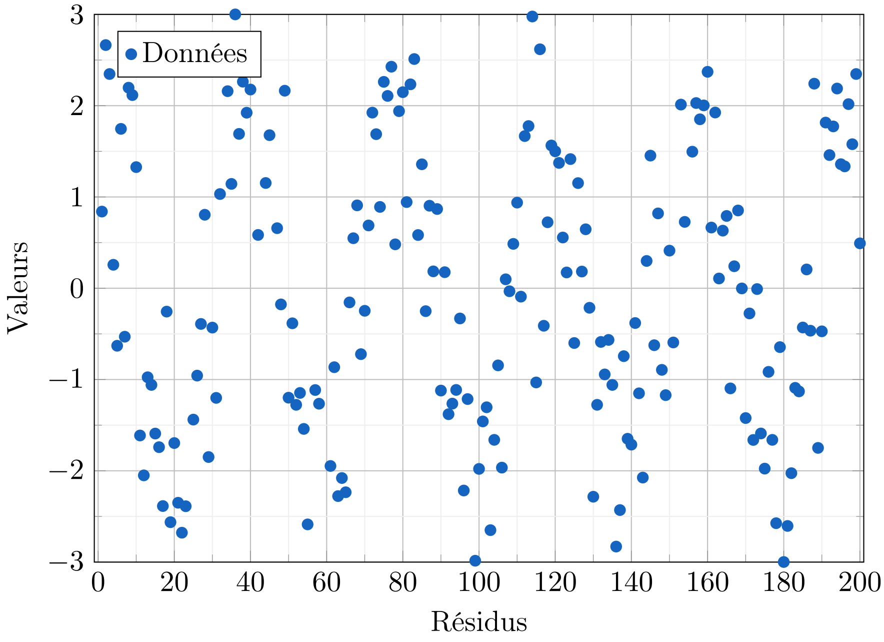


We will assume that the normality hypothesis will not be contradicted if the points are generally aligned.

(iii) Independence of the residuals. This is to check that no patterns emerge when graphing the residuals.
Below, we represent residuals where the independence assumption is verified.



In contrast, the graph below shows that the independence assumption is not verified.



Indeed, we observe a cyclic pattern in the residual values, which is a sign of the presence of *autocorrelation* in the data. This is typical in the context of *time series/data analysis*, a topic not covered in this course.

The study of potential *outliers* involves measuring the impact of each observation in the regression. This measurement can be done in two different ways: *(i)* by calculating the *hat values* or *Cook's distance.*

2.7 Take into account Categorical Variables in the Model

A last thing to study is how to take into account categorical variables in the model, such as the gender of the person. It is very important to study this point because the linear model is only able to work with numerical/quantitative variables and not categorical ones which can be seen as text.
Let us consider two different settings.

(i) Case of two modalities

Suppose that our variable X is categorical variable that takes two different values. Let us assume that $X \in \{\text{M,F}\}$. The usual thing to do is to *encode* this new variable into $0-1$ by choosing one *modility* (*i.e.*, M or F) as a reference.

For instance, if you choose F as the reference, than X will take the value 0 if it is equal to "F" and 1 otherwise.

Assume that we are working with a model that has two variables X_1 and X_2, X_1 is a quantitative variable and X_2 a categorical variables which takes to values. thus it can be written as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

If we apply the process described before, this will lead, indirectly, to two different models: *(i)* when $X_2 = 0$:

$$Y = \beta_0 + \beta_1 X_1 \varepsilon$$

and *(ii)* when $X_2 = 1$

$$Y = \beta_0 + \beta_2 + \beta_1 X_1 + \varepsilon.$$

When looking at these two models, we can see that the categorical variables may only impact the intercept of the model.

But it is also possible to go a little bit further by including an *interaction term:* X_1X_2 ad thus study the following model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

by adding this term $\beta_3 X_1 X_2$, we aim to study if the categorical variable has also on impact on the slope of the model, the one associated to the variable X_1.

(ii) Case of more than two modalities

Imagine that we have a variable a categorical variable X which has, for instance, 4 modalities, such as *summer, automn, winter* and *spring.*

If we follow the same idea as in the *two modalities* setting, we have to assign different figures to each season. However, by replacing each category with numerical values such as 0, 1, 2, and 3, we implicitly impose an ordinal structure on the seasons, which can bias the analysis. This would mean, for example, that summer (coded as 3) is more important than winter (coded as 0), which has no statistical justification in itself.

This is why, in linear regression, categorical variables are encoded using indicator variables (one-hot encoding). Each season is thus represented by a binary variable, and it is the model that determines, based on the data, the relative effect of each season on the target variable. This approach avoids the implicit assumption of an order among categories and ensures a better interpretation of the results.

Thus, a categorical variable S with p modalities will be replaced by $p - 1$ variable that have a binary output.

$$X \in \{\text{winter, summer, spring, automn}\}$$

$$\downarrow$$

$$X_{\text{spring}} \in \{0,1\}, \quad X_{\text{automn}} \in \{0,1\}, \quad X_{\text{summer}} \in \{0,1\}$$

And in this scenario, *winter* is the reference, *i.e.* it means that ifX_{spring}, X_{automn} and X_{summer} are all equal to 0, then the studied is associated to the group *winter.*

References

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

[Gilmour, 1996] Gilmour, S. G. (1996). The interpretation of mallows's cp-statistic. *Journal of the Royal Statistical Society Series D: The Statistician*, 45(1):49–56.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461 – 464.