People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
MOHAMED KHIDER UNIVERSITY, BISKRA
Faculty of Exact Sciences and Natural and Life Sciences
DEPARTEMENT OF MATHEMATICS

Master dissertation presented with a view to obtaining the Diploma:
MASTER of Mathematics

Option: Statistic

By

Abdelhak Aloui

Title :

The relationship between random variables

Examination Committee Members:

Pr.	Benatia Fatah	UMKB	President
Pr.	Brahimi Brahim	UMKB	Supervisor
Dr.	Chine Amel	UMKB	Examinator (rice)

June 2024

Dedication

I dedicate this modest work to my dear parents,

To my dear brothers and sisters,

To my friends along the path,

To all the family members,

I would like to thank all my dear professors,

For all the knowledge, wisdom, and moral values you have taught me.

Finally, I dedicate this thesis to my colleagues and all those who are dear to me.

ACKNOWLEDGEMENTS

Praise be to **Allah**, Lord of the worlds! I thank Him for the knowledge and faith He has granted me. I ask Him to benefit me with what I have learned and to make my knowledge beneficial for me and for others.

The realization of this work would not have been possible without the support of many people whom I would like to thank.

First and foremost, I thank my parents, because it is thanks to them that I was able to continue my studies to this advanced level.

I thank my supervisor, **Pr:. Brahimi Brahim**, who guided me in my project and helped me find solutions to progress. As well as Professors **Benatia Fatah** and **Chine Amel**, who accepted to preside over the defense committees.

Finally, I thank all the people and friends who contributed directly or indirectly to the completion of this work.

Contents

List of Figures

List of Tables

Introduction

In this thesis has thoroughly explored correlation and regression analyses, which are essential tools in statistical research for examining relationships between variables, predicting outcomes, and testing hypotheses. Through the three chapters, we have achieved our objective of studying the relationship between two quantitative variables by assessing their strength, direction, and predictive capacity.

In the first chapter, we detailed the different types of correlation coefficients, methods for calculating them, and interpretation of these coefficients. This section highlighted various aspects of correlation measures and their relevance in statistical analysis.

The second chapter was devoted to types of regression, particularly focusing on estimation using the Ordinary Least Squares method, properties of ordinary least squares estimators, and hypothesis testing. We demonstrated how regression can be used to model relationships between variables and provide precise estimates.

Finally, in the third chapter, we applied the concepts of simple and multiple regression to real data using R software. This practical application illustrated the theoretical principles discussed earlier and showed the importance of using software tools to conduct robust statistical analyses.

Chapter 1

Correlation

In statistical analysis, understanding the relationship between variables is fundamental for interpreting data and drawing meaningful conclusions. One of the key concepts used to explore this relationship is correlation. Correlation measures the degree of association between two variables and helps in understanding how changes in one variable relate to changes in another. By quantifying this relationship, correlation analysis provides valuable insights into the patterns and dependencies within datasets.

Correlation refers to the statistical relationship between two variables. It quantifies the extent to which changes in one variable are associated with changes in another. In other words, correlation measures the strength and direction of the linear relationship between two variables. This relationship can be positive, negative, or zero.

Correlation analysis is commonly used in various fields such as economics, finance, social sciences, and healthcare to understand the relationships between different factors or variables. It helps researchers identify patterns, make predictions, and test hypotheses based on the observed data. By providing insights into the degree of association between variables, correlation analysis plays a crucial role in statistical inference and decision-making processes.

Definition 1.0.1 *A correlation is a relationship between two variables. The data can be represented by the ordered pairs* (X, Y) *where* X *is the independent variable and* Y *is the dependent variable.*

The independent variable is the cause, its value independent of other variables in such study. The dependent variable is the effect, its value depends on changes in

the independent variable. We can use a scatter plot to determine whether a linear (straight line) correlation exists between two variables.

1.1 Types of correlation coefficient

Correlation coefficients provide a synthetic measure of the intensity of the relationship between two characteristics and its direction when this relationship is monotonic. The Pearson correlation coefficient is used to analyze linear relationships, the Spearman correlation coefficient for non-linear monotonic relationships, and the Kendall correlation coefficient for concordant and discordant relationships. There are other coefficients for non-linear and non-monotonic relationships.

Definition 1.1.1 ***(Correlation coefficient)*** *The correlation coefficient is a measure of the strength and the direction of a linear relationship between two variables, denoted by r wich represents the sample correlation coefficient.*

Note that: The range of the correlation coefficient is in $[-1, 1]$.

1.1.1 Positive correlation

- If the value of $r > 0$ then we can say it is positive correlation.
- In positive correlation both the variables changes in the same direction.

Example 1.1.1 *To study relationship between English language proficiency and understanding of mathematics, we took a dataset that includes* 50 *students and their scores in English and mathematics:*

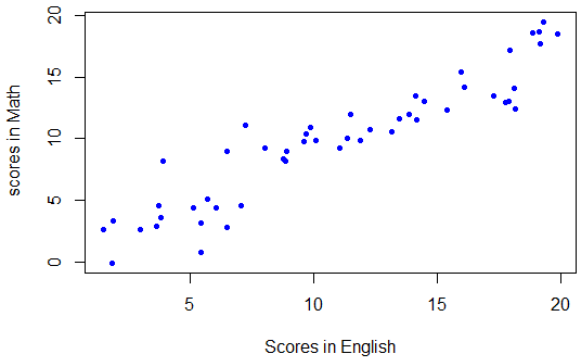

Figure 1.1: Positive correlation

A positive correlation between the two subjects would suggest that students who perform well in English tend to perform well in Math, and vice versa.

1.1.2 Negative correlation

- If the Value of $r < 0$ then we can say it is negative correlation.
- Here, Both the variables changes in the opposite direction.

Example 1.1.2 *The following data represents the number of classes missed by* 24 *different students during the first season, along with the student's final season average:*

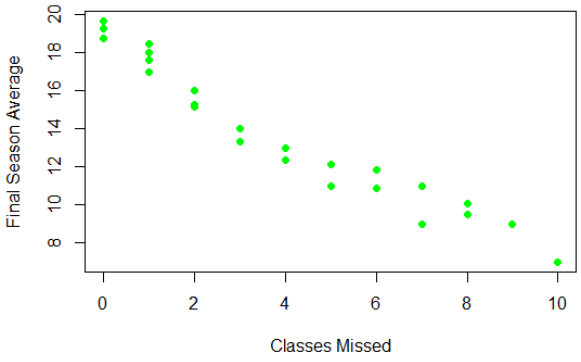

Figure 1.2: Negative correlation

In this example, we can see that as the number of classes missed increases, the final season average tends to decreases, suggesting a negative relationship between the two variables.

1.1.3 Zero correlation

- In Zero correlation the estimation of $r = 0$.
- If whole points spread all over the graph then it is called zero correlation.

Example 1.1.3 *The following data represents the number of hours of sleep per day and the ages of* 27 *different individuals:*

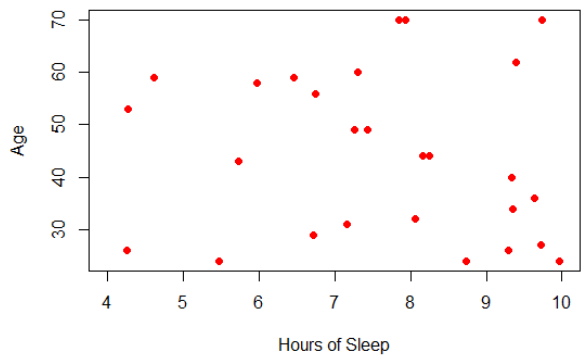

Figure 1.3: Zero correlation

In this example, we can see that people's ages are not affected by the number of hours of sleep, indicating no linear relationship between the two variables.

1.2 Methods for calculating correlation

There are several methods for calculating corelation, including: Pearson and Spearman, Kendall.

1.2.1 Pearson correlation coefficient

The pearson correlation coefficient r indicates the degree of linear correlation between two continuous variables. To calcilate it, you need three different sums of squares

(S): the sum of squares for variable X, the sum of squares for variable Y, and the sum of the cross-product of the variables XY.

- **The sum of squares for variable X is:**

$$S_{XX} = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}.$$

- **The sum of squares for variable Y** is:

$$S_{YY} = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}.$$

- **Finally, the sum of the cross-products** (S_{XY}):

$$S_{XY} = \sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}.$$

- **The correlation cofficient** (r) is:

$$r = \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}} = \frac{n\sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{\sqrt{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}\sqrt{n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2}}$$

where X represents the values of one variable,Y the values of the other variable, and n the number of observations.

Example 1.2.1 *The exam results for* 100 *students in the subjects of mathematics and physics mathematics are displayed in the following table, mathematics (x) physics (y).*

$X \diagdown Y$	3	6	10	14	18	$n_{i,0}$
3	2	1	0	0	0	3
6	5	12	3	1	0	21
10	2	10	28	5	0	45
14	0	3	12	10	1	26
18	0	0	1	2	2	5
$n_{0,j}$	9	26	44	18	3	100

Table 1.1: Distribution Table of Students' Scores in Mathematics and Physics

Using marginal distributions to calculate averages

$$m(T) = \bar{T} = \frac{1}{n}\sum_{i=1}^{n} T_i$$
$$\Rightarrow m(X) = 10.36, m(Y) = 9.2$$

We calculate the Pearson correlation coefficient

- ***The sum of squares for variable*** X ***is:*** $S_{XX} = 12.51$
- ***The sum of squares for variable*** Y *is:* $S_{YY} = 14.08$
- ***Finally, the sum of the cross-products*** (S_{XY})*:* $S_{XY} = 8.608$
- ***The correlation cofficient*** (r) *is:* $r = \frac{8.608}{\sqrt{12.51}\sqrt{14.08}} = 0.6486$

Which tells us that there is a positive relationship between mathematics and physics

1.2.2 Spearman correlation coefficient

The Spearman rank corrlation coefficient, alsodenoted by r_s, is astatistical measure used to assess the monotonic relationship between two variables using the ranks of their values themselves. The calculation processe is based on assignin ranks to each value of X, and likewise to Y, and then considering the bivariate series $\{[R(X_i), R(Y_i)]\, i = 1, 2, ..., n\}$.

The Spearman rank correlation coefficient, denoted by r_s, is defined as the Pearson correlation coefficient of this latter series.

$$r_s = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[R(X_i)R(Y_i)\right] - \left[\bar{R}_X \bar{R}_Y\right]}{\sqrt{\left\{\frac{1}{n}\sum_{i=1}^{n}\left[R(X_i) - \bar{R}_X\right]^2\right\}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[R(Y_i) - \bar{R}_Y\right]^2\right\}}}$$

We have:

$$\bar{R}_X = \bar{R}_Y = \frac{1}{n}\sum_{i=1}^{n} R(X_i) = \frac{n+1}{2},$$

and:

$$\sigma_X^2 = \sigma_Y^2 = \frac{n^2-1}{12},$$

and we have too:

$$\sum_{i=1}^{n} R(X_i)R(Y_i) = -\frac{1}{2}\sum_{i=1}^{n}\left(R(X_i) - R(Y_i)\right)^2 + \frac{1}{2}\sum_{i=1}^{n} R^2(X_i) + \frac{1}{2}\sum_{i=1}^{n} R^2(Y_i).$$

But:

$$\sum_{i=1}^{n} R^2(X_i) = \sum_{i=1}^{n} R^2(Y_i) = \frac{n\,(n+1)\,(2n+1)}{6}$$

The sum of squares of integers is calculated as follows:

$$r_s = -\frac{6\sum_{i=1}^{n}\left[R(X_i) - R(Y_i)\right]^2}{n(n^2-1)} + \frac{\frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2}{\frac{n^2-1}{12}}$$

The seconde term equals 1 after calculation, and we obtain the practical formula:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d^2}{n(n^2-1)}, \text{ avec } d_i = R(X_i) - R(Y_i)$$

Example 1.2.2 *The table below shows the sales (X) and expenses (Y) of 8 compainies in millions of dollar. Calculate the rank correlation coefficient:*

X	50	62	40	50	71	60	66	70
Y	48	40	35	30	48	55	48	60

Table 1.2: Table of Company Sales and Expenses

First, we calculate the value of d_i (the difference between the two rankings) by assigning ranks to each value of X, as well as Y,

$R(X_i)$	6.5	4	8	6.5	1	5	3	2
$R(Y_i)$	4	6	7	8	4	2	4	1
d_i	2.5	-2	1	-1.5	-3	3	-1	1

Table 1.3: Table of Rank Differencess

This implies that

$$\sum_{i=1}^{n}\left[R(X_i) - R(Y_i)\right]^2 = 33.5$$

$$n = 8 \;\Rightarrow\; n^2 - 1 = 63$$

$$r_s = 1 - \frac{(6)(33.5)}{(8)(63)} = 0.6012$$

Which tells us that there is a positive relationship between expenses and sales

1.2.3 Kendall correlation coefficient

this coefficient introduces a fundamentally different structure from the previous one, focusing on the concept of concordant and discordant pairs of observations. Assuming once more that there are no coincidences between the observed values in each marginal series, let's consider a pair of observation (X_i, Y_i) and (X_j, Y_j), where $i \neq j$, from a bivariate observed series.

a) Concordant: if $X_i < X_j$ and $Y_i < Y_j$ or if $X_i > X_j$ and $Y_i > Y_j$

b) Discordant: if $X_i < X_j$ and $Y_i > Y_j$ or if $X_i > X_j$ and $Y_i < Y_j$

- Let N_c be the number of concordant pairs and N_d be the number of dicordant pairs among the $n(n-1)/2$ pairs of observation
- The kendall rank correlation coefficient, denoted by τ, is defined by the expression

$$\tau = \frac{N_c - N_d}{N}; \quad \tau \in [-1, 1]$$
$$\tau = \frac{N_c - N_d}{n(\frac{n-1}{2})}$$

- The value of the cofficient τ ranges between 1 and -1:

$$-1 \leq \tau \leq 1$$

Example 1.2.3 *The following table illustrates how "Seven individuals underwent tests for admission to the police academy. These tests included a written exam and a physical test, and the results were as follows:": Examiners* (A, B, C, D, E, F, G)*:*

$Examiners$	$Writtentest(X)$	$Physicaltest(Y)$
A	14	18
B	5	17
C	12	13
D	10	11
E	9	15
F	16	14
G	8	12

Table 1.4: Table of Written and Physical Test Results for Police Academy Admission

Calculate the kendall correlation.

$Candidate$	A	B	C	D	E	F	G
$R(X_i)$	2	7	3	4	5	1	6
$R(Y_i)$	1	2	5	7	3	4	6
$ArrangeR(X_i)$	1	2	3	4	5	6	7
$ArrangeR(Y_i)$	4	1	5	7	3	6	2
$n_c(X_i, Y_i)$	3	5	2	0	1	0	0
$n_d(X_i, Y_i)$	3	0	2	3	1	1	0

Table 1.5: Kendall Correlation Calculation Table

For each X, the number $n_c(X_i, Y_i)$ of concordant pairs $\{(X_i, Y_i), (X_j, Y_j)\}$, is obtained by counting the ranks $R(Y_j)$ greater than $R(Y_i)$, wich $j > i$. Similarly, **xe can define the numbe**r $n_d(X_i, Y_i)$ of discordant pairs by determining the ranks $R(Y_j)$ such that $R(Y_j) < R(Y_i)$, with $j < i$. thus, we have:

$$N_c = \sum_{i=1}^{n} n_c(X_i, Y_i) = 11$$

$$N_d = \sum_{i=1}^{n} n_d(X_i, Y_i) = 10$$

$$\Rightarrow \tau_a = \frac{11 - 10}{(7)(\frac{6}{2})} = 0.057$$

Which indicates no relationship between the written test and the physical test.

1.3 Interpretation of correlation coefficients

- The interpretation of the linear correlation coefficient can be more complex than is generally believed.
- The correlation coefficient measures the strength and direction of a linear relationship between two variables, X and Y.
- The correlation coefficient laways satisfies $r \in [-1, 1]$.

$$-1 \leq r \leq 1$$

- Perfect correlation, corresponding to the case $|r| = 1$, is very rare in practice but serves as a benchmark. The closer $|r|$ is to 1, the more closely the variables X and Y are related.
- If X and Y are independent, then $r = 0$. However, the converse is not necessarily true.
- If $r = 0$, we can assert that there is no linear relationshpe between X and Y. But there may be a relationship of aother type.

Sign of the correlation coefficient:

- If the corrlation cofficient r is positive, it indicates a positive relationship between the variables X and Y, meaning they vary in the same direction: when X increases, Y increases. In the graph below, the slope of the line is positive.

$$X_i > X_j => Y_i > Y_j$$

- If the correlation coefficient r is negative, it indicates a negative relationship between the variables X and Y, meaning they vary in opposite directions: when X increases, Y decreases. In the graph below, the slope of the line is negative.

$$X_i > X_j => Y_i < Y_j$$

- The sign of r indicates the direction of the relationship, while the absolute value of r indicates the strength of the relationship, i.e., its ability to predict the values of Y based on those of X.

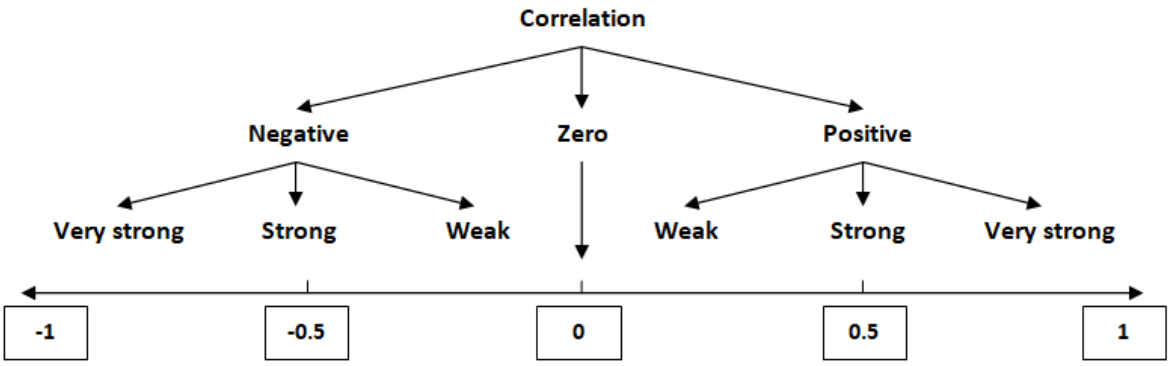

Figure 1.4: A plot illustrating the nature of the relationship

- non-linear data for example, data of the form$(y = ex)$.
- A statistical relationship, detected by the correlation coefficient or by a graph, never demonstrates a causal relationship between two variables. Causality can only be inferred from a non-statistical analysis of the data.

Chapter 2

Regression

In simple regression, we study the relationship between two variables by expressing one variable as a function of the other.In contrast, in multiple regression, we analyze the relationship between one variable and multiple other variables. If there in no relationship between the variables, it is called non-linear regression.

Implementing regression analysis requires the presence of a causal relationship between the variables included in the model. This method can be applied to quantitative data collected from n individuals.

Definition 2.0.1 *Regression is a data analysis process based on observing a random variable Y, known as the response, exogenous, dependent, or target variable, which needs to be explained (modeled) by measurements on p explanatory, control, endogenous, or independent variables. These variables can be quantitative or qualitative. This criterion determines the type of method or model to be used: linear regression, logistic regression, count data and log-linear models. The equation of regression or the regressiion model is written as follows:*

$$Y = f(X) + \epsilon$$

2.1 Types of regression

Based on the type of functions used to represent the relationship between the dependent or output variables, the regression models are categorized into four types.

The regression models are, (Simple linear regression, Multiple regression, Polynomial regression, Logistic regression)

2.1.1 Simple linear regression:

Statistical modeling

- The equation of simple linear regression is written as follows:

$$f(X) = aX + b$$

So:

$$Y_i = aX_i + b + \epsilon_i, \quad \forall i \in \{1, ..., n\}$$

a and b: The unknown parameters of the model (are the coefficients).

X_i:Is the explanatory and independent variable (input value).

Y_i: Is the explained and dependent variable (observed and random value).

n: Is the number of observations.

ϵ_i: Is the random error term in the model.

- The distribution of errors:

$$E(\epsilon_i) = 0, \quad E(\epsilon_i^2) = \sigma_\epsilon^2, \quad i = 1, ..., n$$

$$Cov(\epsilon_i, \epsilon_j) = \delta_{i,j}\sigma_\epsilon^2, \quad \forall (i, j) \text{ tel that } i \neq j$$

- The errors are assumed to be centered, have equal variance (homoscedasticity), and be uncorrelated (δ_{ij} is the Kronecker symbol, i.e., $\delta_{ij} = 1$ if $i = j$, $\delta_{ij} = 0$ if $i \neq j$).

In matrix form

- The simple linear regression model, as defined, can be reformulated using matrices:

$$Y = XB + \epsilon$$

or:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}, \quad B = \begin{bmatrix} b \\ a \end{bmatrix}$$

So:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Y represnts the response vector of size $n \times 1$.

X represents the explanatory matrix of size $n \times 2$.

ϵ represents the error vector of size $n \times 1$.

- This vector notaion will be convenient, especially for the grometric interpretation of the problem.

2.1.2 Exponential adjustment:

When calculating the equation of a linear adjustment line, it is assumed that there is a linear relationship between the variables, significantly constraining its applicability.

Often, there is a relationship between variable Y and variable X, but this relationship is not linear. The curve describing variable Y in terms of variable X is not a straight line.

However, in some cases, we can resort to a linear relationship.

If the scatter plot indicates that the functional relationship between Y and X is of the form

$$Y = \lambda \beta^X$$

Then it is noted that:

$$\ln(Y) = \ln(\lambda) + X \ln(\beta),$$

Therefore, it becomes apparent that there is a linear relationship between the variables $\ln(Y)$ and X. We define $U_i = \ln(Y_i)$ for all possible values of y_i for variable Y, and then compute the equation of the line that adjusts to U with respect to X.

This yields the following equation:

$$U = aX + b.$$

By replacing the value of U with its value $\ln(Y)$, we obtain the following functional relationship between Y and X

$$\ln(Y) = aX + b.$$

Passing to exponentials, we find:

$$Y = e^{(aX+b)} = e^b(e^a)^X.$$

So, returning to the sought relationship

$$\lambda = e^b \quad and \quad \beta = e^a.$$

2.1.3 Multiple regression:

Statistical modeling

In practic, the key steps involved in a multiple regression analysis are as follows:

- Define the dependent variable and the explanatory variables.
- Specify the nature of the relationship between the dependent variable and explanatory variables.
- Estimate the parameters of the model, then quantify its quality and verify its validity.
- if the model is retained, intrepret its significance inrelation to the proplem at hand.
- In the section, we will focus on multiple regression within the framework of the linear model. Multiple linear regression is a generalization of simple linear regression, which considres only one explanatory variable. Consider the multiple linear ragression model in the following from:

$$f(X) = f(X_1, X_2, ..., X_n) = b + a_1X_1 + a_2X_2 + ... + a_kX_k$$

So:

$$Y_i = b + a_1 X_{i,1} + a_2 X_{i,2} + ... + a_k X_{i,k} + \epsilon_i, \qquad i = 1, ..., n$$
$$Y_i = b + \sum_{j=1}^{k} a_j X_{i,j} + \epsilon_i, \quad i = 1, ..., n$$

- For each i, Y_i follows a normal distribution with mean 0 and variance σ^2. since $(\epsilon_1, \epsilon_1, ..., \epsilon_n)$ are independent and identically distributed (i.i.d), Y_i is independent of ϵ_i.

$$\epsilon_i \rightsquigarrow \mathcal{N}(0, \sigma^2)$$

$X_{i,j}$ Is the j-th explanatory (independent) variable for individual i $(j = 1, ..., k)$.

ϵ_i Is independent of X and follows a normal distribution $\mathcal{N}(0, \sigma^2)$.

b and $a_1, a_2, ..., a_k$ are the unknown paramaters of the model.

In matrix form

This model can be written in matrix form as follows:

$$Y = XB + \epsilon$$

or:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_{1,1} & \cdots & X_{1,k} \\ 1 & X_{2,1} & \cdots & X_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,k} \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}, \quad B = \begin{bmatrix} b \\ a_1 \\ \vdots \\ a_k \end{bmatrix}$$

So:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{1,1} & \cdots & X_{1,k} \\ 1 & X_{2,1} & \cdots & X_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,k} \end{bmatrix} \begin{bmatrix} b \\ a_1 \\ \vdots \\ a_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Y represents the response vector of size n.

X represents the design matrix of size $n \times (k + 1)$

ϵ represents the error vector of size n.

2.2 Estimation

In this section, we use the least squares method to estimate the unknown parameters of regression model.

2.2.1 Least squares method

The Ordinary Least Squares (OLS) method minimizes the sum of the squared differences between the observed values and the values predicted by the regression line. This method finds the best-fitting regression line (the regression line) that represents the relationship between the dependent variable Y and the independent variable X, passing through the mean point $G(m(X), m(Y))$.

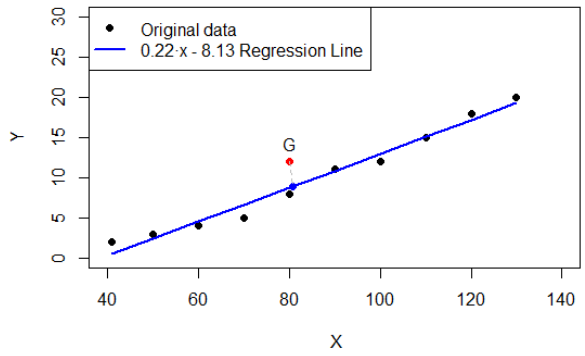

Figure 2.1: Least squares method

Case of simple linear regression

We seek values $\hat{a}$ and $\hat{b}$ of the estimators of a and b. We have:

$$Y_i = aX_i + b + \epsilon_i \Rightarrow \epsilon_i = Y_i - aX_i - b$$

$$\left(\hat{b}, \hat{a}\right) = \underset{(b,a)\in\mathbb{R}*\mathbb{R}}{\arg\min} \mathcal{S}\left(b, a\right)$$

$$\mathcal{S}\left(b, a\right) = \min_{b,a} \sum_{i=1}^{n} \epsilon_i^2 = \min_{b,a} \sum_{i=1}^{n} (Y_i - aX_i - b)^2$$

We solve the system of two equation with two unknowns

$$\nabla \mathcal{S}(b, a) = 0$$

The least squares estimators for b and a, denoted as $\hat{b}$ and $\hat{a}$, respectively, must satify

$$\frac{\partial \mathcal{S}}{\partial b} = -2\sum_{i=1}^{n}(Y_i - \hat{a}X_i - \hat{b}) = 0$$
$$\frac{\partial \mathcal{S}}{\partial a} = -2\sum_{i=1}^{n}(Y_i - \hat{a}X_i - \hat{b})X_i = 0$$

The least squares estimates of the intercept and slope in the simple linear regression model are

$$\nabla \mathcal{S}(b, a) = 0 \Rightarrow \begin{cases} \hat{b} = \bar{Y} - \hat{a}\bar{X} \\ \hat{a} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} = \frac{S_{XY}}{S_{XX}} \end{cases}$$

The fitted or regression line is therefore$\mathcal{S}(b, a)$

$$\hat{Y} = \hat{a}X + \hat{b}$$

Note that each pair of observations stissfies the relationship

$$Y_i = \hat{a}X_i + \hat{b} + \epsilon_i, \quad i = 1, ..., n$$

Algebraic perspective Given n data points $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ in $\mathbb{R}^2$, we try to find the equation of a line that passes through the n points.

This equation is $Y = aX + b$ with $b, a \in \mathbb{R}$.

b and a should be the solutions of the system $Y = XB$.

The solution in terms of least squares is

$$\left(\hat{b}, \hat{a}\right) = \left(X^t X\right)^{-1} X^t Y$$

- Where $\hat{\sigma}_\epsilon^2$, an estimator of σ_ϵ^2, is given by:

$$\hat{\sigma}_\epsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\epsilon}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

Case of exponential adjustment:

We have:

$$Y = \lambda\beta^X$$

We seek values $\hat{\beta}$ and $\hat{\lambda}$ of the estimators of β and λ.

$$\left(\hat{\lambda}, \hat{\beta}\right) = \underset{(\lambda,\beta)\in\mathbb{R}_+^*\times\mathbb{R}_+^*}{\arg\min}\ \mathcal{S}(\lambda,\beta)$$
$$\mathcal{S}(\lambda,\beta) = \min_{\lambda,\beta}\sum_{i=1}^{n}\epsilon_i^2 = \min_{\lambda,\beta}\sum_{i=1}^{n}(\ln(Y_i) - \ln(\beta)X_i - \ln(\lambda))^2$$

So, returning to the sought relationship:

$$\lambda = e^b \quad and \quad \beta = e^a.$$

We solve the system of two equation with two unknowns

$$\nabla\mathcal{S}(\lambda,\beta) = 0$$

The least squares estimators for λ and β, denoted as $\hat{\lambda}$ and $\hat{\beta}$, respectively, must satify

$$\nabla\mathcal{S}(\lambda,\beta) = 0 \Rightarrow \begin{cases} \ln(\hat{\lambda}) = \bar{U} - \ln(\hat{\beta})\bar{X} \\ \ln(\hat{\beta}) = \frac{\sum_{i=1}^{n} X_iU_i - n\bar{X}\bar{U}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} = \frac{S_{XU}}{S_{XX}} \end{cases}, \text{Such as } \ln(Y) = U$$

By replacing the value of U with its value $\ln(Y)$, we conclude that:

$$\hat{\lambda} = e^{\bar{U} - \ln(\hat{\beta})\bar{X}}$$
$$\hat{\beta} = e^{\frac{S_{XU}}{S_{XX}}}$$

Case of multiple regression

The least squares estimator $\hat{B}$ of B is defined as follows:

$$\hat{B} = \arg\min_{B}\sum_{i=1}^{n}\epsilon_i^2$$

We then seek the statistic $\hat{B} = (\hat{b}, \hat{a}_1, ..., \hat{a}_k)$ that minimizes

$$\mathcal{S}(B) = \mathcal{S}(b, a_1, ..., a_k) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - b - \sum_{i=1}^{n} a_j X_{i,j} \right)^2$$

To calculate $\hat{B}$, we propose the following matrix expression:

$$\begin{aligned}
\mathcal{S}(B) &= \epsilon^t \epsilon \\
&= \|Y - XB\|^2 \\
&= (Y - XB)^t (Y - XB) \\
&= (Y^t - X^t B^t)(Y - XB) \\
&= Y^t Y - Y^t XB - B^t X^t Y + B^t X^t XB
\end{aligned}$$

And since $Y^t YB = B^t X^t Y \in \mathbb{R}$, we obtain

$$\mathcal{S}(B) = Y^t Y - 2B^t X^t Y + B^t X^t XB$$

By taking the derivative of $\mathcal{S}(B)$ with respect to each parameter $b, a_1, ..., a_k$, we obtain the system of equations:

$$\begin{aligned}
-2X^t Y + 2X^t X\hat{B} &= 0 \\
&\Rightarrow X^t Y - X^t X\hat{B} = 0 \\
&\Rightarrow X^t Y = X^t X\hat{B} \\
&\Rightarrow (X^t X)^{-1} X^t Y = (X^t X)^{-1} (X^t Y)\hat{B}
\end{aligned}$$

And from there, we finally deduce

$$B = (X^t X)^{-1} X^t Y$$

- The variance of the error σ_ϵ^2 is estimated without bias by:

$$\hat{\sigma}_\epsilon^2 = \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n - k - 1} = \frac{\hat{\epsilon}^t \hat{\epsilon}}{n - k - 1}$$

2.3 Properties of the OLS estimators

2.3.1 Properties of simple linear regression

The statistical properties of the least squares estimators $\hat{b}$ and $\hat{a}$ can be easily descibed. Assuming the error term ϵ in the model $Y = b + aX + \epsilon$ is a random variable with a mean of zero and a variance of σ^2. since the values of X are fixed, Y is a random variable with mean of $\mu = b + aX$ and variance σ^2. Therefor, the values of $\hat{b}$ and $\hat{a}$ depend on the observed y's, making the least squares estimators of the regression coefficients random variables. We will examine the bias and variance properties of $\hat{b}$ and $\hat{a}$.

Starting with $\hat{b}$, as it is a linear combination of the observations Y_i, we can use the properties of expectation to show that the expected value of $\hat{a}$ is:

Proposition 2.3.1 *Unbiased estimators*

$$E(\hat{a}) = a, \qquad E(\hat{b}) = b$$

Preuve. for $\hat{a}$: $\hat{a} = \frac{S_{XY}}{S_{XX}}$

$$\begin{aligned}
&\Rightarrow \hat{a} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
&= a + \frac{\sum_{i=1}^{n}(X_i - \bar{X})(\epsilon_i - \bar{\epsilon})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
&= a + \frac{\sum_{i=1}^{n}(X_i - \bar{X})\epsilon_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
&\Rightarrow E(\hat{a}) = a
\end{aligned}$$

■

and $\hat{b}$: $\hat{b} = \bar{Y} - \hat{a}\bar{X}$

$$\begin{aligned}
&\Rightarrow E(\hat{b}) = E(\bar{Y} - \hat{a}\bar{X}) = E(\bar{Y}) - \bar{X}E(\hat{a}) = b + \bar{X}a - \bar{X}a \\
&\Rightarrow E(\hat{b}) = b
\end{aligned}$$

Proposition 2.3.2 *Variances of the estimators* $\left\{\hat{b}, \hat{a}\right\}$

$$Var(\hat{a}) = \frac{\sigma_\epsilon^2}{S_{XX}}, \qquad Var(\hat{b}) = \frac{\sigma_\epsilon^2 \sum_{i=1}^{n} X_i^2}{nS_{XX}}$$

Preuve.

$$\begin{aligned}
Var(\hat{a}) &= E\left[(\hat{a}-E(\hat{a}))^2\right] = E\left[(\hat{a}-a)^2\right] \\
&= E\left[\left(a+\frac{\sum_{i=1}^{n}(X_i-\bar{X})\epsilon_i}{S_{XX}^2}-a\right)^2\right] \\
&= \frac{1}{S_{XX}^2}E\left[\left(\sum_{i=1}^{n}(X_i-\bar{X})\epsilon_i\right)^2\right] \\
&= \frac{1}{S_{XX}^2}E\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\epsilon_i^2+\sum_{i=1}^{n}(X_i-\bar{X})\epsilon_i(X_j-\bar{X})\epsilon_j\right] \\
&\Rightarrow Var(\hat{a}) = \frac{\sigma_\epsilon^2}{S_{XX}}
\end{aligned}$$

$$\begin{aligned}
Var(\hat{b}) &= E\left[(\hat{b}-b)^2\right] = E\left[\left(\bar{Y}-\hat{b}-\hat{a}\bar{X}\right)^2\right] \\
&= E\left[\left(b+a\bar{X}+\bar{\epsilon}-\hat{a}\bar{X}-b\right)^2\right] = E\left[\left((a-\hat{a})\bar{X}+\bar{\epsilon}\right)^2\right] \\
&= E\left[\left((a-\hat{a})^2\bar{X}^2+\bar{\epsilon}^2+2\bar{\epsilon}(a-\hat{a})\bar{X}\right)\right] \\
&= \bar{X}^2E\left[(\hat{b}-b)^2\right]+E(\bar{\epsilon}^2)+2\bar{X}E\left[\bar{\epsilon}(a-\hat{a})\right] \\
&= \bar{X}^2Var(\hat{a})+E\left[\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right)^2\right]
\end{aligned}$$

■

$$= \frac{\bar{X}^2\sigma_\epsilon^2}{S_{XX}}+\frac{1}{n}\sigma_\epsilon^2, \quad E\left[\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right)^2\right] = \frac{1}{n^2}\sum_{i=1}^{n}\left(E\left(\epsilon_i^2\right)\right)+\frac{1}{n}\sum_{i=1}^{n}E\left(\epsilon_i\right) = \frac{1}{n}\sigma_\epsilon^2$$

$$= \frac{n\bar{X}^2\sigma_\epsilon^2+\sigma_\epsilon^2S_{XX}}{nS_{XX}} = \frac{n\bar{X}^2\sigma_\epsilon^2+\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)\sigma_\epsilon^2}{nS_{XX}} = \frac{n\bar{X}^2\sigma_\epsilon^2+\sigma_\epsilon^2\sum_{i=1}^{n}X_i^2-n\bar{X}^2\sigma_\epsilon^2}{nS_{XX}}$$

$$\Rightarrow Var(\hat{b}) = \frac{\sigma_\epsilon^2\sum_{i=1}^{n}X_i^2}{nS_{XX}}$$

Proposition 2.3.3 *Covariance of the estimators* $\left(\hat{b},\hat{a}\right)$

$$Cov(\hat{b},\hat{a}) = -\frac{\sigma_\epsilon^2\bar{X}}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

Preuve.

$$Cov(\hat{b},\hat{a}) = Cov\left(\bar{Y} - \hat{a}\bar{X}, \hat{a}\right) = Cov\left(\bar{Y},\hat{a}\right) - \bar{X}Var(\hat{a})$$

The variance between $\bar{Y}$ and $\hat{a}$ is written as:

$$Cov\left(\bar{Y},\hat{a}\right) = Cov\left(\sum_{i=1}^{n}\frac{Y_i}{n}, \frac{\sum_{i=1}^{n}(X_i-\bar{X})\epsilon_i}{S_{XX}^2}\right) = 0$$

We finally conclude:

$$Cov(\hat{b},\hat{a}) = -\frac{\sigma_\epsilon^2\bar{X}}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

■

2.3.2 Properties of multiple regression

The statistical properties of the least squares estimators $\hat{b},\hat{a}_1,...,\hat{a}_k$ canbe easily determined under ceertain assumption on the error terms $\epsilon_1,...,\epsilon_n$ in the regression model.We assume that the errors are statistically independent with mean zero and variance σ^2.Under these assumptions, the least squares estimators $\hat{b},\hat{a}_1,...,\hat{a}_k$ are unbiased estimators of the regression coefficients $b,a_1,...,a_k$. This property can be shown as follows:

Proposition 2.3.4 *The estimator $\hat{B}$ est best unbiased estimator of B.*

$$E(\hat{B}) = B$$

Preuve.

$$\begin{aligned}\hat{B} &= (X^tX)^{-1}X^tY \\ &= (X^tX)^{-1}X^t(XB+\epsilon) \\ &= B + (X^tX)^{-1}X^t\epsilon\end{aligned}$$

■

So:

$$\begin{aligned}E\left[\hat{B}\right] &= B + E\left[\left(X^tX\right)^{-1}X^t\epsilon\right] \\ &= B + \left(X^tX\right)^{-1}X^tE\left[\epsilon\right]\end{aligned}$$

And Since $E[\epsilon] = 0$,

$$E(\hat{B}) = B$$

Proposition 2.3.5 *The variance of $\hat{B}$ is:*

$$Var(\hat{B}) = \sigma_\epsilon^2 (X^t X)^{-1}$$

Preuve.

$$\begin{aligned}
Var(\hat{B}) &= E\left[\left(\hat{B} - E\left(\hat{B}\right)\right)\left(\hat{B} - E\left(\hat{B}\right)\right)^t\right] \\
&= E\left[\left(\hat{B} - B\right)\left(\hat{B} - B\right)^t\right] \\
&= E\left[\left(X^t X\right)^{-1} X^t \epsilon \left(\left(X^t X\right)^{-1} X^t \epsilon\right)^t\right] \\
&= E\left[\left(X^t X\right)^{-1} X^t \epsilon \epsilon^t X \left(X^t X\right)^{-1}\right] \\
&= \left(X^t X\right)^{-1} X^t E\left(\epsilon \epsilon^t\right) X \left(X^t X\right)^{-1} \\
&= \left(X^t X\right)^{-1} X^t \sigma_\epsilon^2 I_n X \left(X^t X\right)^{-1} \\
&= \sigma_\epsilon^2 \left(X^t X\right)^{-1} X^t X \left(X^t X\right)^{-1}
\end{aligned}$$

$$\Rightarrow Var(\hat{B}) = \sigma_\epsilon^2 (X^t X)^{-1}$$

■

2.4 Hypothesis testing

2.4.1 Hypothesis tests in simple linear regression

The complete assumption are that the errors are normally and independently distributed with mean 0 and variance σ^2, abbreviated i.i.d $\mathcal{N}(0, \sigma^2)$.

Use of t-test

Suppose that we wish to test the hypothesis that the slop equals a constant, say, β_0. The appropriate hypotheses are

$$H_0 : a = \beta_0, \qquad H_1 : a \neq \beta_0$$

Assuming a two-sided alternative, with errors ϵ_i bring i.i.d $\mathcal{N}(0, \sigma^2)$, the observations Y_i are directly assumed to be i.i.d $\mathcal{N}(b + aX_i, \sigma^2)$.

The estimator $\hat{a}$ is a linear combination of independent normal random variables, making it $\mathcal{N}(a, \frac{\sigma^2}{S_{XX}})$ based on the bais and variance properties of the slope.

Additionally, $\frac{(n-2)SSE}{\sigma^2} \rightsquigarrow \mathcal{X}^2_{n-2}$, and a is independent of σ^2.

These properties lead to the statistic

Test Statistic for the slope

$$T_0 = \frac{\hat{a} - \beta_0}{\sqrt{\hat{\sigma}^2 / S_{XX}}}$$

follows the t distribution with $n - 2$ degrees of freedom under $H_0 : a = \beta_0$.

We would reject $H_0 : a = \beta_0$ if

$$|t_0| > t_{\frac{\alpha}{2}, n-2}$$

We can write the test statistic as follows.

$$T_0 = \frac{\hat{a} - \beta_0}{se(\hat{a})}$$

A similar procedure can be used to test hypotheses about the intercept.

To test

$$H_0 : b = \beta_1, \qquad H_1 : b \neq \beta_1$$

Test Statistic for the Intercept

$$T_0 = \frac{\hat{b} - \beta_1}{\sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right]}} = \frac{\hat{b} - \beta_1}{se(\hat{b})}$$

And reject the null hypothesis if the computed value of this test statistic, t_0, is such that $|t_0| > t_{\frac{\alpha}{2}, n-2}$.

Note that the denominator of the test statistic in equation is the standard error of the intercept.

A very important special case of the hypotheses of equation is

$$H_0 : a = 0, \qquad H_1 : a \neq 0$$

these hypotheses pertain to the significance of regression.

Not rejecting $H_0 : a = 0$ equates to concluding that there is no linear relationship between X and Y.

It should be noted that this may imply that X has little value in explaining the variance in Y and that the best estimator of Y for X is $Y = \hat{Y}$ or that true relationship between X and Y is not linear.

Conversly, if $H_0 : a = 0$ is rejected, this implies that X is valuble in explaining the variability in Y.

Rejecting $H_0 : a = 0$ could be mean either that the straight-line model is adequate or that, although there is a linear effect of X, better results could be obtianed by adding higher oder polynomial terms in X.

Analysis of variance approach to test significance of regression

A method called the analysis of variance can be used to test for significance of regression.

The procedure partitions the total variability in the response variable into meaningful components as the basis for the test.

Analysis of variance Identity The analysis of variance identity is as follows:

$$\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

The two components on the right-hand-side of Equation measure, respectively, the amount of variability in Y_i accounted for by the regression line and the residual variation left unexplained by the regression line.

We usually call $SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$ the error sum of squares and $SSR = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2$ the regression sum of squares and $SST = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$ the total corrcted sum of squares.

Symbolically, Equation may be written as

$$SST = SSR + SSE$$

The analysis of variance table is presented in the following format to summarize the results of the regression test.

Source of Varia.	Degr. of Free.	Sum of Squares	Mean Square	F
Regression	1	$SSR = \hat{a}S_{XY}$	$MSR = \frac{SSE}{1}$	$F = \frac{MSR}{MSE}$
Error	$n-2$	$SSE = SST - \hat{a}S_{XY}$	$MSE = \frac{SSE}{n-2}$	
Total	$n-1$	SST		

Table 2.1: Analysis of variance table for testing significance of regression in simple linear regression.t

2.4.2 Hypothesis tests in Multiple linear regression

In multiple linear regression problems, certian tests of hypotheses about the model parameters are useful in measuring model adequacy.

In this section, we describe several important hypothesis-testing procedures.

As in the simple linear regression case, hypothesis testing requires that the error terms ϵ_i in the regression model are normally and independently distributed with mean zero and variance σ^2.

Test for significance of regression

The test for significance of regression is a test to determine whether a linear relationship exists between the response variable Y and a subset of the regressor variables $X_1, X_2, ..X_k$.

The appropriate hypotheses are

$$H_0 : a_1 = a_2 = ... = a_k = 0$$
$$H_0 : a_k \neq 0 \text{ for at least one k}$$

Hypotheses for ANOVA Test

Rejecting the null hypothesis $H_0 : a_1 = a_2 = ... = a_k = 0$ indicates that at least one of the regressor variables $X_1, X_2, ..X_k$ contributes significantly to the model.

the test for the significance of regression is an extension.

The total sum of squares, SST, is divided into a sum of squares, attributable to the model or regression, and a sum of squares attributable to error.

$$SST = SSR + SSE$$

The test statistic for $H_0 : a_1 = a_2 = ... = a_k = 0$ is

Source of Varia.	Degr. of Free.	Sum of Squares	Mean Square	F
Regression	k	SSR	MSR	$F = \frac{MSR}{MSE}$
Error	$n-k-1$	SSE	MSE	
Total	$n-1$	SST		

Table 2.2: Analysis of variance for testing significance of regression in multiple regression.

Chapter 3

Application

R is a free and open-source software and a statistical processing language that has been introduced here. The R language (R Development Core Team, 2013) is object-oriented, like Python or Ruby. One of R's advantages is its ability to communicate through written scripts, as it is commonly used in console mode. This avoids concerns about technical issues arising from the variety of operating system versions and the need for screenshots, but most importantly, it allows for communicating calculations and statistical analyses in just a few lines of text.

3.1 Simple linear regression with R

Example 3.1.1 *As part of research on the duration of the growing season in the mountains, meteorological stations are installed at different altitudes. The average temperature (variable Y in degrees Celsius) and the altitude (variable X in meters) of each station are given in the table below (see [8]):*

Altitude	1040	1230	1500	1600	1740	1950	2200	2530	2800	3100
Temperature	7.4	6	4.5	3.8	2.9	1.9	1	-1.2	-1.5	-4.5

Table 3.1: Table of Study on the Duration of the Growing Season in the Mountains

Scatter plot:

The `plot()` function produces different default graphics depending on the type of data. Thus, if we are dealing with the relationship between two variables, we get a scatter plot.

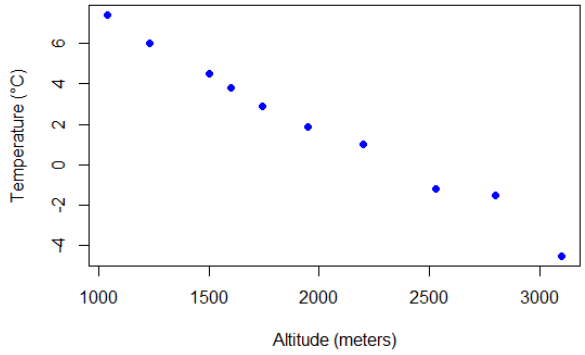

Figure 3.1: Relationship Between Altitude and Temperature in the Mountains

Linear correlation expresses the strength of the relationship between two variables. The function cor(.,.) is used to calculate the correlation coefficient between these two variables.

```
 # Calculating correlation coefficients
> cor(altitude, temperature, method="pearson")
[1] -0.9936053
> cor(altitude, temperature, method="spearman")
[1] -1
> cor(altitude, temperature, method="kendall")
[1] -1
```

The correlation coefficient is an indicator of this relationship. We seek to explain the variations of Y through the variations of a linear function of X, based on n observations of each of the variables. In other words, to adjust the data according to the model.

$$Y_i = aX_i + b, \quad i = 1, 2, ..., n$$

lm() Function

This function creates the relationship model between the predictor and the response variable.

Syntax

The basic syntax for `lm()` function in linear regression is -

```
lm(formula,data)
```

Following is the description of the parameters used -

formula is a symbol presenting the relation between x and y.

Data is the vector on which the formula will be applied.

Create relationship model & get the coefficients

```
Call:
lm(formula = y ~x)
Coefficients:
(Intercept)      altitude
 12.5953      -0.0054
```

$$Y_i = aX_i + b, \quad i = 1, 2, ..., n \quad With, \quad \hat{a} = -0.0054, \hat{b} = 12.5953$$

Get the summary of the relationship

```
Call:
lm(formula = y ~x)
Residuals:
 Min       1Q        Median     3Q        Max
-0.4613  -0.2289  -0.1283     0.1583     0.9290
Coefficients:
            Estimate Std.  Error t value Pr(> | t | )
(Intercept) 12.5953  0.4466   28.20    2.70e-09
  altitude    -0.0054   0.0002  -24.89   7.26e-09
---
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'  0.1 ' ' 1
Residual standard error:  0.4395 on 8 degrees of freedom
Multiple R-squared:  0.9873,    Adjusted R-squared:  0.9857
F-statistic:  619.5 on 1 and 8 DF, p-value:  7.26e-09
```

Relationship Between Altitude and Temperature in the Mountains with Regression Line

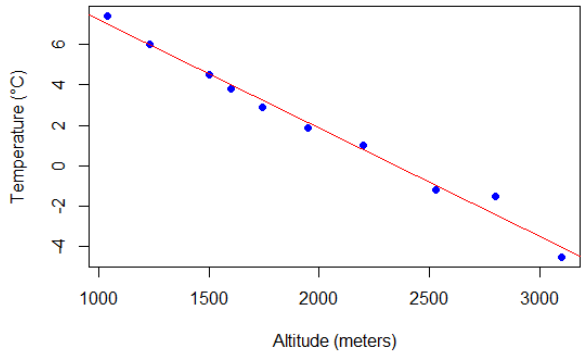

Figure 3.2: Relationship Between Altitude and Temperature in the Mountains with Regression Line

The analysis reveals a strong negative linear relationship between altitude and temperature. As altitude increases, temperature decreases. The correlation coefficients support this strong inverse relationship. The linear regression model further quantifies this relationship, showing that for every meter increase in altitude, the temperature decreases by approximately 0.0054 degrees Celsius. The high R-squared value indicates that the model fits the data very well, and the statistical significance of the coefficients suggests that the relationship is robust.

3.2 Multiple regression with R

Multiple regression is an extension of linear regression into relationship between more than two variables. In simple linear relation we have one predictor and one response variable, but in multiple regression we have more than one predictor variable and one response variable.

The general mathematical equation for multiple regression is

$$Y = b + a_1X_1 + a_2X_2 + ... + a_nX_n,$$

Following is the description of the parameters used -

Y is the response variable.

$b, a_1, ..., a_n$ are the coefficients.

$X_1, ..., X_n$ are the predictor variables.

com.	1	2	3	4	5	6	7	8	9	10	11	12	13
X_1	7	1	11	11	7	11	3	1	2	21	1	11	10
X_2	26	29	56	31	52	55	71	31	54	47	40	66	68
X_3	6	15	8	8	6	9	17	22	18	4	23	9	8
X_4	60	52	20	47	33	22	6	44	22	26	34	12	12
Y	78.5	74.5	104.3	87.6	95.9	109.2	102.7	72.5	93.1	115.9	83.8	113.3	109.4

Table 3.2: Table of the Relationship between Cement Solidification Heat and Its Chemical Compositions

We create the regression model using the `lm()` function in **R**. The model determines the value of the coefficients using the input data. Next we can predict the value of the response variable for a given set of predictor variables using these coefficients.

lm() Function

This function creates the relationship model between the predictor and the response variable.

Syntax

The basic syntax for `lm()` function in multiple regression is -

```
lm(y ~x1+x2+x3...,data)
```

Following is the description of the parameters used -

formula is a symbol presenting the relation between the response variable and predictor variables.

data is the vector on which the formula will be applied.

Example 3.2.1 *A certain kind of cement release the heat y (cal/g) when it is solidified is related to four chemical compositions (see [6]). The data are shown in table*

Creating the Regression Model

Extracting the Coefficients

After creating the model, we extract the coefficients (intercept and coefficients for each predictor variable):

```
Call:
```

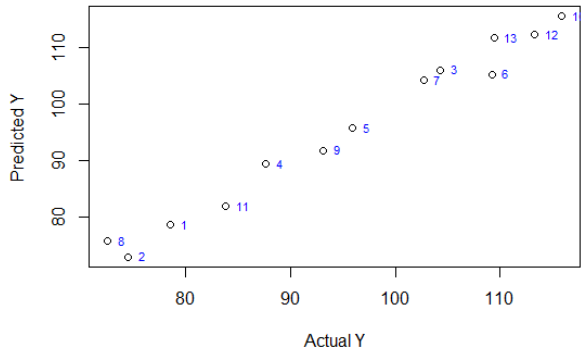

Figure 3.3: Prediction of Cement Heat Release Values

```
lm(formula = Y ~X1 + X2 + X3 + X4, data = compositions)

Coefficients:

(Intercept)      X1       X2       X3          X4

 63.6759      1.5345   0.4974   0.0872    -0.1556
```

Creating the Mathematical Equation of the Model

```
[1] "Y = 63.68 + 1.53 * X1 + 0.5 * X2 + 0.09 * X3 + -0.16 *X4"
```

Predicting New Values

	Actual_Y	Predicted_Y
1	78.5	78.54072
2	74.5	72.85545
3	104.3	105.9978
4	87.6	89.3625
5	95.9	95.6734
6	109.2	105.2765
7	102.7	104.1458
8	72.5	75.7053
9	93.1	91.7536
10	115.9	115.5838
11	83.8	81.8248
12	113.3	112.3036
13	109.4	111.6767

Table 3.3: Table of Predicting New Values

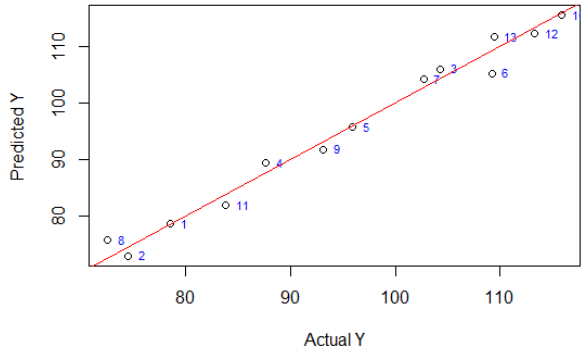

Figure 3.4: Prediction of Cement Heat Release Values with Regression Line

The analysis shows that there is a positive relationship between the factors X_1, X_2, X_3, X_4 and the dependent variable Y. In other words, as the values of X_1, X_2, X_3, X_4 increase, the value of Y also increases. The high R-squared value indicates that the

multiple linear model explains a large proportion of the variance in Y, making it a good fit for the data used.

Conclusion

In conclusion, correlation and regression analyses have proven their utility in providing valuable insights into the underlying dynamics of the studied variables. The use of these techniques has not only enabled a better understanding of the relationships between variables but also facilitated informed predictions and rigorous hypothesis testing. This thesis thus contributes to a better understanding of statistical methods and their practical application, providing a solid foundation for future research in this field.

Bibliography

[1] BENABDERREZAK, A. (2018). Régresion linéaire multiple: Théorie et Applications, UNIVERSITÉ MOHAMED KHIDER, BISKRA.

[2] CHERFAOUI, M. (2016/2017). Statistiques Appliquées à l'Expérimentation En Science Biologique, UNIVERSITÉ MOHAMED KHIDER, BISKRA.

[3] CHOUQUTE, C. (2009/2010). Laboratoire de Statistique et Probabilités, UNIVERSITÉ PAUL SABATIER, TOULOUSE.

[4] Douglas C. M, George C. R. (2013). Applied Statistics and Probability for Engineers, Arizona state University.

[5] Li, X. M. (2013). Multivariate Regression Analysis Using Statistics with R. Advanced Materiales Research, 765, 1572-1575.

[6] MANSOURI, B. (2008). Analyse des Données, UNIVERSITÉ MOHAMED KHIDER, BISKRA.

[7] Montgomery, D. C. (2003). Applied Statistics and Probability for Engineers.

[8] Saporta, G. (2006). Probabilités, analyse des données et statistique. Editions technip.

[9] TOUFOUTA, N, BOUAMAR, Z. (2016/2017). Modèles linéaires généralisés (estimations et prédictions), Mémoire de Master, UNIVERSITÉ A. MIRA de Béjaia.

[10] Veysseyre, R. (2006). Statistique et probabilités pour l'ingénieur. 2eme édition Dunod.

[11] YAHIA, D, BENAMEUR, S. (2021/2022). Analyse de la Variance, Cours de L'analyse des données, UNIVERSITÉ MOHAMED KHIDER, BISKRA.

Notation

r_s	:	Spearman's Rho
τ	:	Kendall's Tau coefficient
N_c	:	number of concordant pairs
N_d	:	number of dicordant pairs
$R(X_i)$	:	The rank of the value X_i
$\left\{\hat{b}, \hat{a}\right\}$	:	the estimated values (b, a)
ϵ	:	Is the random error termin the model.
$\delta_{i,j}$	:	Kronecker delta, which equals 1 if $i = j$ and 0 otherwise
$\mathcal{N}(0, \sigma^2)$	:	Normal distribution with mean 0 and variance σ^2.
$E(X)$	:	Mathematical expectation of a random variable X.
$Var(X)$	:	Variance of a random variable X.
$Cov(X, Y)$	:	Covariance between random variables X and Y.
$\mathcal{S}(b, a)$	:	Sum of Squares due to Regression
SSR	:	Sum of Squares due to Regression
SSE	:	Sum of Squares due to Error
SST	:	Sum of Squares due to Total
MSR	:	Mean Square due to Regression
MSE	:	Mean Square due to Error
T_0	:	Test Statistic
$\mathbb{R}$	:	the set of real numbers

Résumé

Le sujet de recherche "Sur la dépendance entre les variables aléatoires" vise à étudier la relation entre les variables aléatoires en utilisant des techniques de corrélation et de régression. Nous passerons en revue les types de coefficients de corrélation tels que Pearson, Spearman et Kendall, et expliquerons comment les calculer et les interpréter. De plus, nous explorerons différents modèles de régression, y compris la régression linéaire simple et la régression multiple, en mettant l'accent sur l'estimation des coefficients en utilisant la méthode des moindres carrés et le test des hypothèses pour assurer la validité du modèle. Enfin, en utilisant le programme R, nous fournirons des interprétations graphiques des résultats théoriques appliqués à des exemples de régression simple et multiple.

ABSTRACT

The research topic "On the Dependence between Random Variables" aims to study the relationship between random variables using correlation and regression techniques. We will review types of correlation coefficients such as Pearson, Spearman, and Kendall, and explain how to calculate and interpret them. Additionally, we will explore different regression models, including simple linear regression and multiple regression, focusing on estimating coefficients using the least squares method and testing hypotheses to ensure model validity. Finally, using the R program, we will provide graphical interpretations of the theoretical results applied to examples of simple and multiple regression.

.

ملخص

يتضمن موضوع البحث "حول التبعية بين المتغيرات العشوائية" التي تهدف إلى دراسة العلاقة بين المتغيرات العشوائية باستخدام تقنيات الارتباط والانحدار. سنقوم بالتذكير بأنواع معاملات الارتباط مثل: بيرسون، سبيرمان، وكيندال، ونقوم بتوضيح كيفية حسابها وتفسيرها. كما نستعرض نماذج الانحدار المختلفة بما في ذلك الانحدار الخطي البسيط والانحدار المتعدد، ونركز على تقدير المعاملات باستخدام طريقة المربعات الصغرى واختبار الفرضيات لضمان صحة النماذج. أخيرًا، باستخدام البرنامج R سنقوم بإعطاء تفسيرات بيانية للنتائج النظرية المطبقة على أمثلة الانحدار البسيط والانحدار المتعدد.