Crash Course - Statistics

Beniamino Sartini

2025-07-13

Table of contents

Introduction

General notations

Probability

- Ω abstract set representing the sample space of a random experiment. The elements in $\omega \in \Omega$ are the possible outcomes of the experiment.
- $\mathcal{P}(\Omega)$: power set of Ω, the set of all possible subsets of Ω.

- Most of subsets $A, B, ...$ will be thought as *events*.
- Collection of subsets $\mathcal{A}, \mathcal{B}, ...$
- The empty set $\emptyset$.
- $\mathcal{B}$ is a σ-field, usually connected with the sample space Ω.
- $\mathcal{B}(\mathbb{R})$ is the Borel σ-field of $\mathbb{R}$.
- $\mathbb{P}$ is a probability measure function $\mathbb{P} : \mathcal{B} \to [0, 1]$.
- $(\Omega, \mathcal{B}, \mathbb{P})$ is a probability space.
- $\sqcup$ is a shortcut to denote a disjoint union, for example writing $A \sqcup B$ means that the sets A and B are disjoint, while writing $A \cup B$ means that the sets A and B are not disjoint.

Linear algebra

- Bold and capital letter stands for a matrix, e.g. $\mathbf{X}$.
- Bold with small letter stands for a vector, e.g. $\mathbf{x}$.
- Small letter not bold denotes a scalar, e.g. x.

Part I

Probability

1 Set and events

In probability, an event is interpreted as a collection of possible outcomes of a random experiment.

Definition 1.1. (Random experiment)
A **random experiment** is any repeatable procedure that results in one out of a well-defined set of possible outcomes.

- The set of possible outcomes is called *sample space* and denoted as Ω.
- A set of zero or more outcomes is an *event*.
- A map that goes from events to probabilities is called a probability function and it is denoted as $\mathbb{P}$. Together, sample space, event space and probability function characterize a **random experiment**.

1.1 Set operations

There are several definitions related to sets and their operation.

Definition 1.2. (Complementation)
The complement of a set A is denoted by A^c and represents the set of elements that do not belong to A, i.e.

$$A^c = \{\omega \in \Omega : \omega \notin A\}. \tag{1.1}$$

Definition 1.3. (Containment)
A set A is said to be **contained** in a set B if every element of A is also an element of B. Formally,

$$A \subset B \iff \omega \in A \implies \omega \in B \quad \forall \omega \in \Omega. \tag{1.2}$$

Definition 1.4. (Equality)
Given two sets, A is equal to B, written $A = B$, if and only if every element of A is an element of B and every element of B is an element of A. Formally,

$$A \subset B \quad \text{and} \quad B \subset A. \tag{1.3}$$

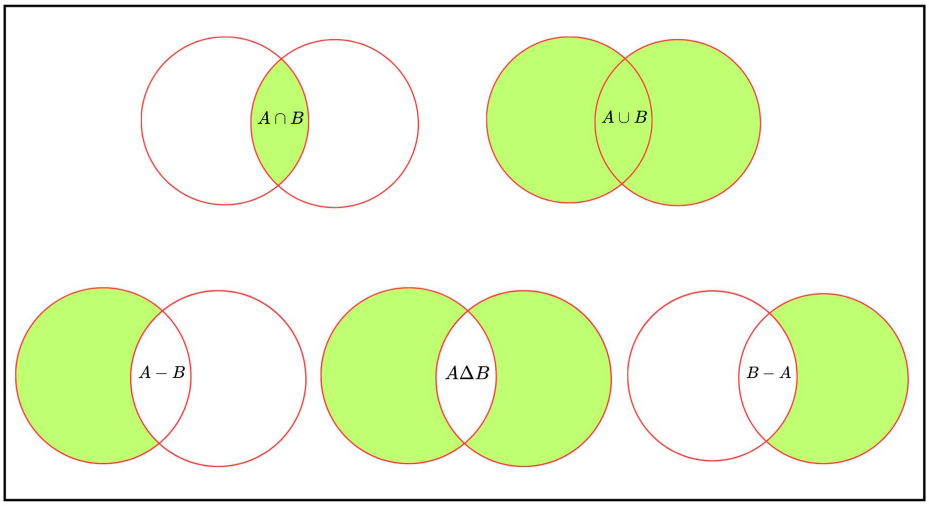


Figure 1.1: Elementary set operations.

Let's now state some elementary operations between sets.

Definition 1.5. (**Union**)
The **union** of two sets, written $A \cup B$, is the set of ω that belongs either to A or B, i.e.

$$A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}. \tag{1.4}$$

As a consequence of the definition of the union the following relations holds true, i.e.

$$\begin{aligned} A \cup A &= A \quad & A \cup \Omega &= \Omega \\ A \cup \emptyset &= A \quad & A \sqcup A^c &= \Omega \end{aligned}$$

Definition 1.6. (**Intersection**)
The **intersection** of A and B is written $A \cap B$ and is the set of elements that belongs at the same time to A and B.

$$A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}.$$

As a consequence of the definition of the intersection the following relations holds, i.e.

$$\begin{aligned} A \cap A &= A \quad & A \cap \Omega &= A \\ A \cap \emptyset &= \emptyset \quad & A \sqcap A^c &= \emptyset \end{aligned}$$

Moreover, let's state the **distributive laws** of the union and the intersection, i.e.

$$\begin{aligned} &\textbf{Intersection}. \quad && (A \cup B) \cap C = (A \cap C) \cup (B \cap C) \\ &\textbf{Union}. \quad && (A \cap B) \cup C = (A \cup C) \cap (B \cup C) \end{aligned} \tag{1.5}$$

And the De Morgan's laws:

$$\begin{aligned} &\textbf{Intersection.} && (A\cap B)^c = (A^c \cup B^c) \\ &\textbf{Union.} && (A\cup B)^c = (A^c \cap B^c) \end{aligned} \tag{1.6}$$

Definition 1.7. (Difference)
The **difference** between two sets A and B, written $A-B$ (or also A/B), is the set of elements of A that do not belong to B. Formally

$$A - B = A\cap B^c = \{\omega \in \Omega : \omega \in A \text{ and } \omega \notin B\}. \tag{1.7}$$

⚠ Disjoint reppresentation of a set

Given two set A and B, then each one can be written as the union of disjoint sets. In fact, their union can be decomposed into the union of three disjoint sets, i.e.

$$A \sqcup B = (A\cap B)\sqcup(A\cap B^c)\sqcup(A^c\cap B), \tag{1.8}$$

and therefore for example the set A can be written as

$$A = (A\cap B)\sqcup(A - B) = (A\cap B)\sqcup(A\cap B^c). \tag{1.9}$$

Definition 1.8. (Symmetric difference)
The **symmetric difference** between two sets A and B is written $A\Delta B$ and is the union of elements of A that do not belong to B and of elements of B that do not belong to A, i.e.

$$\begin{aligned} A\Delta B &= (A - B)\sqcup(B - A) = \\ &= (A\cap B^c)\sqcup(A^c\cap B) = \\ &= \{\omega : \omega \in A,\, \omega \notin B\}\sqcup\{\omega : \omega \in B,\, \omega \notin A\} \end{aligned}$$

Proposition 1.1. *Given two set A, B, the symmetric difference can be written as*

$$A\Delta B = (A\sqcup B)\cap(A^c\sqcup B^c).$$

Proof: Proposition 1.1

Proof. Let's denote with $C = A^c\cap B$, then apply the distributive law of the union twice

(Equation 1.5) and develop the computations, i.e.

$$\begin{aligned} A\Delta B &= (A\cap B^c)\sqcup(A^c\cap B) = \\ &= (A\cap B^c)\sqcup C = \\ &= (A\cup C)\cap(B^c\cup C) = \\ &= [A\cup(A^c\cap B)]\cap[B^c\cup(A^c\cap B)] = \\ &= [((A\sqcup A^c)\cap(A\cup B)]\cap(B^c\cup A^c)\cap(B^c\sqcup B) = \\ &= (A\cup B)\cap(A^c\cup B^c) \end{aligned}$$

□

1.2 Indicator function

Definition 1.9. (**Indicator function**)
An indicator function is a function that associate an $\omega \in A \subset \Omega$ to a real number, i.e. either 0 or 1. It is a tool that allows to transfer a computation from the set domain into the real numbers domain, i.e. $\{0,1\}$. Formally, $\mathbb{1}_A(\omega) : \Omega \to \{0,1\}$, i.e.

$$\mathbb{1}_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \in A^c \end{cases}$$

Proposition 1.2. *The containment between two sets can be equivalently written in terms of indicator functions:*

$$A \subset B \iff \mathbb{1}_A(\omega) \leq \mathbb{1}_B(\omega), \quad \forall \omega \in \Omega.$$

Proof: Proposition 1.2

Proof. In order to prove the results in Proposition 1.2, let's start by assuming $A \subset B$ and let's distinguish two main cases.

1. Assuming $\omega \in A$ implies that $\omega \in B$, and therefore one have an equality $1 = \mathbb{1}_A \leq \mathbb{1}_B = 1$.
2. Assuming $\omega \in A^c$ implies $[\omega \in B]\cup[\omega \in B^c]$. In this situation for both cases one will have that $\mathbb{1}_A \leq \mathbb{1}_B$, in fact:
   - Considering $\omega \in B$ implies that $0 = \mathbb{1}_A < \mathbb{1}_B = 1$.
   - Considering $\omega \in B^c$ implies that $0 = \mathbb{1}_A \leq \mathbb{1}_B = 0$.

Hence, assuming $A \subset B$ implies that $\mathbb{1}_A(\omega) \leq \mathbb{1}_B(\omega)$ for all $\omega \in \Omega$. Now let's assume the contrary: $\mathbb{1}_A \leq \mathbb{1}_B$ and let's again distinguish in two main cases:

1. Assuming $\omega \in A$, i.e. $\mathbb{1}_A = 1$, the inequality $\mathbb{1}_A \leq \mathbb{1}_B$ holds and since the indicator function is bounded by 1 by definition it is possible to write $1 = \mathbb{1}_A \leq \mathbb{1}_B \leq 1$. Therefore, one obtain $\mathbb{1}_B = 1$ and so $\omega \in B$.
2. Assuming $\omega \in A^c$, i.e. $\mathbb{1}_A = 0$, the inequality $\mathbb{1}_A \leq \mathbb{1}_B$ holds and it is possible to write $0 = \mathbb{1}_A \leq \mathbb{1}_B \leq 1$. Hence, when $\omega \in A^c$, there are two possible cases, i.e.
   - $\mathbb{1}_B = 1$, but this implies that $\omega \in B$.
   - $\mathbb{1}_B = 0$, but this implies that $\omega \in B^c$.

When an $\omega \in A$ implies that $\omega \in B$, but the contrary do not holds true. Hence, it is possible to conclude that $A \subset B$. □

1.3 Limits of sets

Let's define the infimum (inf) and the lim inf of a sequence of sets A_k as:

$$\inf_{k\geq n} = \bigcap_{k=n}^{\infty} A_k, \quad \lim_{n\to\infty} \inf A_n = \sup_{n\geq 1} \inf_{k\geq n} A_k = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k.$$

Informally, the infimum of a sequence of sets is the *smallest set in* $k = n, \dots, \infty$, it follows that the limit of the infimum (lim inf) is the *biggest (union) among all the smallest (intersection) sets*. Instead the supremum (sup) and the lim sup are defined as:

$$\sup_{k\geq n} = \bigcup_{k=n}^{\infty} A_k, \quad \lim_{n\to\infty} \sup A_n = \inf_{n\geq 1} \sup_{k\geq n} A_k = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k.$$

On the other hand, the supremum of a sequence of sets is the *biggest set in* $k = n, \dots, \infty$, it follows that the limit of the supremum (lim sup) is the *smallest (intersection) among all the biggest (union) sets*. Moreover, by De Morgan's laws:

$$(\lim_{n\to\infty} \sup A_n)^c = \left(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \right)^c = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c = \lim_{n\to\infty} \inf A_n^c,$$

and similarly

$$(\lim_{n\to\infty} \inf A_n)^c = \left(\bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k \right)^c = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k^c = \lim_{n\to\infty} \sup A_n^c.$$

1.4 Fields and σ-fields

Definition 1.10. (Field)
Let's consider the sample space Ω, then a field is a non-empty class of subsets of Ω closed under finite union, finite intersection and complementation. Formally, $\mathcal{A}$ is a field if and only if:

1. $\Omega \in \mathcal{A}$.
2. $A \in \mathcal{A} \implies A^c \in \mathcal{A}$.
3. $A, B \in \mathcal{A} \implies A \cup B \in \mathcal{A}$.

Definition 1.11. (σ-field)
Let's consider the sample space Ω, then a σ-field is a non-empty class of subsets of Ω closed under countable union, countable intersection and complementation. Formally, $\mathcal{B}$ is a σ-field if and only if:

1. $\Omega \in \mathcal{B}$.
2. $B \in \mathcal{B} \implies B^c \in \mathcal{B}$.
3. $B_i \in \mathcal{B}, i \geq 1 \implies \bigcup_{n \geq 1} B_n \in \mathcal{B}$.

⚠ Field vs σ-Field

The main difference between a field and a σ-field is in the third property of the definitions. A field is closed under **finite union**, namely the union of a finite sequence of events A_n indexed by $n \in \{0, 1, 2, \dots, n\}$ (property 3 of Definition 1.10). On the other hand, a σ-field is closed under **countable union**, namely the union of an infinite sequence of events A_n indexed by $n \in \{0, 1, 2, \dots, n, n+1, \dots\}$ (property 3. of the Definition 1.11).

2 Probability measure

A **probability space** is a triple $(\Omega, \mathcal{B}, \mathbb{P})$ where

1. Ω, the sample space.
2. $\mathcal{B}$, a σ-field of subsets of Ω where each element is called *event*.
3. $\mathbb{P}$ is a probability measure.

Definition 2.1. (**Probability measure**)
A probability measure $\mathbb{P}$ is any function $\mathbb{P} : \mathcal{B} \to [0, 1]$ such that

1. $\mathbb{P}(A) \geq 0$ for all sets $A \in \mathcal{B}$.
2. $\mathbb{P}(\Omega) = 1$.
3. $\mathbb{P}$ is σ-additive: if $\{A_n\}_{n\geq 1}$ are a sequence of disjoint events in $\mathcal{B}$, then:

$$\mathbb{P}\left(\bigsqcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n) \tag{2.1}$$

In general, a probability measure $\mathbb{P}$ is a function that always goes from a σ-field of subsets of Ω to $[0, 1]$.

2.1 Consequences of the axioms

Here we list some consequences of the axioms.

1. **Probability of the complement** of a set A:

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A). \tag{2.2}$$

Proof: probability of the complement

Proof. Since it is possible to write $\Omega = A \cup A^c$ as the union of disjoint set, we can apply

σ-additivity (Equation 2.1) to obtain:

$$\begin{aligned} \Omega = A \sqcup A^c \xrightarrow{\mathbb{P}} \mathbb{P}(\Omega) &= \mathbb{P}(A) + \mathbb{P}(A^c) \\ \implies 1 &= \mathbb{P}(A) + \mathbb{P}(A^c) \\ \implies \mathbb{P}(A^c) &= 1 - \mathbb{P}(A) \end{aligned}$$

□

2. Probability of the **empty set** $\emptyset$: $\mathbb{P}(\emptyset) = 0$.

Proof: probability of the empty set

Proof. Using the fact that $\mathbb{P}(\Omega) = 1$ by assumption and applying Equation 2.2:

$$\mathbb{P}(\emptyset) = 1 - \mathbb{P}(\emptyset^c) = 1 - \mathbb{P}(\Omega) = 0.$$

□

3. Probability of the **union** of two sets:

$$\mathbb{P}(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Proof: probability of the union of two sets

Proof. Let's write the sets A and B in terms of **union of disjoint events** (Equation 1.9) and apply $\mathbb{P}$ on both side and σ-additivity (Equation 2.1).

$$\begin{aligned} \mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) &\implies \mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A \cap B) \\ \mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(B \cap A^c) &\implies \mathbb{P}(B \cap A^c) = \mathbb{P}(B) - \mathbb{P}(A \cap B) \end{aligned} \tag{2.3}$$

Let's now decompose $A \cup B$ in the disjoint union of 3 events (Equation 1.8) and again, apply $\mathbb{P}$ on both side and σ-additivity:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) + \mathbb{P}(A^c \cap B).$$

Substituting $\mathbb{P}(A \cap B^c)$ and $\mathbb{P}(A \cap B^c)$ from Equation 2.3 gives the result:

$$\begin{aligned} \mathbb{P}(A \cup B) &= \mathbb{P}(A \cap B) + \mathbb{P}(B) - \mathbb{P}(A \cap B) + \mathbb{P}(A) - \mathbb{P}(A \cap B) = \\ &= \mathbb{P}(B) + \mathbb{P}(A) - \mathbb{P}(A \cap B) \end{aligned}$$

□

4. **Monotonicity property**: the measure $\mathbb{P}$ is *non-decreasing*. Given two events A and B,

$$A \subset B \implies \mathbb{P}(A) \leq \mathbb{P}(B).$$

Proof: monotonicity property

Proof. The proof of the statements follows once the set B is written as disjoint union of subsets of A and B (Equation 1.9). Then, applying the probability $\mathbb{P}$ and σ-additivity on both sides one obtain:

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B - A) \geq \mathbb{P}(A).$$

□

6. **Subadditivity**: the measure $\mathbb{P}$ is *σ-subadditive*. For a sequence of events $\{A_n\}_{n\geq 1}$ in $\mathcal{B}$ then:

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n). \tag{2.4}$$

7. **Continuity**: the measure $\mathbb{P}$ is **continuous** for a monotone sequence of sets $A_n \in \mathcal{B}$, i.e.

$$A_n \uparrow A \implies \mathbb{P}(A_n) \uparrow \mathbb{P}(A), \quad A_n \downarrow A \implies \mathbb{P}(A_n) \downarrow \mathbb{P}(A). \tag{2.5}$$

8. **Fatou's lemma**: consider a sequence of events $\{A_n\}_{n\geq 1}$ in $\mathcal{B}$, then we have the following result:

$$\mathbb{P}(\liminf_{n\to\infty} A_n) \leq \liminf_{n\to\infty} \mathbb{P}(A_n) \leq \limsup_{n\to\infty} \mathbb{P}(A_n) \leq \mathbb{P}(\limsup_{n\to\infty} A_n). \tag{2.6}$$

2.2 Maps and inverse maps

Let's be very general and consider a probability space $(\Omega, \mathcal{B}, \mathbb{P})$ and consider a map X that associate an $\omega \in \Omega$ to an outcome $\omega' \in \Omega'$, i.e.

$$X : (\Omega, \mathcal{B}) \to (\Omega', \mathcal{B}').$$

Then, X determine a function X^{-1} called **inverse map**, i.e.

$$X^{-1} : (\Omega', \mathcal{B}') \to (\Omega, \mathcal{B}).$$

In general, given a subset $A' \subset \mathcal{B}'$, its inverse map is defined as

$$X^{-1}(A') = \{\omega \in \Omega : X(\omega) \in A'\}.$$

Example: Map and inverse map

Example 2.1. Let's consider a deck of poker cards with 52 cards in total. We have 4 groups of 13 distinct cards, where the Jack (J) is 11, the Queen (Q) is 12, the King (K) is 13 and Ace (A) is 14. Then, let's consider a very general experiment setup in which we define a map

$$X(\omega) = \begin{cases} +1 & \text{if } \omega \in \{2,3,4,5,6\} \\ 0 & \text{if } \omega \in \{7,8,9\} \\ -1 & \text{if } \omega \in \{10,11,12,13,14\} \end{cases}$$

In this case the sample space will composed by 54 elements, i.e. all the cards, and $\Omega' = \{-1,0,1\}$. Let's say that we observe the value $X(\omega) = \{+1\} \subset \Omega'$. Then, the inverse map is the set $X^{-1}(\{+1\})$, i.e.

$$X^{-1}(\{+1\}) = \{\omega \in \Omega : X(\omega) \in \{+1\}\} = \{2,3,4,5,6\}.$$

In practice, we have to search those ω's such that $X(\omega) = \{1\}$.

Here we list some properties of inverse maps.

1. $X^{-1}(\Omega') = \Omega$.
2. $X^{-1}(\emptyset) = \emptyset$.
3. $X^{-1}(A'^c) = (X^{-1}(A'))^c$.
4. $X^{-1}(\Omega' \cap A') = \Omega \cap X^{-1}(A'^c)$.
5. $X^{-1}(\bigcup_n A'_n) = \bigcup_n X^{-1}(A'_n)$ for all $A'_n \in \mathcal{B}'$.

Properties of inverse maps

Let's consider two sets A' and B' both in Ω'. Then, by definition:

$$\begin{aligned} X^{-1}(A' \cup B') &= \{\omega \in \Omega : X(\omega) \in A' \cup B'\} = \\ &= \{\omega \in \Omega : X(\omega) \in A' \text{ OR } X(\omega) \in B'\} = \\ &= \{\omega \in \Omega : X(\omega) \in A'\} \cup \{\omega \in \Omega : X(\omega) \in B'\} = \\ &= X^{-1}(A') \cup X^{-1}(B') \end{aligned}$$

Similarly for the intersection, i.e.

$$\begin{aligned} X^{-1}(A' \cap B') &= \{\omega \in \Omega : X(\omega) \in A' \cap B'\} = \\ &= \{\omega \in \Omega : X(\omega) \in A' \text{ AND } X(\omega) \in B'\} = \\ &= \{\omega \in \Omega : X(\omega) \in A'\} \cap \{\omega \in \Omega : X(\omega) \in B'\} = \\ &= X^{-1}(A') \cap X^{-1}(B') \end{aligned}$$

Proposition 2.1. *If $\mathcal{B}'$ is a σ-field of subsets of Ω', then $X^{-1}(\mathcal{B}')$ is a σ-field of subsets of Ω. Moreover, if $\mathcal{C}'$ is a class of subsets of Ω', then*

$$X^{-1}(\sigma(C')) = \sigma(X^{-1}(C')),$$

that is. the inverse image of the σ-field generated by the class $\mathcal{C}' \in \Omega'$ is the same as the σ-field generated in Ω by the inverse image X^{-1}. In practice, the counter image and the generators commute. Usually can be difficult to know all about the σ-field $\mathcal{B}'$, however if we know a class of subset that generate it, namely $\mathcal{C}' \in \Omega'$, we are able to recreate the σ-field.

2.2.1 Measurable maps

A measurable space is composed by a sample space Ω and a σ-field of subsets of Ω, namely $\mathcal{B}$.

Definition 2.2. (Measurable map)
Let's consider the function $X : (\Omega, \mathcal{B}) \to (\Omega', \mathcal{B}')$, then X is $\mathcal{B}$-**measurable**, namely $X \in \mathcal{B}/\mathcal{B}'$, iff:

$$X \in \mathcal{B}/\mathcal{B}' \iff X^{-1}(\mathcal{B}') \in \mathcal{B}.$$

Note that the measurability concept is very important since only if X is measurable it is possible to make probability statements about X. Since, $X^{-1}(B') \in \mathcal{B}$ for all $B' \in \mathcal{B}'$ it is possible to assign probabilities to the events that are in $\mathcal{B}$.

Definition 2.3. (Test for measurability)
Consider a map $X : (\Omega, \mathcal{B}) \to (\Omega', \mathcal{B}')$ and the class $\mathcal{C}'$ that generates the σ-field $\mathcal{B}'$, i.e. $\mathcal{B}' = \sigma(\mathcal{C}')$. Then X is $\mathcal{B}$-measurable iff:

$$X^{-1}(\mathcal{C}') \subset \mathcal{B}.$$

3 Random variables

Definition 3.1. (Random variable)
Let's consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then a random variable is a map where $(\Omega', \mathcal{B}') = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and therefore the map takes values on the real line, i.e.

$$X : (\Omega, \mathcal{B}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$$

and such that:

$$\forall B \in \mathcal{B}(\mathbb{R}) \quad X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \subset \mathcal{B}$$

Note that when X is a random variable the test for measurability (Definition 2.3), became:

$$X^{-1}((-\infty, y]) = [X(\omega) \leq y] \subset \mathcal{B} \quad \forall y \in \mathbb{R}$$

! Discrete vs continuous random variables

Definition 3.2. Let X be a random variable with set of possible outcomes Ω'. Then, X is called **discrete** random variable if Ω' is either a finite set or a countably infinite set. X is called **continuous** random variable if Ω' is either a an uncountable infinite set.

3.1 Induced distribution function

Consider a probability space $(\Omega, \mathcal{B}, \mathbb{P})$ and a measurable map $X : (\Omega, \mathcal{B}) \to (\Omega', \mathcal{B}')$, then the composition $\mathbb{P} \circ X^{-1}$ is again a map. In this way at each element $\omega \in \Omega$ is attached a probability measure. In fact, the composition is a map such that

$$\mathbb{P} \circ X^{-1} : (\Omega', \mathcal{B}') \to [0, 1] \iff (\Omega', \mathcal{B}') \xrightarrow{X^{-1}} (\Omega, \mathcal{B}) \xrightarrow{\mathbb{P}} [0, 1]$$

In general, the probability of a subset $A' \in \mathcal{B}'$ is denoted equivalently as:

$$\mathbb{P} \circ X^{-1}(A') = \mathbb{P}(X^{-1}(A')) = \mathbb{P}(X(\omega) \in A')$$

Example: Map and inverse map (continued)

Example 3.1. Let's continue from the Example 2.1 and compute the probability of $\mathbb{P} \circ X^{-1}(\{+1\})$. Let's consider one random extraction from the 52 cards, then for each distinct number we have 4 copies. Therefore the probability is computed as:

$$\begin{aligned}\mathbb{P}(X^{-1}(\{+1\})) &= \mathbb{P}(\{\omega \in \Omega : X(\omega) \in \{-1\}\}) = \\ &= \mathbb{P}(\{2,3,4,5,6\}) = \\ &= \frac{5 \cdot 4}{52} = \frac{5}{13} \approx 38.46\%\end{aligned}$$

Let's now consider the probability of observing either $\{+1\}$ or $\{-1\}$, then

$$\begin{aligned}\mathbb{P}(X^{-1}(\{-1,+1\})) &= \mathbb{P}(\{\omega \in \Omega : X(\omega) \in \{-1,+1\}\}) = \\ &= \mathbb{P}(\{2,3,4,5,6,10,11,12,13,14\}) = \\ &= \frac{10 \cdot 4}{52} = \frac{10}{13} \approx 76.92\%\end{aligned}$$

Finally, by property of the probability measure $\mathbb{P}(X(\omega) \in \{0\}) = 1 - \mathbb{P}(X(\omega) \in \{-1,+1\}) \approx 23.08\%$.

3.1.1 Distribution function on $\mathbb{R}$

When X is a random variable the composition $\mathbb{P}(X^{-1}(A'))$ is a probability measure induced on $\mathbb{R}$ by the distribution:

$$\mathbb{P}(X^{-1}((-\infty, y]) = \mathbb{P}(X \leq y) \quad \forall y \in \mathbb{R}$$

Hence, it is possible to attach to a random variable a distribution function of X that is a measure induced on the real line $\mathbb{R}$ and defined as:

$$F_X(y) = \mathbb{P}(X(\omega) \in [-\infty, y)) = \mathbb{P}(X \leq y)$$

The distribution of X is a function that goes from $F_X : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to [0,1]$. If a random variable has a continuous and it's distribution function is differentiable, then it is possible to define the density as:

$$f_X(y) = \frac{dF_X(y)}{dy} \iff dF_X(y) = f_X(y)dy \tag{3.1}$$

4 Independence

Definition 4.1. (**Independent events**)
Given a probability space $(\Omega, \mathcal{B}, \mathbb{P})$, two events $A, B \in \mathcal{B}$ are said to be independent if:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$$

A finite sequence of events, namely $A_1, A_2, \ldots, A_n \in \mathcal{B}$, is said to be independent, if for all $2 \leq j \leq n$ and $1 \leq k_1 \leq k_2 \leq \ldots \leq k_j \leq n$ we have:

$$\mathbb{P}(A_{k_1} \cap A_{k_2} \cap \ldots \cap A_{k_n}) = \prod_{j=1}^{n} \mathbb{P}(A_{k_j})$$

Events not pairwise independent.

Consider the probability space $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$, where $\Omega = \{1, 2, 3, 4, 5, 6\}$ and for every $\omega_i \in \Omega$ we have a constant probability $\mathbb{P}(\omega_i) = \frac{1}{6} \ \ \forall i$. Consider the events $A_1 = \{1, 2, 3, 4\}$ and $A_2 = A_3 = \{4, 5, 6\}$, are these events independent? Note that the events have probabilities $\mathbb{P}(A_1) = \frac{2}{3}$, $\mathbb{P}(A_2) = \mathbb{P}(A_3) = \frac{1}{2}$. Consider all the events A_1, A_2, A_3, then the intersection of those sets gives $[A_1 \cap A_2 \cap A_3] = \{4\}$ that has probability $\mathbb{P}(\{4\}) = \frac{1}{6}$. Then we can compute the product of the probabilities of the single events:

$$\frac{1}{2} = \mathbb{P}([A_1 \cap A_2 \cap A_3]) = \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3) = \frac{2}{3}\frac{1}{2}\frac{1}{2} = \frac{1}{2}$$

Hence the events A_1, A_2, A_3 are **pairwise independent**. Consider now only the events A_2, A_3, the probability of the joint set, namely $[A_2 \cap A_3] = \{4, 5, 6\}$, is $\mathbb{P}(\{4, 5, 6\}) = \frac{1}{2}$. However the product of the probabilities of the single events gives a different result:

$$\frac{1}{2} = \mathbb{P}([A_2 \cap A_3]) \neq \mathbb{P}(A_2)\mathbb{P}(A_3) = \frac{1}{2}\frac{1}{2} = \frac{1}{4}$$

Hence the events A_2, A_3 are **NOT pairwise independent**.

Proposition 4.1. *(**Independence and complementation**)*
If two events A and B are independent, then also are A and B^c, B and A^c, A^c and B^c are independent.

Independence and complementation

Proof. If two events A and B are independent, then also are A and B^c, B and A^c, A^c and B^c. In order to prove that A and B^c are independent let's write the event A as union of disjoint events (Equation 1.9). Then since A and B are assumed to be independent:

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) = \\ &= \mathbb{P}(A)\mathbb{P}(B) + \mathbb{P}(A \cap B^c)\end{aligned}$$

Recovering $\mathbb{P}(A \cap B^c)$ one obtain:

$$\begin{aligned}\mathbb{P}(A \cap B^c) &= \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B) = \\ &= \mathbb{P}(A)(1 - \mathbb{P}(B)) = \\ &= \mathbb{P}(A)\mathbb{P}(B^c)\end{aligned}$$

The same follows for A^c and B. Now let's consider the case of A^c and B^c. Using the same trick done previously $A^c = [A^c \cap B] \sqcup [A^c \cap B^c]$. Since we have already proven that B^c and A are independent, we can write:

$$\begin{aligned}\mathbb{P}(A^c) &= \mathbb{P}([A^c \cap B] \sqcup [A^c \cap B^c]) = \\ &= \mathbb{P}(A^c \cap B) + \mathbb{P}(A^c \cap B^c) = \\ &= \mathbb{P}(A^c)\mathbb{P}(B) + \mathbb{P}(A^c \cap B^c)\end{aligned}$$

Recovering $\mathbb{P}(A^c \cap B^c)$ one obtain:

$$\begin{aligned}\mathbb{P}(A^c \cap B^c) &= \mathbb{P}(A^c) - \mathbb{P}(A^c)\mathbb{P}(B) = \\ &= \mathbb{P}(A^c)(1 - \mathbb{P}(B)) = \\ &= \mathbb{P}(A^c)\mathbb{P}(B^c)\end{aligned}$$

□

5 Expectation

The **expectation** represents a central value of a random variable and has a measure theory counterpart as a Lebesgue-Stieltjes integral of X with respect to a (probability) measure $\mathbb{P}$. This kind of integration is defined in steps. First it is shown the integration of **simple functions** and then extended to more general random variables. In general, let's consider a probability space $(\Omega, \mathcal{B}, \mathbb{P})$ and a random variable X such that

$$X : (\Omega, \mathcal{B}) \longrightarrow (\bar{\mathbb{R}}, \mathcal{B}(\bar{\mathbb{R}}))$$

where $\bar{\mathbb{R}} = [-\infty, \infty]$. Then, the **expectation** of X is denoted as:

$$\mathbb{E}\{X\} = \int_{\Omega} X d\mathbb{P} = \int_{\Omega} X(\omega)\mathbb{P}(d\omega)$$

as the Lebesgue-Stieltjes integral of X with respect to the (probability) measure $\mathbb{P}$.

5.1 Simple functions

In general a random variable $X(\omega)$ is *simple* if it has a *finite range*. Let's consider a probability space $(\Omega, \mathcal{B}, \mathbb{P})$ and consider a $\mathcal{B}/\mathcal{B}(\mathbb{R})$-measurable **simple function** $X : \Omega \to \mathbb{R}$, i.e.

$$X(\omega) = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}(\omega), \tag{5.1}$$

where $a_i \in \mathbb{R}$ and $A_i \in \mathcal{B}$ are a disjoint partition of the sample space, i.e. $\bigsqcup_{i=1}^{n} A_i = \Omega$. Let's denote the set of all simple functions on Ω as $\mathcal{E}$. In this settings, $\mathcal{E}$ is a **vector space**. This implies that he following two properties holds.

1. **Constant**: given a simple function $X \in \mathcal{E}$, then $\alpha X \in \mathcal{E}$. In fact:

$$\alpha X = \sum_{i=1}^{n} \alpha a_i \mathbb{1}_{A_i} = \sum_{i=1}^{n} a_i^* \mathbb{1}_{A_i} \in \mathcal{E} \tag{5.2}$$

   where $a_i^* = \alpha a_i$.

2. **Linearity**: given two simple function $X, Y \in \mathcal{E}$, then $X + Y \in \mathcal{E}$. In fact:

$$\begin{aligned} X + Y &= \sum_{i=1}^{n} a_i \mathbb{1}_{A_i} + \sum_{j=1}^{m} b_j \mathbb{1}_{B_j} = \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} (a_i + b_j) \mathbb{1}_{A_i} \mathbb{1}_{B_j} = \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} (a_i + b_j) \mathbb{1}_{A_i \cap B_j} \end{aligned} \tag{5.3}$$

   where the sequence of sets $\{A_i B_j \; 1 \leq i \leq n \text{ and } 1 \leq j \leq m\}$ form a disjoint partition of Ω.

3. **Product**: given two simple function $X, Y \in \mathcal{E}$, then $XY \in \mathcal{E}$. In fact:

$$\begin{aligned} XY &= \sum_{i=1}^{n} a_i \mathbb{1}_{A_i} \sum_{j=1}^{m} b_j \mathbb{1}_{B_j} = \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} (a_i b_j) \mathbb{1}_{A_i} \mathbb{1}_{B_j} = \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} (a_i b_j) \mathbb{1}_{A_i \cap B_j} \end{aligned} \tag{5.4}$$

5.1.1 Measurability

Simple functions are the building blocks in the definition of the expectation in terms of Lebesgue-Stieltjes integral. In fact a known theorem called **Measurability theorem** shows that any measurable function can be approximated by a sequence of simple functions.

Theorem 5.1. *Suppose that $X(\omega) \geq 0$ for all $\omega \in \Omega$. Then, X is $\mathcal{B}/\mathcal{B}(\mathbb{R})$ measurable if and only if there exists simple functions $X_n \in \mathcal{E}$ and*

$$0 \leq X_n \uparrow X \iff X = \lim_{n \to \infty} \uparrow X_n$$

5.2 Expectation of Simple Functions

The expectation of a simple function X is defined as:

$$\mathbb{E}\{X\} = \sum_{i=1}^{n} a_i \mathbb{P}(A_i)$$

where $|a_i| < \infty$.

5.2.1 Properties

1. **Non-negativity**: If $X \geq 0$ and $X \in \mathcal{E}$ then $\mathbb{E}\{X\} \geq 0$

> **i** Expectation of a simple function is non-negative
>
> *Proof.* By definition of simple functions □

2. **Linearity**: the expectation of simple function is linear, i.e.

$$\mathbb{E}\{\alpha X + \beta Y\} = \alpha \mathbb{E}\{X\} + \beta \mathbb{E}\{Y\}$$

> **i** Expectation of a simple function is linear
>
> *Proof.* Let's consider two simple functions, i.e.
>
> $$X(\omega) = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}(\omega) \quad \text{and} \quad Y(\omega) = \sum_{j=1}^{m} b_j \mathbb{1}_{B_j}(\omega),$$
>
> and let's fix $\alpha, \beta \in \mathbb{R}$. Then, by the second property of the vector space $\mathcal{E}$ (Equation 5.3) it is possible to write:
>
> $$\alpha X + \beta Y = \sum_{i=1}^{n} \sum_{j=1}^{m} (\alpha a_i + \beta b_j) \mathbb{1}_{A_i \cap B_j}$$
>
> Then, taking the expectation on both sides:
>
> $$\begin{aligned} \mathbb{E}\{\alpha X + \beta Y\} &= \sum_{i=1}^{n} \sum_{j=1}^{m} (\alpha a_i + \beta b_j) \mathbb{P}(A_i \cap B_j) = \\ &= \sum_{i=1}^{n} \alpha a_i \sum_{j=1}^{m} \mathbb{P}(A_i \cap B_j) + \sum_{j=1}^{m} \beta b_j \sum_{i=1}^{n} \mathbb{P}(A_i \cap B_j) \end{aligned}$$
>
> Fixing i, the sequence $A_i \cap B_j$ for $j = 1, \dots, n$ is composed by disjoint events since by definition B_j are disjoint. Hence, applying σ-additivity it is possible to write:
>
> $$\begin{aligned} \sum_{j=1}^{m} \mathbb{P}(A_i \cap B_j) &= \mathbb{P}\left(\bigsqcup_{j=1}^{m} A_i \cap B_j \right) = \\ &= \mathbb{P}\left(A_i \cap \left(\bigsqcup_{j=1}^{m} B_j \right) \right) = \\ &= \mathbb{P}\left(A_i \cap \Omega \right) = \mathbb{P}\left(A_i \right) \end{aligned}$$

Therefore, the expectation simplifies in:

$$\mathbb{E}\{\alpha X + \beta Y\} = \sum_{i=1}^{n} \alpha a_i \mathbb{P}(A_i) + \sum_{j=1}^{m} \beta b_j \mathbb{P}(B_j) =$$
$$= \alpha \mathbb{E}\{X\} + \beta \mathbb{E}\{Y\}$$

□

5.3 Review of inequalities

5.3.1 Modulus inequality

Definition 5.1. (Modulus Inequality)
Let's consider a random variable $X \in \mathcal{L}_1$, where $\mathcal{L}_1$ stands for the set of integrable random variables, i.e.

$$\mathcal{L}_1 = \{X : \Omega \to \mathbb{R} : X \text{ is a r.v. }, \mathbb{E}\{|X|\} < \infty\}$$

Then, the modulus inequality states that:

$$|\mathbb{E}\{X\}| \leq \mathbb{E}\{|X|\}$$

5.3.2 Markov inequality

Definition 5.2. (Markov Inequality)
Let's consider a random variable $X \in \mathcal{L}_1$ and fix a $\lambda > 0$, then by the Markov inequality:

$$\mathbb{P}(|X| \geq \lambda) \leq \frac{1}{\lambda} \mathbb{E}\{|X|\}$$

5.3.3 Chebychev inequality

Definition 5.3. (Chebychev Inequality)
Consider a random variable X with first and second moment finite, i.e.

$$\mathbb{E}\{|X|\} < \infty, \quad \mathbb{V}\{|X|\} < \infty$$

then by the Chebychev inequality:

$$\mathbb{P}(X \geq \lambda) \leq \frac{1}{\lambda^2} \mathbb{E}\{|X|^2\} \tag{5.5}$$

5.3.4 Holder inequality

Definition 5.4. (Holder Inequality)
Let's consider two numbers p and q such that

$$p > 1, q > 1, \ \frac{1}{p} + \frac{1}{q} = 1$$

and let's consider two random variables X and Y such that:

$$\mathbb{E}\{|X|^p\} < \infty, \quad \mathbb{E}\{|Y|^q\} < \infty$$

Then,

$$|\mathbb{E}\{XY\}| \leq \mathbb{E}\{|XY|\} \leq (\mathbb{E}\{|X|^p\})^{\frac{1}{p}} (\mathbb{E}\{|Y|^q\})^{\frac{1}{q}} \tag{5.6}$$

In terms of norms:

$$||XY||_1 \leq ||X||_p ||Y||_q$$

5.3.5 Schwartz inequality

Definition 5.5. (Schwartz Inequality)
Consider two random variables $X, Y \in \mathcal{L}_2$, i.e. with first and second moment finite, i.e.

$$\mathbb{E}\{|X|\} < \infty, \quad \mathbb{E}\{X^2\} < \infty$$

Then

$$|\mathbb{E}\{XY\}| \leq \mathbb{E}\{|XY|\} \leq \sqrt{\mathbb{E}\{X^2\}\mathbb{E}\{Y^2\}} \tag{5.7}$$

In terms of norms:

$$||XY||_1 \leq ||X||_2 ||Y||_2$$

Note that this is a special case of Holder inequality (Equation 5.6) with $p = q = 2$.

5.3.6 Minkowski inequality

Definition 5.6. (Minkowski Inequality)
For $1 \leq p < \infty$ let's consider two random variables $X, Y \in \mathcal{L}_p$, then $X + Y \in \mathcal{L}_p$ and

$$||X + Y||_p \leq ||X||_p + ||Y||_p \tag{5.8}$$

Note that the triangular inequality is a special case of Minkowski inequality with $p = 1$, i.e.

$$|X + Y| \leq |X| + |Y| \tag{5.9}$$

5.3.7 Jensen inequality

Definition 5.7. (Jensen Inequality)
Let's consider a **convex** function $u : \mathbb{R} \to \mathbb{R}$. Suppose that $\mathbb{E}\{X\} < \infty$ and $\mathbb{E}\{|u(X)|\} < \infty$, then

$$\mathbb{E}\{u(X)\} \geq u(\mathbb{E}\{X\}) \tag{5.10}$$

if u is **concave** the results revert, i.e.

$$\mathbb{E}\{u(X)\} \leq u(\mathbb{E}\{X\}) \tag{5.11}$$

6 Conditional expectation

Theorem 6.1. *(Radom Nikodym)*
Consider a measure space $(\Omega, \mathcal{B})$ and two measures μ, ν such that μ is σ-finite (Definition 31.4) and $\mu << \nu$ (Definition 31.1). Then there exists a measurable function $X : \Omega \to \mathbb{R}$ such that:

$$\mu(B) = \int_A X d\nu \quad \forall B \in \mathcal{B}$$

Definition 6.1. (Conditional expectation)
Given a probability space $(\Omega, \mathcal{B}, \mathbb{P})$, consider $\mathcal{G}$ as a sub σ-field of $\mathcal{B}$, i.e. $\mathcal{G} \subset \mathcal{B}$. Let's consider a random variable $X : \Omega \to \mathbb{R}$ with finite expectation $\mathbb{E}\{|X|\} < +\infty$. We define a conditional expectation for X given $\mathcal{G}$, any random variable $Y = \mathbb{E}\{X|\mathcal{G}\}$ such that:

1. Y has finite expectation, i.e. $\mathbb{E}\{|Y|\} < +\infty$.
2. Y is $\mathcal{G}$-measurable.
3. $\mathbb{E}\{\mathbb{1}_A Y\} = \mathbb{E}\{\mathbb{1}_A X\}$, $\forall A \in \mathcal{G}$, namely if X and Y are restricted to $A \in \mathcal{G}$, then their expectation coincides.

A σ-field can be used to describe our state of information. It means that, $\forall A \in \mathcal{G}$ we already know if the event A has occurred or not. Therefore, when we insert in $\mathcal{G}$ the events that we know were already occurred, we are saying that the random variable Y is $\mathcal{G}$-measurable, i.e. the value of Y is not stochastic once we know the information contained in $\mathcal{G}$. Moreover, the random variable $Y = \mathbb{E}\{X|\mathcal{G}\}$ represent a **prediction** of the random variable X, given the information contained in the sub σ-field $\mathcal{G}$.

Definition 6.2. (Predictor)
Consider Z any $\mathcal{G}$-measurable random variable. Then Z can be interpreted as a predictor of another random variable X under the information contained in the σ-field $\mathcal{G}$. However, when we substitute X with its prediction, namely Z, we make an error given by the difference $X - Z$. In the special case in which $\mathbb{E}\{|Z|^2\} < \infty$, we can take as error function the mean squared error, i.e.

$$\mathbb{E}\{\text{error}^2\} = \mathbb{E}\{(X - Z)^2\}$$

We say that the conditional expectation $\mathbb{E}\{X|\mathcal{G}\}$ is the **best predictor** in the sense that:

$$\mathbb{E}\{(X - \mathbb{E}\{X|\mathcal{G}\})^2\} = \min_{Z \in \mathcal{Z}} \mathbb{E}\{(X - Z)^2\}$$

Hence, $\mathbb{E}\{X|\mathcal{G}\}$ is the best predictor that minimize the mean squared error over the class $\mathcal{Z}$ composed by $\mathcal{G}$-measurable functions with finite second moment, formally

$$\mathcal{Z} = \{Z \;\; \mathcal{G}\text{-measurable and } \mathbb{E}\{|Z|^2\} < \infty\}$$

6.1 Properties of conditional expectation

Here we state some useful properties of conditional expectation:

1. **Linearity**: $\mathbb{E}\{aX + bY|\mathcal{G}\} = a\mathbb{E}\{X|\mathcal{G}\} + b\mathbb{E}\{Y|\mathcal{G}\}$, for all constants $a, b \in \mathbb{R}$.
2. **Positive**: $X \geq 0 \implies \mathbb{E}\{X|\mathcal{G}\} \geq 0$.
3. **Measurability**: If Y is $\mathcal{G}$-measurable, then $\mathbb{E}\{XY|\mathcal{G}\} = Y\mathbb{E}\{X|\mathcal{G}\}$.
4. **Constant**: $\mathbb{E}\{a|\mathcal{G}\} = a \;\; \forall a \in \mathbb{R}$. In general, if X is $\mathcal{G}$-measurable then $\mathbb{E}\{X|\mathcal{G}\} = X$, i.e. is not stochastic.
5. **Independence**: If X is independent from the σ-field $\mathcal{G}$, then $\mathbb{E}\{X|\mathcal{G}\} = \mathbb{E}\{X\}$.
6. **Chain rule**: consider two two sub σ-fields of $\mathcal{B}$ such that $\mathcal{G}_1 \subset \mathcal{G}_2$, then we can write:

   $$\mathbb{E}\{X|\mathcal{G}_1\} = \mathbb{E}\{\mathbb{E}\{X|\mathcal{G}_2\}|\mathcal{G}_1\}$$

   Remember that, when using this property it is mandatory to take the conditional expectation before with respect to the greatest σ-field, i.e. the one that contains more information (in this case $\mathcal{G}_2$), and then with respect to the smallest one (in this case $\mathcal{G}_1$).

6.2 Conditional probability

Definition 6.3. (Conditional probability)
Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, consider $\mathcal{G}$ as a sub σ-field of $\mathcal{F}$, i.e. $\mathcal{G} \subset \mathcal{F}$. Then the **general definition** of the conditional probability of an event A given $\mathcal{G}$ is:

$$\mathbb{P}(A|\mathcal{G}) = \mathbb{E}(\mathbb{1}_A|\mathcal{G}) \tag{6.1}$$

Instead, the **elementary definition** do not consider the conditioning with respect to a σ-field, but instead with respect to a single event B. In practice, take an event $B \in \mathcal{F}$ such that $0 < \mathbb{P}(B) < 1$, then $\forall A \in \mathcal{F}$ the conditional probability of A given B is defined as:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad \mathbb{P}(A|B^c) = \frac{\mathbb{P}(A \cap B^c)}{\mathbb{P}(B^c)} \tag{6.2}$$

Elementary and the general definition are equivalent

The elementary (Equation 6.2) and the general (Equation 6.1) definitions are equivalent, in fact consider a sub σ-field $\mathcal{G}$ which provides only the information concerning whenever ω is in B or not. A σ-field of this kind will have the form $\mathcal{G}_B = \{\Omega, \emptyset, B, B^c\}$. Then, consider a $\mathcal{G}_B$-measurable function, $f : \Omega \to \mathbb{R}$, such that:

$$f(\omega) = \begin{cases} \alpha & \omega \in B \\ \beta & \omega \in B^c \end{cases}$$

It remains to find α and β in the following expression:

$$\mathbb{P}(A|\mathcal{G}_B) = \mathbb{E}\{\mathbb{1}_A|\mathcal{G}_B\} = \alpha\mathbb{1}_B + \beta\mathbb{1}_{B^c}$$

Note that, the joint probability of A and B can be obtained as:

$$\begin{aligned} \mathbb{P}(A \cap B) &= \mathbb{E}\{\mathbb{1}_A\mathbb{1}_B\} = \mathbb{E}\{\mathbb{E}\{\mathbb{1}_A\mathbb{1}_B|\mathcal{G}_B\}\} = \mathbb{E}\{\mathbb{E}\{\mathbb{1}_A|\mathcal{G}_B\}\mathbb{1}_B\} = \\ &= \mathbb{E}\{\mathbb{P}(A|\mathcal{G}_B)\mathbb{1}_B\} = \\ &= \mathbb{E}\{(\alpha\mathbb{1}_B + \beta\mathbb{1}_{B^c})\mathbb{1}_B\} = \\ &= \alpha\mathbb{E}\{\mathbb{1}_B\} + \beta\mathbb{E}\{\mathbb{1}_{B^c}\mathbb{1}_B\} = \\ &= \alpha\mathbb{P}(B) \end{aligned}$$

Hence, we obtain:

$$\mathbb{P}(A \cap B) = \alpha\, \mathbb{P}(B) \implies \alpha = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Equivalently for $\mathbb{P}(A \cap B^c)$ it is possible to prove that:

$$\mathbb{P}(A \cap B^c) = \beta\, \mathbb{P}(B^c) \implies \beta = \frac{\mathbb{P}(A \cap B^c)}{\mathbb{P}(B^c)}$$

Finally it is possible to write the conditional probability in the general definition as a linear combination of conditional probabilities defined accordingly to the elementary one, i.e.

$$\mathbb{P}(A|\mathcal{G}_B) = \mathbb{P}(A|B)\mathbb{1}_B + \mathbb{P}(A|B^c)\mathbb{1}_{B^c}$$

Conditional probability

Example 6.1. Let's continue from the example Example 2.1, let's say that we observe $X(\omega) = \{+1\}$, then we ask ourselves, **what is the probability that in the next extraction** $X(\omega) = \{0\}$? The chances that with 52 cards we obtain $X(\omega) = \{0\}$ is approximately $\frac{3}{13} \approx 23.08\%$ (see Example 3.1). Then, given the fact that the extracted

card originates $X(\omega) = \{+1\}$ we have that the probability, conditional to the fact that in the first extraction we had a card $\{+1\}$, that in the next extraction we have $\{0\}$ is $\frac{12}{51} = 23.52\%$. Let's now investigate the chances that in the next extraction $X(\omega) = \{+1\}$ given that in the previous was $\{+1\}$. The unconditional probability is $\frac{20}{52} \approx 38.46\%$, the conditional probability will be $\frac{19}{51} \approx 37.25\%$.

Conditional probability: numerical example

Example 6.2. Let's consider two random variables $X(\omega)$ and $Y(\omega)$ taking values in $\{0, 1\}$. The marginal probabilities $\mathbb{P}(X = 0) = 0.6$ and $\mathbb{P}(Y = 0) = 0.29$. Let's consider the matrix of joint events and probabilities, i.e.

$$\begin{pmatrix} [X = 0] \cap [Y = 0] & [X = 0] \cap [Y = 1] \\ [X = 1] \cap [Y = 0] & [X = 1] \cap [Y = 1] \end{pmatrix} \xrightarrow{\mathbb{P}} \begin{pmatrix} 0.17 & 0.43 \\ 0.12 & 0.28 \end{pmatrix}$$

Then, by definition the conditional probabilities are defined as:

$$\mathbb{P}(X = 0|Y = 0) = \frac{\mathbb{P}(X = 0 \cap Y = 0)}{\mathbb{P}(Y = 0)} = \frac{0.17}{0.29} \approx 58.63\%$$

and

$$\mathbb{P}(X = 0|Y = 1) = \frac{\mathbb{P}(X = 0 \cap Y = 1)}{\mathbb{P}(Y = 1)} = \frac{0.43}{1 - 0.29} \approx 60.56\%$$

Considering Y instead:

$$\mathbb{P}(Y = 0|X = 0) = \frac{\mathbb{P}(Y = 0 \cap X = 0)}{\mathbb{P}(X = 0)} = \frac{0.17}{0.6} \approx 28.33\%$$

and

$$\mathbb{P}(Y = 0|X = 1) = \frac{\mathbb{P}(Y = 0 \cap X = 1)}{\mathbb{P}(X = 1)} = \frac{0.12}{1 - 0.6} \approx 30\%$$

Then, it is possible to express the marginal probability of X as:

$$\begin{aligned} \mathbb{P}(X = 0) &= \mathbb{E}\{\mathbb{P}(X = 0|Y)\} = \\ &= \mathbb{P}(X = 0|Y = 0)\mathbb{P}(Y = 0) + \mathbb{P}(X = 0|Y = 1)\mathbb{P}(Y = 1) = \\ &= 0.5863 \cdot 0.29 + 0.6056 \cdot (1 - 0.29) \approx 60\% \end{aligned}$$

And similarly for Y

$$\begin{aligned} \mathbb{P}(Y = 0) &= \mathbb{E}\{\mathbb{P}(Y = 0|X)\} = \\ &= \mathbb{P}(Y = 0|X = 0)\mathbb{P}(X = 0) + \mathbb{P}(Y = 0|X = 1)\mathbb{P}(X = 1) = \\ &= 0.2833 \cdot 0.6 + 0.30 \cdot (1 - 0.6) \approx 29\% \end{aligned}$$

7 Characteristic functions

Definition 7.1. (**Characteristic function**)
Consider an n-dimensional vector $\mathbf{X}$, then the characteristic function is defined $\forall t \in \mathbb{R}^n$ as:

$$\Phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}\{e^{i\mathbf{t}^T\mathbf{X}}\} \quad \text{where} \quad \mathbf{t}^T\mathbf{X} = \begin{pmatrix} t_1, t_2, \dots, t_n \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

The characteristic function always exists when treated as a function of a real-valued argument, unlike the moment-generating function. The characteristic function uniquely determines the probability distribution of the correspondent random vector $\mathbf{X}$. More precisely, saying that two random variables has the same distribution is equivalent to say that their characteristic functions are equal. It follows that we can always work under characteristic functions to prove that two distribution of some random vectors are equal or that a distribution converges to another distribution. Formally $\mathbf{X}$ and $\mathbf{Y}$ ave same distribution, i.e.

$$\mathbf{X} \sim \mathbf{Y} \iff \Phi_{\mathbf{X}}(\mathbf{t}) = \Phi_{\mathbf{Y}}(\mathbf{t}) \quad \forall \mathbf{t} \in \mathbb{R}^n$$

Here, we list some properties considering the random variable case, i.e. $n = 1$, with $t \in \mathbb{R}$.

1. **Independence**: X and Y are independent iff:

$$X \perp Y \iff \Phi_{X+Y}(t) = \Phi_X(t)\Phi_Y(t) \quad \forall t \in \mathbb{R}$$

2. **Existence of the j-th moment**: If the j-th moment of the random variable is finite then the characteristic function is j-times differentiable and continuous in 0, i.e. $\Phi_X \in \mathbb{C}^{(j)}$. Formally,

$$\mathbb{E}\{|X|^r\} < \infty \implies \Phi_X \in \mathbb{C}^{(r)} \quad \text{and} \quad \Phi_X^{(r)}(t) = \mathbb{E}\{(i\mathbf{X})^r e^{itX}\} \quad r = 1, 2, \dots, j$$

   Note that, if j is even, it became an if:

$$\mathbb{E}\{|X|^r\} < \infty \implies \Phi_X \in \mathbb{C}^{(r)}, \quad \Phi_X^{(r)}(t) = \mathbb{E}\{(iX)^r e^{itX}\} \quad r = 2, 4, \dots, j$$

3. **Inversion theorem**: The characteristic function uniquely determines the probability distribution of the correspondent random vector $\mathbf{X}$:

$$\mathbb{F}(b) - \mathbb{F}(a) = \frac{1}{2\pi i} \lim_{c\to\infty} \int_{-c}^{c} \frac{e^{-ita} - e^{itb}}{t} \Phi(t) dt \quad \forall a < b$$

Then, the density function is obtained as:

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-it\mathbf{X}} \Phi_{\mathbf{X}}(t) dt$$

4. **Convergence in distribution**:

$$X_n \xrightarrow[n\to\infty]{d} X \iff \Phi_X(t) = \lim_{n\to\infty} \Phi_{X_n}(t) \quad \forall t \in \mathbb{R}$$

5. **Scaling and centering**: Given $Y = a + bX$, the effect of scaling and centering on the characteristic function is such that:

$$\Phi_Y(t) = \mathbb{E}\{e^{itY}\} = \mathbb{E}\{e^{it(a+bX)}\} = e^{itb}\mathbb{E}\{e^{i(ta)X}\} = e^{itb}\Phi_X(at) \quad \forall a, b, t \in \mathbb{R}$$

6. **Weak Law of Large Numbers**: consider a sequence of IID random variables $\{X_n\}_{n\geq 1}$, such that exists the first derivative of the characteristic function in zero, namely $\exists \phi'_X(0)$ and the first moment of X_1 is finite, i.e. $\mathbb{E}\{|X_1|\} = \mu < \infty$, then the sample mean converges in probability to a degenerate random variable $\alpha \in \mathbb{R}$.

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\text{p}} \alpha$$

for some $\alpha \in \mathbb{R}$.

i WLLN and characteristic function

Proof. To prove the above statement, let's compute the characteristic function of $\bar{\mathbf{X}}_n$:

$$\phi_{\bar{\mathbf{X}}_n}(t) = \mathbb{E}\left\{\exp\left(\frac{it}{n}\sum_{i=1}^{n}\mathbf{X}_n\right)\right\} \overset{\text{IID}}{=} \left[\phi_{\mathbf{X}_1}\left(\frac{t}{n}\right)\right]^n$$

Let's now apply the Taylor series of a function $f(x)$ around the point a (Equation 30.1) to expand till the first order term the function $\phi_{\mathbf{X}_1}(\frac{t}{n})$ around zero ($a = 0$ and $x = \frac{t}{n}$),

i.e.

$$\phi_{\bar{\mathbf{X}}_n}\left(\frac{t}{n}\right) = \left(\phi_{\mathbf{X}_1}(0) + \frac{t}{n}\phi'_{\mathbf{X}_1}(0) + o\left(\frac{t}{n}\right)\right)^n =$$

$$= \left(1 + \frac{t\phi'_{\mathbf{X}_1}(0) + no(\frac{t}{n})}{n}\right)^n \underset{n\to\infty}{\longrightarrow} \exp\left\{t\phi'_{\mathbf{X}_1}(0)\right\}$$

The convergence going to the limit as $n \to \infty$ follows from the fact that in general if $a_n \underset{n\to\infty}{\longrightarrow} a$, then the following limit holds:

$$\left(1 + \frac{a_n}{n}\right)^n \underset{n\to\infty}{\longrightarrow} e^a$$

Therefore, since $\phi'_{\mathbf{X}_1}(0) = i\alpha$ for some $\alpha \in \mathbb{R}$, it follows that:

$$\lim_{n\to\infty} \phi_{\bar{\mathbf{X}}_n}(t) = e^{t\phi'_{\mathbf{X}_1}(0)} = e^{it\alpha} \quad \forall t \in \mathbb{R}$$

Hence, since $e^{it\alpha}$ is the characteristic function of a degenerate random variable, namely a random variable that is constant almost surely, it is possible to conclude that the sample mean converges in distribution to a degenerate random variable α. Moreover, in this specific case in which the limit is a degenerate random variable it can be proved that having convergence in distribution implies also convergence in probability, something that in general is not true. □

7.1 Moment generating function

Definition 7.2. (Moment Generating Function)
Consider an uni dimensional random variable $\mathbf{X}$, then the moment generating function is defined as:

$$\psi_{\mathbf{X}}(t) = \mathbb{E}\{e^{t\mathbf{X}}\} \quad \forall t \in \mathbb{R} - \{0\}$$

Proposition 7.1. *(Moment generating function and sequence of moments)*
Consider a random variable $\mathbf{X}$, such that it's moment generating function exists and it's finite around zero, i.e.

$$\psi_{\mathbf{X}}(t) = \mathbb{E}\{e^{t\mathbf{X}}\} < \infty \quad \epsilon > 0, \forall t \in (-\epsilon, \epsilon)$$

Then this implies that the sequence of moments are finite $\mathbb{E}\{|\mathbf{X}|^n\} < \infty$ for all n and the sequence of moments uniquely determine the distribution of $\mathbf{X}$. According to this result, if we consider another random variable $\mathbf{Y}$ such that $\mathbb{E}\{|\mathbf{X}|^n\} = \mathbb{E}\{|\mathbf{Y}|^n\}$ for all n, then the distribution of $\mathbf{X}$ and $\mathbf{Y}$ is the same, i.e. $\mathbf{X} \sim \mathbf{Y}$.

8 Convergence concepts

Let's consider a sequence of real number, say a_n, then stating that the associated series converges, formally $\sum_{k=1}^{\infty} \alpha_k < \infty$, implies that from a certain k awards $a_k = 0$, i.e.

$$\sum_{k=1}^{\infty} \alpha_k < \infty \iff \lim_{N\to\infty} \sum_{k=n}^{N} a_k = 0$$

8.1 Types of convergence

Definition 8.1. (Pointwise)
A sequence of random variables $\{X_n\}_{n\geq 1}$ is said to be convergent **point wise** to a limit X iff for all $\omega \in \Omega$:

$$X_n(\omega) \underset{n\to\infty}{\longrightarrow} X(\omega) \iff \lim_{n\to\infty} X_n(\omega) = X(\omega)$$

This kind of definition requires that convergence happen for every $\omega \in \Omega$.

Definition 8.2. (Almost Surely)
A sequence of random variables $\{X_n\}_{n\geq 1}$ is said to be convergent **almost surely** to a limit X iff:

$$\mathbb{P}\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\} = 1$$

Usually, such kind of convergence is denoted as:

$$X_n(\omega) \overset{\text{a.s.}}{\underset{n\to\infty}{\longrightarrow}} X(\omega)$$

In other terms, an almost surely convergence implies the relation must holds for all $\omega \in \Omega$ with the exception of some ω's, that are in Ω, but whose probability of occurrence is zero.

Definition 8.3. (In Probability)
A sequence of random variables $\{X_n\}_{n\geq 1}$ is said to be convergent **in probability** to a limit X if, for a fixed $\epsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \epsilon\} = 0$$

Usually, such kind of convergence is denoted as:

$$X_n(\omega) \overset{\text{p}}{\underset{n\to\infty}{\longrightarrow}} X(\omega)$$

Definition 8.4. (L_p)
A sequence of events X_n such that:

$$\mathbb{E}\{|X_n|^p\} < \infty, \quad \mathbb{E}\{|X|^p\} < \infty,$$

is said to be convergent in L_p, with $p > 0$, to a random variable X iff

$$X_n(\omega) \xrightarrow[n\to\infty]{L_p} X(\omega) \iff \lim_{n\to\infty} \mathbb{E}\{|X_n - X|^p\} = 0$$

Usually, such kind of convergence is denoted as:

$$X_n \xrightarrow[n\to\infty]{L_p} X$$

Note that, it can be proved that there is no relation between almost sure convergence and L_p convergence, i.e. one do not imply the other and viceversa. However, a convergence in a bigger space, say $q > s$ implies the convergence in the smaller space, i.e.

$$X_n \xrightarrow[n\to\infty]{L_q} X \implies X_n \xrightarrow[n\to\infty]{L_p} X, \quad 0 < p < q$$

Definition 8.5. (**In Distribution**)

A sequence of random variables X_n is said to be convergent **in distribution** to a random variable X if the distribution of $\mathbb{F}_{X_n}$ *weakly converges* to $\mathbb{F}_X$, i.e.

$$\lim_{n\to\infty} \mathbb{F}_{X_n}(x) = \mathbb{F}_X(x) \quad \forall x$$

where x is a continuity point of $\mathbb{F}$. Usually, such kind of convergence is denoted as:

$$X_n(\omega) \xrightarrow[n\to\infty]{\mathrm{d}} X(\omega)$$

In other terms, we have convergence **in distribution** if the distribution of X_n, namely $\mathbb{F}_{X_n}$, converges as $n \to \infty$ to the distribution of X, namely $\mathbb{F}_X$. Note that the convergence in distribution is not related with probability space but involves only the distribution functions.

8.2 Laws of Large Numbers

There are many versions of laws of large numbers (**LLN**). In general, a sequence $\{X_n\}_{n\geq 1}$ is said to satisfy a LLN iff:

$$\bar{X}_n = \frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow X$$

! Strong vs weak laws of large numbers

In general, if convergence happens **almost surely** (Definition 8.2) we speak about strong laws of large numbers (**SLLN**). Otherwise, if convergence happens **in probability** we speak about weak laws of large numbers (**WLLN**). A crucial difference to be noted is that when convergence happens almost surely we are dealing with a limit of a sequence of sets (limit is inside $\mathbb{P}$), instead if convergence happens in probability we are dealing with a limit of a sequence of real numbers in $[0, 1]$ (limit is outside $\mathbb{P}$).

8.2.1 Strong Laws of Large Numbers

Definition 8.6. (Kolmogorov SLLN)
Let's consider a sequence of IID random variables $\{X_n\}_{n\geq 1}$. Then, there exist a constant $c \in \mathbb{R}$ such that:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{S_n}{n} \xrightarrow[n\to\infty]{\text{a.s.}} c$$

Then, if $\mathbb{E}\{|X_1|\} < \infty$ in which case $c = \mathbb{E}\{|X_1|\}$.

Definition 8.7. (SLLN without independence)
Let's consider a sequence of **identically distributed** random variables $\{X_n\}_{n\geq 1}$, i.e. $\mathbb{E}\{X_n\} = \mathbb{E}\{X_1\}$ for all n, such that:

1. $\mathbb{E}\{X^2\} < c$ where $c > 0$ is a constant independent from n.
2. $\mathbb{C}v\{X_i, X_j\} = 0 \quad \forall i \neq j$.

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{S_n}{n} \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}\{X_1\}$$

Note that the existence of the first moment and the fact that it is finite, i.e. $\mathbb{E}\{X_1\} < \infty$, implies that there exists the characteristic function of the random variable in zero, i.e. $\exists \phi'_{X_1}(0)$. On the other hand, the existence of the characteristic function in zero do not ensure that the first moment is finite.

8.2.2 Weak Laws of Large Numbers

Let's repeat a random experiment many times, every time ensuring the same conditions in such a way that the sequence of the experiment are IID. Then, each random variable X_i comes from the same population with a unknown mean $\mathbb{E}\{X\}$ and variance $\mathbb{V}\{X\}$. Thanks to the WLLN and repeating the experiment many times we have that the sample mean of the experiment

converges in probability to the true mean in population. Convergence in probability means that:

$$\lim_{n\to\infty} \mathbb{P}\left\{\omega \in \Omega : \left|\frac{1}{n}\sum_{i=1}^{n} X_i(\omega) - \mathbb{E}\{X(\omega)\}\right| > \epsilon\right\} = 0$$

Definition 8.8. (**WLLN with variances**)
Given a sequence of **independent** and **identically distributed** random variables $\{X_n\}_{n\geq 1}$ such that:

1. $\mathbb{E}\{X_1\} = \mu$.
2. $\mathbb{E}\{X_1^2\} < \infty$.

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{S_n}{n} \xrightarrow[n\to\infty]{\text{P}} \mathbb{E}\{X_1\} = \mu$$

Proof WLLN with variances

Proof. Let's consider the random variable $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, then since by assumption the mean and variance are finite, let's apply the Chebychev inequality (Equation 5.5), i.e.

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \lambda) \leq \frac{1}{\lambda^2}\mathbb{V}\{\bar{X}_n - \mu\}$$

Using a well known scaling property of variance let's simplify it as:

$$\begin{aligned}
\mathbb{V}\{\bar{X}_n - \mu\} = \mathbb{V}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right\} = & \quad \text{(Constant)} \\
= \mathbb{V}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i\right\} = & \quad \text{(Scaling)} \\
= \frac{1}{n^2}\mathbb{V}\left\{\sum_{i=1}^{n} X_i\right\} = & \quad \text{(Independence)} \\
= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\{X_i\} = & \quad \text{(Identically distribution)} \\
= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} &
\end{aligned}$$

Therefore the Chebychev inequality became

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \lambda) \leq \frac{\sigma^2}{n\lambda^2}$$

Taking the limit as $n \to \infty$ proves the convergence in probability, i.e.

$$\lim_{n\to\infty} \mathbb{P}(|\bar{X}_n - \mu| \geq \lambda) \leq \lim_{n\to\infty} \frac{\sigma^2}{n\lambda^2} = 0$$

□

Definition 8.9. (**Khintchin's WLLN under first moment hypothesis**)
Given a sequence of **independent** and **identically distributed** random variables $\{X_n\}_{n\geq 1}$ such that:

1. $\mathbb{E}\{X_1\} < \infty$.
2. $\mathbb{E}\{X_n\} = \mu$.

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{S_n}{n} \xrightarrow[n\to\infty]{\text{p}} \mathbb{E}\{X_1\} = \mu$$

Definition 8.10. (**Feller's WLLN without first moment**)
Given a sequence of **independent** and **identically distributed** random variables $\{X_n\}_{n\geq 1}$ such that:

$$\lim_{x\to\infty} x\mathbb{P}\{|X_1| > x\} = 0$$

then

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{S_n}{n} \xrightarrow[n\to\infty]{\text{p}} \mathbb{E}\{X_1 \mathbb{1}_{[|X_1|\leq n]}\}$$

Note that this result makes not assumptions about a finite first moment.

i SLLN (without independence) implies WLLN

Let's verity that under the assumptions of the SLLN without independence (Definition 8.7) we will always have convergence in probability, i.e.

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\text{p}} \mathbb{E}\{X_1\}$$

Proof. Using Chebychev inequality (Equation 5.5), fix an $\varepsilon > 0$ such that:

$$\mathbb{P}(|\bar{X}_n - \mathbb{E}\{X_1\}| > \varepsilon) \leq \frac{\mathbb{V}\{\bar{X}_n\}}{\varepsilon^2}$$

Let's explicit the computations, i.e.

$$\frac{\mathbb{V}\{\bar{X}_n\}}{\varepsilon^2} = \frac{1}{n^2\varepsilon^2}\mathbb{V}\left\{\sum_{i=1}^{n} X_i\right\} = \\ = \frac{1}{n^2\varepsilon^2}\left[\sum_{i=1}^{n}\mathbb{V}\{X_i\} + \sum_{i=1}^{n}\sum_{j\neq i}^{n}\mathbb{C}v\{X_i, X_j\}\right]$$

By assumption the covariances are zero $\mathbb{C}v\{X_i, X_j\} = 0 \forall i \neq j$. Moreover, since $\mathbb{V}\{X_i\} = \mathbb{E}\{X_i^2\} - \mathbb{E}\{X_i\}^2$ it is possible to upper bound the variance with the second moment, namely $\mathbb{V}\{X_i\} \leq \mathbb{E}\{X_i^2\}$, i.e.

$$\frac{1}{n^2\varepsilon^2}\sum_{i=1}^{n}\mathbb{V}\{X_i\} \leq \frac{1}{n^2\varepsilon^2}\sum_{i=1}^{n}\mathbb{E}\{X_i^2\}$$

Since by the assumption of the SLLN we have that $\mathbb{E}\{X^2\} < c$ where $c > 0$ is a constant independent from n we can further upper bound the probability by:

$$\frac{1}{n^2\varepsilon^2}\sum_{i=1}^{n}\mathbb{E}\{X_i^2\} \leq \frac{1}{n^2\varepsilon^2}\sum_{i=1}^{n} c = \frac{nc}{n^2\varepsilon^2} = \frac{c}{n\varepsilon^2}$$

Finally if we take the limit for $n \to \infty$ it is equal to zero implying convergence in probability:

$$0 \leq \lim_{n\to\infty}\mathbb{P}(|\bar{X}_n - \mathbb{E}\{X_1\}| > \varepsilon) \leq \lim_{n\to\infty}\frac{c}{n\varepsilon^2} = 0$$

□

8.3 Central Limit Theorem

Theorem 8.1. *(Central Limit Theorem (CLT) - IID case)*
Let's consider a sequence of n random variables, $X_n = (X_1, \dots, X_n)$, where each X_i is independent and identically distributed (IID), i.e.

$$X_i \sim IID(\mu, \sigma^2) \implies \mathbb{E}\{X_i\} = \mathbb{E}\{X_1\} = \mu \\ \implies \mathbb{V}\{X_i\} = \mathbb{V}\{X_1\} = \sigma^2$$

Then, let's define a random variable, namely S_n, given by the sum of all the X_i, i.e.

$$S_n = \sum_{i=1}^{n} X_i$$

It is easy to see that due to the fact that the random variables are IID the moments of S_n are:

$$\mathbb{E}\{S_n\} = n\mathbb{E}\{X_1\} = n\mu, \quad \mathbb{V}\{S_n\} = n\mathbb{V}\{X_1\} = n\sigma^2$$

Hence, the standardized variable Z_n on large samples is normally distributed, i.e.

$$Z_n = \frac{S_n - \mathbb{E}\{S_n\}}{\sqrt{\mathbb{V}\{S_n\}}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\,\sigma} \underset{n\to\infty}{\overset{d}{\sim}} \mathcal{N}(0,1)$$

Part II

Statistics

9 Population, sample and moments

A **population** refers to the entire group of individuals or instances about whom we hope to learn. It encompasses all possible subjects or observations that meet a set of criteria. The population is the complete set of items that interest the researcher, and it can be **finite** (e.g. the students in a particular school) or **infinite** (e.g. the number of times a die can be rolled). A population **size** is given by the number of distinct elements and it includes every individual or observation of interest.

A **sample** is a subset of the population that is used to represent the population. Since studying an entire population is often impractical due to constraints like time, cost, and accessibility, samples provide a manageable and efficient way to gather data and make inferences about the population. It is important that the sample is **representative** of the population of interest to allow for valid inferences. It is always important to distinguish between a **random sample**, e.g. a random group of students in 5th year from a school to make inference about the students at the 5th year of such school, and a **convenience sample**, e.g. a class of 5th year students who are easily accessible to the researcher, but that can be not representative of all the 5th year students in the school.

Aspect	**Population**	**Sample**
Definition	Entire group of interest	Subset of the population
Size	Large, potentially infinite	Small, manageable
Data Collection	Often impractical to study directly	Practical and feasible
Purpose	To understand the whole group	To make inferences about the population

9.1 Expectation

The expectation of a random variable X is it's first moment, also called *statistical average*. In general, it is denoted as $\mathbb{E}\{X\}$. Let's consider a discrete random variable X with distribution function $P(X = x_j) = p_j$. Then the **expectation** of X is the weighted average between all

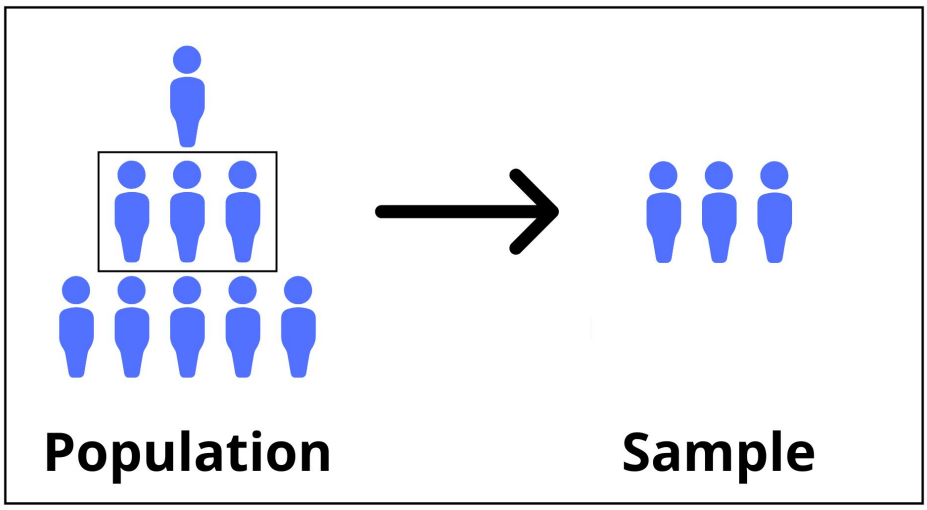


Figure 9.1: Population vs sample.

the possible m-states that the random variable can assume by it's respective probability of occurrence, i.e.

$$\mathbb{E}\{X\} = \sum_{j=1}^{m} x_j p_j.$$

In the continuous case, i.e. when X takes values in $\mathbb{R}$ and admits a density function, the expectation is computed as an integral, i.e.

$$\mathbb{E}\{X\} = \int_{-\infty}^{\infty} x dF_X(x) = \int_{-\infty}^{\infty} x f_X(x) dx.$$

9.1.1 Sample statistic

Let's consider a sample of IID observations, i.e. $X_n = (x_1, \dots, x_i, \dots, x_n)$. Then the sample expectation is computed as:

$$\hat{\mu}(X_n) = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Population vs sample

In general, the notation X_n refers to a finite sample, e.g. $\hat{\mu}(X_n)$ is the sample mean. Instead the notation without n, i.e. X, stands for the random variable in population, e.g. $\mathbb{E}\{X\}$ is the mean in population. A population can be **finite** or **non-finite**. In the case of a finite population with N element it is useful to distinguish between:

- Extraction **with reimmission** of n elements for the sample gives N^n possible combinations.
- Extraction **without readmission** of n elements for the sample gives $\binom{N}{n}$ possible combinations.

Table 9.2: Expectation in a discrete and continuous population and in a sample X_n.

Population (*continuous*)	**Population** (*discrete*)	*Sample*
$\int_{-\infty}^{\infty} x f(x) dx$	$\sum_{j=1}^{m} x_j p_j$	$\frac{1}{n}\sum_{i=1}^{n} x_i$

9.1.2 Sample moments

Let's consider an the moments of the sample mean of an IID sample. Since all the variables has the same expected value, i.e. $\mathbb{E}\{x_i\} = \mathbb{E}\{X\}$, the expected value of the sample mean is computed as:

$$\mathbb{E}\left\{\hat{\mu}(X_n)\right\} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\{x_i\} = \mathbb{E}\{X\}. \tag{9.1}$$

The variance of the sample mean is computed as:

$$\begin{aligned}\mathbb{V}\left\{\hat{\mu}(X_n)\right\} &= \frac{1}{n^2}\mathbb{V}\left\{\sum_{i=1}^{n} x_i\right\} = \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\left\{x_i\right\} = \frac{\mathbb{V}\{X\}}{n}\end{aligned} \tag{9.2}$$

9.1.3 Sample distribution

Proposition 9.1. *Let's consider a* ***sample*** X_n *of* n *IID random variables. If* n *is sufficiently large, independently from the distribution of the* X*, by the central limit theorem (CLT) the distribution of the sample expectation converges to the distribution of a normal random variable, i.e.*

$$\hat{\mu}(X_n) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(\mathbb{E}\{X\}, \frac{\mathbb{V}\{X\}}{n}\right).$$

Proof: Distribution of sample expectation (Proposition 9.1)

Proof. In order to prove Proposition 9.1 it is useful to compute the expectation and the

variance of the following random variable, i.e.

$$S_n = \sum_{i=1}^{n} x_i.$$

The expectation and the variance of S_n can be easily obtained from Equation 9.1 and Equation 9.2 respectively and read:

$$\mathbb{E}\{S_n\} = n \cdot \mathbb{E}\{X\} \qquad \mathbb{V}\{S_n\} = n \cdot \mathbb{V}\{X\}$$

Applying the central limit theorem (Theorem 8.1) one obtain:

$$\frac{S_n - n \cdot \mathbb{E}\{X\}}{\sqrt{n} \cdot \mathbb{S}d\{X\}} = \frac{\frac{S_n}{n} - \mathbb{E}\{X\}}{\frac{\mathbb{S}d\{X\}}{\sqrt{n}}} \sim N(0,1).$$

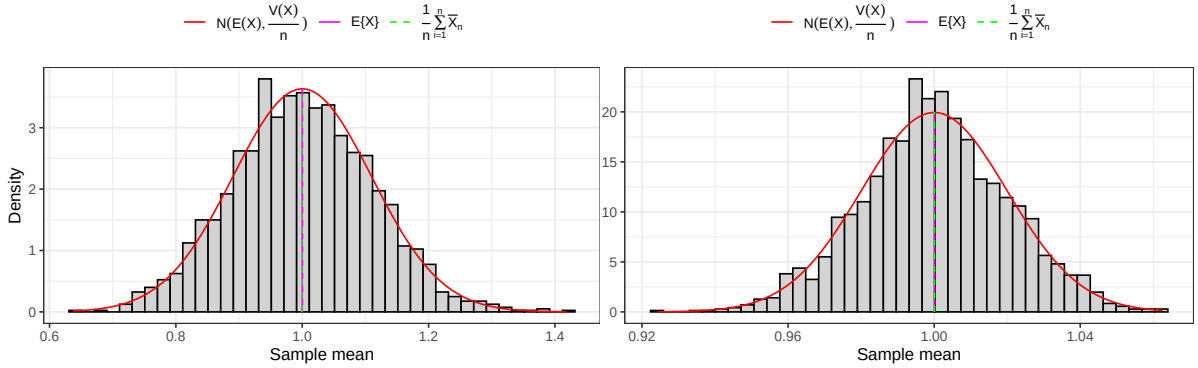
Hence the random variable mean $\hat{\mu}(X_n) = \frac{S_n}{n}$ on large samples is distributed as a normal random variable, i.e.

$$\hat{\mu}(X_n) = \frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow[n\to\infty]{d} \mathcal{N}\left(\mathbb{E}\{X\}, \frac{\mathbb{V}\{X\}}{n}\right).$$

Note that *on small sample* this results holds true if and only if X is normally distributed also in population. Under normality also in population we have that independently from the sample size:

$$X_i \sim \mathcal{N}(\mathbb{E}\{X\}, \mathbb{V}\{X\}),\ \forall i \implies \hat{\mu}(X_n) \sim \mathcal{N}\left(\mathbb{E}\{X\}, \frac{\mathbb{V}\{X\}}{n}\right).$$

□



(a) Small sample (166).

(b) Large sample (5000).

Figure 9.2: Distribution of the sample mean.

9.2 Variance and covariance

In general the variance of a random variable in population defined as:

$$\mathbb{V}\{X\} = \mathbb{E}\{(X - \mathbb{E}\{X\})^2\}.$$

Let's consider a discrete random variable X with distribution function $P(X = x_j) = p_j$. Then the variance of X is the weighted average between all the possible m-centered and squared states that the random variable can assume by it's respective probability of occurrence, i.e.

$$\mathbb{V}\{X\} = \sum_{j=1}^{m} (x_j - \mathbb{E}\{X\})^2 p_j.$$

In the continuous case, i.e. when X admits a density function and takes values in $\mathbb{R}$, the expectation is computed as:

$$\mathbb{V}\{X\} = \int_{-\infty}^{\infty} (x - \mathbb{E}\{X\})^2 f_X(x) dx.$$

Let's consider two random variables X and Y. Then, in general their covariance is defined as:

$$\mathbb{C}v\{X, Y\} = \mathbb{E}\{(X - \mathbb{E}\{X\})(Y - \mathbb{E}\{Y\})\}.$$

In the discrete case where X and Y have a joint distribution $\mathbb{P}(X = x_i, Y = y_j) = p_{ij}$, their covariance is defined as:

$$\mathbb{C}v\{X, Y\} = \sum_{i=1}^{m} \sum_{j=1}^{s} (x_i - \mathbb{E}\{X\}) \left(y_j - \mathbb{E}\{Y\}\right) p_{ij}.$$

In the continuous case, if the joint distribution of X and Y admits a density function the covariance is computed as:

$$\mathbb{C}v\{X, Y\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mathbb{E}\{X\})(y - \mathbb{E}\{Y\}) f_{X,Y}(x, y) dx dy.$$

9.2.1 Properties

There are several properties connected to the variance.

1. The variance can be computed as:

$$\mathbb{V}\{X\} = \mathbb{E}\{X^2\} - \mathbb{E}\{X\}^2. \tag{9.3}$$

2. The variance is invariant with respect to the addition of a constant a, i.e.

$$\mathbb{V}\{a + X\} = \mathbb{V}\{X\}. \tag{9.4}$$

3. The variance scales upon multiplication with a constant a, i.e.

$$\mathbb{V}\{aX\} = a^2\mathbb{V}\{X\}. \tag{9.5}$$

4. The variance of the sum is computed as:

$$\mathbb{V}\{X+Y\} = \mathbb{V}\{X\} + \mathbb{V}\{Y\} + 2\mathbb{C}v\{X,Y\}. \tag{9.6}$$

5. The covariance can be expressed as:

$$\mathbb{C}v\{X,Y\} = \mathbb{E}\{XY\} - \mathbb{E}\{X\}\mathbb{E}\{Y\}. \tag{9.7}$$

6. The covariance scales upon multiplication with a constant a and b, i.e.

$$\mathbb{C}v\{aX,bY\} = ab\mathbb{C}\{X,Y\}. \tag{9.8}$$

Proof: Properties of the variance

Proof. The property 1. (Equation 9.3) follows easily developing the definition of variance, i.e.

$$\begin{aligned}\mathbb{V}\{X\} &= \mathbb{E}\{(X-\mathbb{E}\{X\})^2\} = \\ &= \mathbb{E}\{X^2\} + \mathbb{E}\{X\}^2 - 2\mathbb{E}\{X\}^2 = \\ &= \mathbb{E}\{X^2\} - \mathbb{E}\{X\}^2\end{aligned}$$

The property 2. (Equation 9.4) follows from the definition, i.e.

$$\begin{aligned}\mathbb{V}\{a+X\} &= \mathbb{E}\{(a+X-\mathbb{E}\{a+X\})^2\} = \\ &= \mathbb{E}\{(X-\mathbb{E}\{X\})^2\} = \\ &= \mathbb{V}\{X\}\end{aligned}$$

The property 3. (Equation 9.5) follows using the expression of the variance in Equation 9.3, i.e.

$$\begin{aligned}\mathbb{V}\{aX\} &= \mathbb{E}\{(aX)^2\} - \mathbb{E}\{aX\}^2 = \\ &= a^2\mathbb{E}\{X^2\} - a^2\mathbb{E}\{X\}^2 = \\ &= a^2(\mathbb{E}\{X^2\} - \mathbb{E}\{X\}^2) = \\ &= a^2\mathbb{V}\{X\}\end{aligned}$$

The property 4. (Equation 9.6), i.e. the variance of the sum of two random variables is:

$$\begin{aligned}\mathbb{V}\{X+Y\} &= \mathbb{E}\{(X+Y-\mathbb{E}\{X+Y\})^2\} = \\ &= \mathbb{E}\{([X-\mathbb{E}\{X\}]+[Y-\mathbb{E}\{Y\}])^2\} = \\ &= \mathbb{E}\{(X-\mathbb{E}\{X\})^2\} + \mathbb{E}\{(Y-\mathbb{E}\{Y\})^2\} + 2\mathbb{E}\{(X-\mathbb{E}\{X\})(Y-\mathbb{E}\{Y\})\} = \\ &= \mathbb{V}\{X\} + \mathbb{V}\{Y\} + 2\mathbb{C}v\{X,Y\}\end{aligned}$$

where in the case in which there is no linear connection between X and Y the covariance is zero, i.e. $\mathbb{C}v\{X,Y\} = 0$. Developing the computation of the covariance it is possible

to prove property 5. (Equation 9.7), i.e.

$$\begin{aligned}\mathbb{C}v\{X,Y\} &= \mathbb{E}\{(X-\mathbb{E}\{X\})(Y-\mathbb{E}\{Y\})\} = \\ &= \mathbb{E}\{XY - X\mathbb{E}\{Y\} - Y\mathbb{E}\{X\} + \mathbb{E}\{X\}\mathbb{E}\{Y\}\} = \\ &= \mathbb{E}\{XY\} - 2\mathbb{E}\{X\}\mathbb{E}\{Y\} + \mathbb{E}\{X\}\mathbb{E}\{Y\} = \\ &= \mathbb{E}\{XY\} - \mathbb{E}\{X\}\mathbb{E}\{Y\}\end{aligned}$$

Finally, using the result in property 5. (Equation 9.7) the result in property 6. (Equation 9.8) follows easily:

$$\begin{aligned}\mathbb{C}v\{aX,bY\} &= \mathbb{E}\{aXbY\} - \mathbb{E}\{aX\}\mathbb{E}\{bY\} = \\ &= ab\mathbb{E}\{XY\} - ab\mathbb{E}\{X\}\mathbb{E}\{Y\} = \\ &= ab\mathbb{C}v\{X,Y\}\end{aligned}$$

□

9.2.2 Sample statistic

The **sample's variance** on $X_n = (x_1, \dots, x_i, \dots, x_n)$ is computed as:

$$\mathbb{V}\{X_n\} = \hat{\sigma}^2(X_n) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mathbb{E}\{X_n\}\right)^2. \tag{9.9}$$

Equivalently, in terms of the first and second moment:

$$\hat{\sigma}^2(X_n) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2. \tag{9.10}$$

In general, the variance computed as in Equation 9.9 is **not correct** for the population value. Hence, to correct the estimator let's define the **corrected sample's variance**:

$$\hat{s}^2(X_n) = \frac{n}{n-1}\hat{\sigma}^2(X_n). \tag{9.11}$$

9.2.3 Sample moments

Let's consider an the moments of the sample variance on an IID sample. The expected value of the corrected sample variance:

$$\mathbb{E}\left\{\hat{s}^2(X_n)\right\} = \sigma^2. \tag{9.12}$$

The variance of the corrected sample variance is:

$$\mathbb{V}\left\{\hat{s}^2(X_n)\right\} = \frac{\sigma^4}{n}\left((\frac{\mu_4}{\sigma^4} - 3) + 2\frac{n}{n-1}\right), \tag{9.13}$$

where $\frac{\mu_4}{\sigma^4}$ is the kurtosis of X_n. If the population is normal, $\frac{\mu_4}{\sigma^4} = 3$ and the variance simplifies in:

$$\mathbb{V}\{\hat{s}^2(X_n)\} = \frac{2\sigma^4}{n-1}. \tag{9.14}$$

9.2.4 Sample distribution

The distribution of the sample variance is available when we consider the sum of n-IID standard normal random variables. Notably, from Cochran's theorem:

$$T_n = (n-1)\frac{\hat{s}^2(X_n)}{\sigma^2} \sim \chi^2(n-1) \tag{9.15}$$

Going to the limit as $\nu \to \infty$ a χ^2_ν random variable converges to a standard normal random variable, i.e.

$$\frac{\chi^2(n) - n}{\sqrt{2n}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$$

therefore, on large samples the statistic T_n converges to a normal random variable, i.e.

$$T_n \xrightarrow[n\to\infty]{d} \mathcal{N}(n, 2n) \iff \frac{T_n - n}{\sqrt{2n}} \sim \mathcal{N}(0,1) \tag{9.16}$$

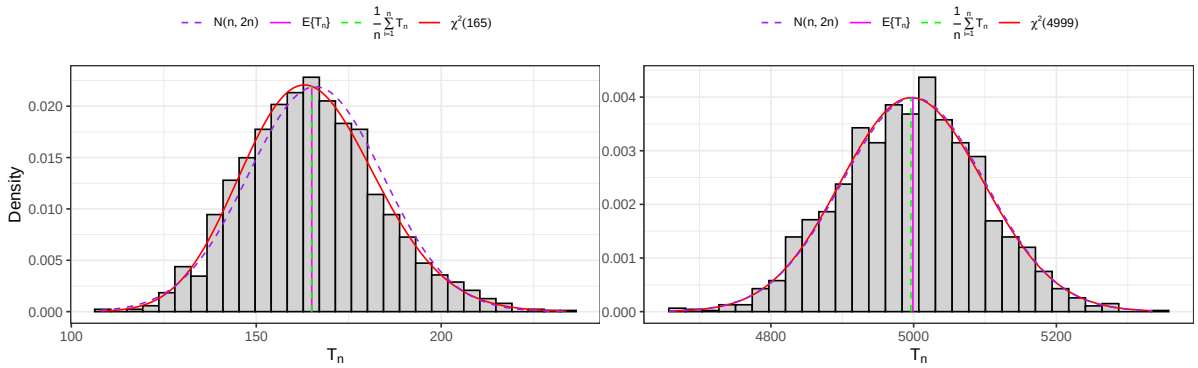
Distribution of $\hat{s}^2(X_n)$ under normality.

If the population X_n is normal, then the distribution of $\hat{s}^2(X_n)$ is proportional to the distribution of a χ^2_{n-1}. In fact, from Equation 9.15 the expectation of $\hat{s}^2(X_n)$ is:

$$\begin{aligned}
\mathbb{E}\{T_n\} &= (n-1)\frac{\mathbb{E}\{\hat{s}^2(X_n)\}}{\sigma^2} \\
&\implies \mathbb{E}\{\hat{s}^2(X_n)\} = \frac{\sigma^2\mathbb{E}\{T_n\}}{n-1} = \frac{\sigma^2(n-1)}{n-1} \\
&\implies \mathbb{E}\{\hat{s}^2(X_n)\} = \sigma^2
\end{aligned}$$

Similarly, computing the variance of Equation 9.15 and knowing that $\mathbb{V}\{T_n\} = 2(n-1)$ one obtain:

$$\begin{aligned}
\mathbb{V}\{T_n\} &= (n-1)^2\frac{\mathbb{V}\{\hat{s}^2(X_n)\}}{\sigma^4} \\
&\implies \mathbb{V}\{\hat{s}^2(X_n)\} = \frac{\sigma^4\mathbb{V}\{T_n\}}{(n-1)^2} = \frac{\sigma^4 2(n-1)}{(n-1)^2} \\
&\implies \mathbb{V}\{\hat{s}^2(X_n)\} = \frac{2\sigma^4}{n-1}
\end{aligned}$$



(a) Small sample (166). (b) Large sample (5000).

Figure 9.3: Distribution of the statistic T_n under normality.

9.3 Skewness

The skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive, zero, negative, or undefined. For a uni modal distribution, negative skew commonly indicates that the tail is on the left side of the distribution, and positive skew indicates that the tail is on the right.

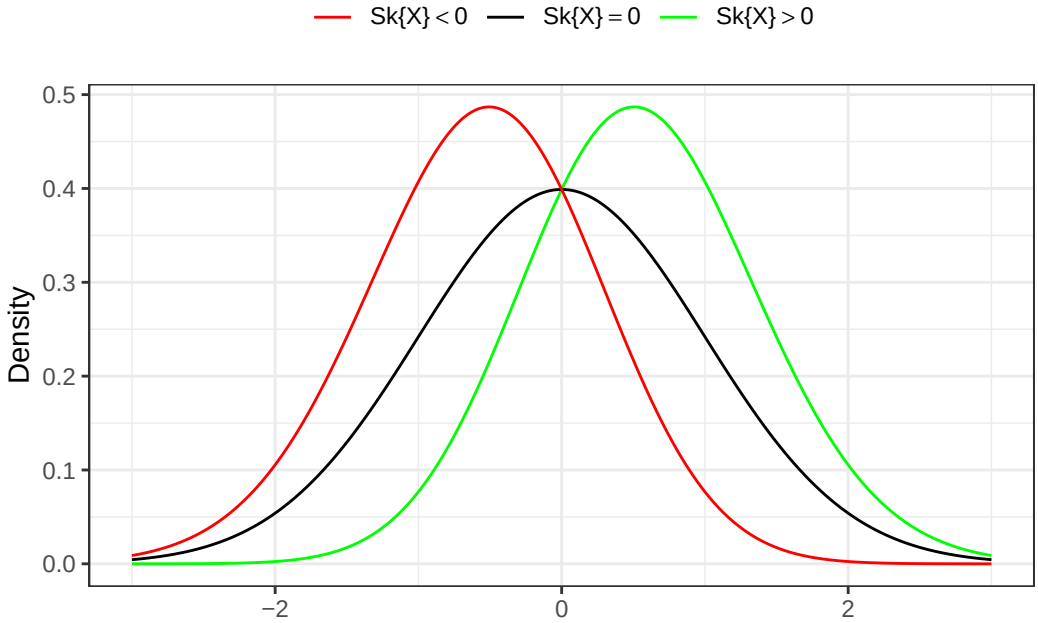


Figure 9.4: Skewness of a random variable.

Following the same notation as in Ralph B. D'agostino and Jr. (1990), let's define and denote

the **population skewness** of a random variable X as:

$$\mathbb{S}k\{X\} = \beta_1(X) = \mathbb{E}\left\{\left(\frac{X - \mathbb{E}\{X\}}{\sqrt{\mathbb{V}\{X\}}}\right)^3\right\},$$

9.3.1 Sample statistic

Let's consider an IID sample $X_n = (x_1, \dots, x_i, \dots, x_n)$, then the **sample's skewness** is estimated as:

$$\mathbb{S}k\{X_n\} = b_1(X_n) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mathbb{E}\{X_n\}}{\sqrt{\mathbb{V}\{X_n\}}}\right)^3. \tag{9.17}$$

The estimator in Equation 9.17 is **not correct**. Hence, let's define the **correct** sample estimator of the skewness as:

$$g_1(X_n) = \frac{\sqrt{n(n-1)}}{(n-2)} b_1(X_n).$$

9.3.2 Sample moments

Under normality, the **asymptotic moments** of the sample skewness are:

$$\mathbb{E}\{b_1(X_n)\} = 0, \quad \mathbb{V}\{b_1(X_n)\} = \frac{6}{n}.$$

In Urzúa (1996) are also reported the exact mean of the estimator in Equation 9.17 for **small normal samples**, i.e.

$$\mathbb{E}\{b_1(X_n)\} = 0,$$

and variance

$$\mathbb{V}\{b_1(X_n)\} = \frac{6(n-2)}{(n+1)(n+3)}. \tag{9.18}$$

9.3.3 Sample distribution

Under normality, the asymptotic distribution of the sample skewness is normal i.e.

$$b_1(X_n) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \frac{6}{n}\right). \tag{9.19}$$

9.4 Kurtosis

The kurtosis is a measure of the *tailedness* of the probability distribution of a real-valued random variable. The standard measure of a distribution's kurtosis, originating with Karl Pearson is a scaled version of the fourth moment of the distribution. This number is related to the tails of the distribution. For this measure, higher kurtosis corresponds to greater extremity of deviations from the mean (or outliers). In general, it is common to compare the excess kurtosis of a distribution with respect to the normal distribution (with kurtosis equal to 3). It is possible to distinguish 3 cases:

1. A **negative excess kurtosis** or **platykurtic** are distributions that produces less outliers than the normal. distribution.
2. A **zero excess kurtosis** or **mesokurtic** are distributions that produces same outliers than the normal.
3. A **positive excess kurtosis** or **leptokurtic** are distributions that produces more outliers than the normal.

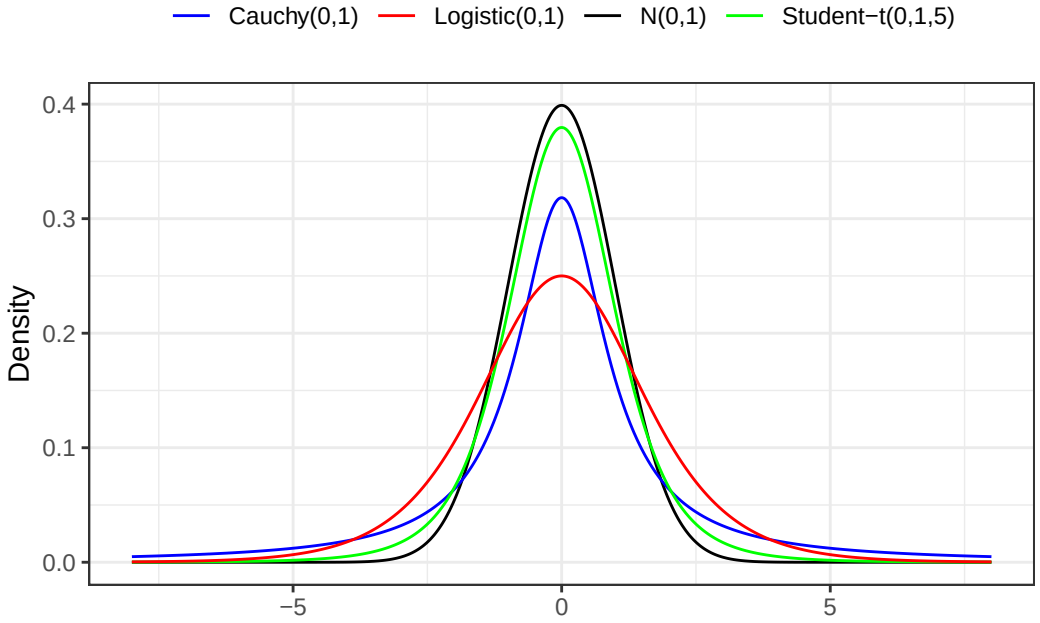


Figure 9.5: Kurtosis of a different leptokurtic distributions.

Let's define and denote the **population kurtosis** of a random variable X as:

$$\mathbb{K}t\{X\} = \beta_2(X) = \mathbb{E}\left\{\left(\frac{X - \mathbb{E}\{X\}}{\sqrt{\mathbb{V}\{X\}}}\right)^4\right\},$$

or equivalently the **excess kurtosis** as $\mathbb{K}t\{X\} - 3$.

9.4.1 Sample statistic

Let's consider an IID sample $X_n = (x_1, \dots, x_i, \dots, x_n)$, then the **sample's kurtosis** is denoted as:

$$\mathbb{K}t\{X_n\} = b_2(X_n) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mathbb{E}\{X_n\}}{\sqrt{\mathbb{V}\{X_n\}}}\right)^4. \tag{9.20}$$

From Pearson (1931), we have a correct the version of $b_1(X_n)$ defined as:

$$g_2(X_n) = \left[b_2(X_n) - \frac{3(n+1)}{n+1}\right]\frac{(n+1)(n-1)}{(n-2)(n-3)}.$$

9.4.2 Sample moments

Under normality, the **asymptotic moments** of the sample kurtosis are:

$$\mathbb{E}\{b_2(X_n)\} = 3, \quad \mathbb{V}\{b_2(X_n)\} = \frac{24}{n}.$$

Notably in Urzúa (1996) are reported also the exact mean and variance for a small normal sample, i.e.

$$\mathbb{E}\{b_2(X_n)\} = \frac{3(n-1)}{(n+1)}, \tag{9.21}$$

and the variance as:

$$\mathbb{V}\{b_2(X_n)\} = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}. \tag{9.22}$$

9.4.3 Sample distribution

Under normality, the asymptotic distribution of the sample kurtosis is normal, i.e.

$$b_2(X_n) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(3, \frac{24}{n}\right). \tag{9.23}$$

10 Likelihood

In general, we define the likelihood of a sample $X_1 ... X_n$ their joint density, function of a general parameter θ and denoted as

$$\mathcal{L}(\theta) = \mathcal{L}(\theta \mid X_n) = \mathcal{L}(\theta \mid x_1, ..., x_n) = f_X(x_1, ..., x_n \mid \theta)$$

For a given value of the parameter θ, the likelihood tells us how likely it is that the data are generated under the distributive law implied by f_X.

10.1 Maximum likelihood estimators

In statistics, thw maximum likelihood estimation (MLE) is a method of estimating the parameters of an assumed probability distribution, given some observed data. This is achieved by maximizing a likelihood function so that, under the assumed statistical model, the observed data is most probable. For example, let's consider a generic sample $X_n = (x_1, ..., x_n)$ drown from a parametric distribution with unknown parameters θ. Then, given the likelihood function, if the observations are independent and identically distributed, then the following factorization of the joint density holds true, i.e.

$$\begin{aligned}\mathcal{L}(\theta \mid X_n) &= f_X(x_1, ..., x_n \mid \theta) = \\ &= f_X(x_1 \mid \theta) ... f_X(x_n \mid \theta) = \\ &= \prod_{i=1}^{n} f_X(x_i \mid \theta)\end{aligned}$$

Then, the **log-likelihood function** is computed taking the logarithm of the likelihood, i.e.

$$\ell(\theta \mid X_n) = \log \mathcal{L}(\theta \mid X_n) = \sum_{i=1}^{n} \log f_X(x_i \mid \theta)$$

If the likelihood function is differentiable, the derivative test for finding maxima can be applied. Since the logarithm is a monotonic function, the maximum of $\ell(\theta \mid X_n)$ occurs at the same value of $\mathcal{L}(\theta \mid X_n)$. Considering a vector of k-parameters the **first order conditions**(**FOC**) and if

the log-likelihood function is is differentiable in θ, a sufficient conditions for the occurrence of a maximum (or a minimum) are

$$\begin{cases} \partial_{\theta_1} \ell(\theta \mid X_n) = 0 \\ \partial_{\theta_2} \ell(\theta \mid X_n) = 0 \\ \quad \vdots \\ \partial_{\theta_k} \ell(\theta \mid X_n) = 0 \end{cases}$$

Whether the identified the optimal solution $\hat{\theta}$ of the likelihood equations is indeed a (local) maximum depends on whether the matrix of second-order partial and cross-partial derivatives, the so-called **Hessian matrix**, i.e.

$$\mathcal{H}(\hat{\theta}) = \begin{pmatrix} \partial_{\theta_1}\partial_{\theta_1} \ell(\theta \mid X_n) \Big|_{\theta=\hat{\theta}} & \partial_{\theta_1}\partial_{\theta_2} \ell(\theta \mid X_n) \Big|_{\theta=\hat{\theta}} & \dots & \partial_{\theta_1}\partial_{\theta_k} \ell(\theta \mid X_n) \Big|_{\theta=\hat{\theta}} \\ \partial_{\theta_2}\partial_{\theta_1} \ell(\theta \mid X_n) \Big|_{\theta=\hat{\theta}} & \partial_{\theta_2}\partial_{\theta_2} \ell(\theta \mid X_n) \Big|_{\theta=\hat{\theta}} & \dots & \partial_{\theta_2}\partial_{\theta_k} \ell(\theta \mid X_n) \Big|_{\theta=\hat{\theta}} \\ \vdots & \vdots & \ddots & \vdots \\ \partial_{\theta_k}\partial_{\theta_1} \ell(\theta \mid X_n) \Big|_{\theta=\hat{\theta}} & \partial_{\theta_k}\partial_{\theta_2} \ell(\theta \mid X_n) \Big|_{\theta=\hat{\theta}} & \dots & \partial_{\theta_k}\partial_{\theta_k} \ell(\theta \mid X_n) \Big|_{\theta=\hat{\theta}} \end{pmatrix}$$

that has to be negative semi-definite at $\hat{\theta}$ denoting a local concavity. In some cases, the first-order conditions of the likelihood function can be solved analytically.

10.2 Example: MLE in the Gaussian case

In the context of maximum likelihood estimation (MLE) for a normal (Gaussian) random variable, let's consider a set of n independent and identically distributed (i.i.d.) random variables $X_n = \{x_1, x_2, \dots, x_n\}$ drawn from a normal distribution with unknown mean μ and variance σ^2. The probability density function (pdf) of a normal distribution is given by:

$$f(x_i \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

The likelihood function $\mathcal{L}(\mu, \sigma^2; X_n)$ is the joint probability of the observed data, viewed as a function of the parameters μ and σ^2, i.e.

$$\mathcal{L}(\mu, \sigma^2 \mid X_n) = \prod_{i=1}^{n} f(x_i \mid \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

The log-likelihood function $\ell(\mu, \sigma^2)$ is the natural logarithm of the likelihood function:

$$\ell(\mu, \sigma^2 \mid X_n) = \log \mathcal{L}(\mu, \sigma^2 \mid X_n) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)\right)$$

Simplifying the log-likelihood function:

$$\ell(\mu, \sigma^2 \mid X_n) = \sum_{i=1}^{n} \left(-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{(x_i - \mu)^2}{2\sigma^2} \right) =$$
$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

To find the maximum likelihood estimates, we have to solve the partial derivatives of $\ell(\mu, \sigma^2 \mid X_n)$ with respect to μ and σ^2, setting it equal to zero.

1. **Condition for the mean (μ)**:

$$\frac{\partial \ell(\mu, \sigma^2 \mid X_n)}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0$$
$$\implies \mu^{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

2. **Condition for the variance (σ^2)**:

$$\frac{\partial \ell(\mu, \sigma^2 \mid X_n)}{\partial \sigma^2} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 =$$
$$\implies -n\sigma^2 + \sum_{i=1}^{n}(x_i - \mu)^2 = 0$$
$$\implies (\sigma^{MLE})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

MLE in the Gaussian case

Example 10.1. Let's consider a normal random sample with n-observation. Let's consider the variance of the distribution known. Then, we can estimate the maximum likelihood mean maximizing the log-likelihood.

Table 10.1: Moments estimate on the sample

Mean	Variance	Std.deviation
0.9936231	4.216398	2.053387

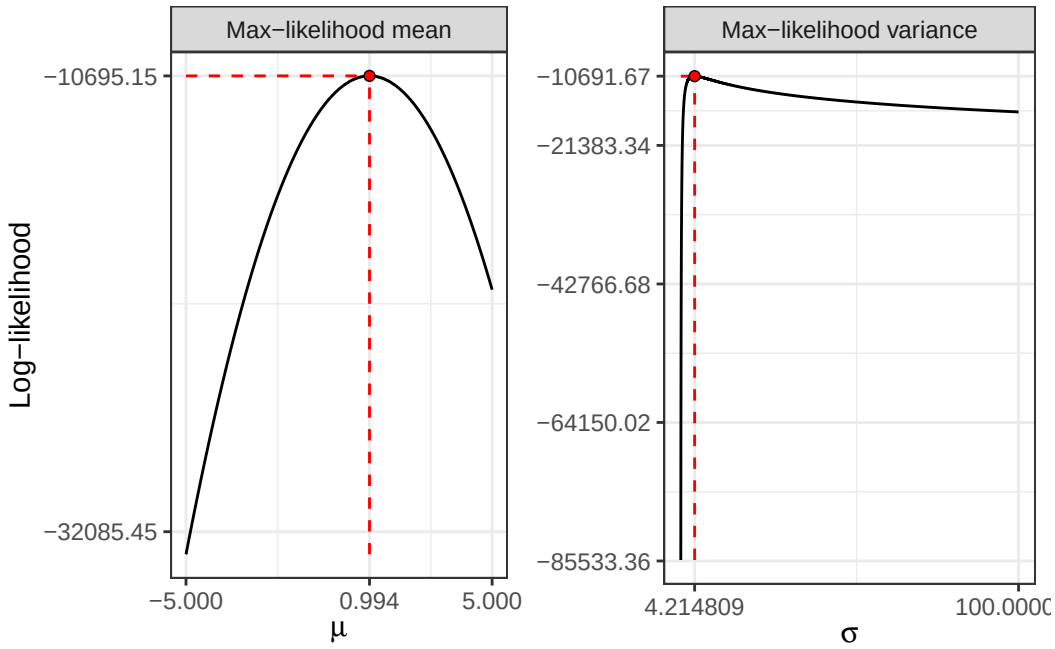


Figure 10.1: Log-likelihood function for a normal sample.

11 Multivariate data

Let's consider a matrix $\mathbf{X}$ with n-observations and k-variables. Then, let's define some useful operations that can be performed on the matrix.

$$\underset{n\times k}{\mathbf{X}} = \begin{pmatrix} x_{1,1} & \dots & x_{1,j} & \dots & x_{1,k} \\ \vdots & & \vdots & & \vdots \\ x_{i,1} & \dots & x_{i,j} & \dots & x_{i,k} \\ \vdots & & \vdots & & \vdots \\ x_{n,1} & \dots & x_{n,j} & \dots & x_{n,k} \end{pmatrix} \tag{11.1}$$

11.1 Vector of means

Let's consider the matrix $\mathbf{X}$ (Equation 11.1), then the vector of means for each column is computed as:

$$\underset{k\times 1}{\bar{\mathbf{x}}} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_k \end{pmatrix} = \left(\tfrac{1}{n}\mathbf{J}_{1,n}\mathbf{X}\right)^\top = \frac{1}{n}\mathbf{X}^\top\mathbf{J}_{n,1} \tag{11.2}$$

where $\mathbf{J}_{n,1}$ is defined as in Equation 32.2.

11.2 Deviation matrix

Let's compute the matrix of centered observations, where each element is computed as $\tilde{x}_{i,j} = x_{i,j} - \bar{x}_j$, i.e.

$$\underset{n\times k}{\tilde{\mathbf{X}}} = \begin{pmatrix} \tilde{x}_{1,1} & \dots & \tilde{x}_{1,j} & \dots & \tilde{x}_{1,k} \\ \vdots & & \vdots & & \vdots \\ \tilde{x}_{i,1} & \dots & \tilde{x}_{i,j} & \dots & \tilde{x}_{i,k} \\ \vdots & & \vdots & & \vdots \\ \tilde{x}_{n,1} & \dots & \tilde{x}_{n,j} & \dots & \tilde{x}_{n,k} \end{pmatrix} \tag{11.3}$$

In matrix notation, it is possible to compute $\tilde{\mathbf{X}}$ as:

$$\begin{aligned}\tilde{\mathbf{X}} &= \mathbf{X} - \mathbf{J}_{1,n}\bar{\mathbf{x}}^\top = \\ &= \mathbf{X} - \mathbf{J}_{1,n}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{J}_{1,n}\right)^\top = \\ &= \mathbf{X} - \frac{1}{n}\mathbf{J}_{n,1}\mathbf{J}_{1,n}\mathbf{X} = \\ &= \left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n\right)\mathbf{X} = \mathbf{A}\mathbf{X}\end{aligned}$$

where $\mathbf{I}_n$ the identity matrix (Equation 32.3) and $\mathbf{J}_{1,n}$ a matrix of ones (Equation 32.2). $\mathbf{A}$ is called *centering matrix* and it is formally defined as:

$$\mathbf{A} = \mathbf{I}_n - \frac{1}{n}\mathbf{J}_n \tag{11.4}$$

11.3 Variance-covariance matrix

Remembering the case for two vectors $\mathbf{x}_k$ and $\mathbf{x}_h$, the covariance is defined as:

$$\begin{aligned}\mathbb{C}v\{\mathbf{x}_k, \mathbf{x}_h\} &= \frac{1}{n}\sum_{i=1}^{n}(x_{i,k} - \bar{x}_k)(x_{i,h} - \bar{x}_h) = \\ &= \frac{1}{n}\sum_{i=1}^{n}\tilde{x}_{i,k}\tilde{x}_{i,h} = \frac{\tilde{\mathbf{x}}_k^\top\tilde{\mathbf{x}}_h}{n}\end{aligned}$$

For a matrix $\mathbf{X}_{n\times k}$, the covariance became a matrix $k \times k$ of the form:

$$\underset{k\times k}{\mathbb{C}v\{\mathbf{X}\}} = \begin{pmatrix} \mathbb{V}\{\mathbf{x}_1\} & \dots & \mathbb{C}v\{\mathbf{x}_1, \mathbf{x}_j\} & \dots & \mathbb{C}v\{\mathbf{x}_1, \mathbf{x}_k\} \\ \vdots & & \vdots & & \vdots \\ \mathbb{C}v\{\mathbf{x}_j, \mathbf{x}_1\} & \dots & \mathbb{V}\{\mathbf{x}_j\} & \dots & \mathbb{C}v\{\mathbf{x}_j, \mathbf{x}_k\} \\ \vdots & & \vdots & & \vdots \\ \mathbb{C}v\{\mathbf{x}_k, \mathbf{x}_1\} & \dots & \mathbb{C}v\{\mathbf{x}_k, \mathbf{x}_j\} & \dots & \mathbb{V}\{\mathbf{x}_k\} \end{pmatrix}$$

In matrix notation, it can be computed as

$$\begin{aligned}\mathbb{C}v\{\mathbf{X}\} &= \frac{1}{n}\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} = \\ &= \frac{1}{n}(\mathbf{A}\mathbf{X})^\top\mathbf{A}\mathbf{X} = \\ &= \frac{1}{n}\mathbf{X}^\top\mathbf{A}^\top\mathbf{A}\mathbf{X}\end{aligned}$$

where $\mathbf{A}$ is the centering matrix (Equation 11.4). The variance-covariance matrix is:

1. **Squared**, i.e. $k \times k$, and **symmetric**.
2. **Semi-definite positive**.
3. Has the variances on the diagonal. Hence the trace (Equation 32.8) is $trace(\mathbb{C}v\{\mathbf{X}\}) = \sum_{j=1}^{k}\mathbb{V}\{\mathbf{x}_j\}$.

11.4 Stardardized variables

In order to remove the effect of the unit of measure in the different variables it is possible to work under the matrix of standardized variables, i.e.

$$\underset{n\times k}{\mathbf{Z}} = \begin{pmatrix} z_{1,1} & \cdots & z_{1,j} & \cdots & z_{1,k} \\ \vdots & & \vdots & & \vdots \\ z_{i,1} & \cdots & z_{i,j} & \cdots & z_{i,k} \\ \vdots & & \vdots & & \vdots \\ z_{n,1} & \cdots & z_{n,j} & \cdots & z_{n,k} \end{pmatrix} \tag{11.5}$$

where each element $z_{i,j}$ is defined as:

$$z_{i,j} = \frac{\tilde{x}_{i,j}}{\sqrt{\mathbb{V}\{x_j\}}} = \frac{x_{i,j} - \bar{x}_j}{\sqrt{\mathbb{V}\{x_j\}}}.$$

In matrix notation, $\mathbf{Z}$ can be rewritten as:

$$\mathbf{Z} = \tilde{\mathbf{X}} \cdot \mathbf{D}^{-\frac{1}{2}}$$

where the matrix $\mathbf{D}$ is defined as:

$$\underset{k\times k}{\mathbf{D}} = \mathbb{C}v\{\mathbf{X}\} \cdot \mathbf{I}_k = \begin{pmatrix} \mathbb{V}\{\mathbf{x}_1\} & \ldots & 0 & \ldots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \ldots & \mathbb{V}\{\mathbf{x}_j\} & \ldots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \ldots & 0 & \ldots & \mathbb{V}\{\mathbf{x}_k\} \end{pmatrix} \tag{11.6}$$

The standardized variables have mean equal to zero and unitary variance. Moreover, the numbers do not depend anymore from the unit of measure of the variables.

11.5 Correlations matrix

The correlation is a statistic that measure the linear dependence between two variables, in the simplest case the correlation between two vectors X_k and X_h is computed as:

$$\begin{aligned} \mathbb{C}r\{\mathbf{x}_k, \mathbf{x}_h\} &= \frac{\mathbb{C}v\{\mathbf{x}_k, \mathbf{x}_h\}}{\sqrt{\mathbb{V}\{\mathbf{x}_k\}\mathbb{V}\{\mathbf{x}_h\}}} = \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{i,k} - \bar{x}_k}{\sqrt{\mathbb{V}\{\mathbf{x}_k\}}}\right)\left(\frac{x_{i,h} - \bar{x}_h}{\sqrt{\mathbb{V}\{\mathbf{x}_h\}}}\right) = \\ &= \frac{1}{n}\sum_{i=1}^{n} z_{i,k}\, z_{i,h} = \frac{1}{n}\mathbf{z}_k^\top \mathbf{z}_h \end{aligned}$$

In matrix notation, $\mathbb{C}r\{\mathbf{X}\}$ can be rewritten as a $k \times k$ matrix:

$$\begin{aligned}\mathbb{C}r\{\mathbf{X}\} &= \mathbf{D}^{-\frac{1}{2}}\mathbb{C}v\{\mathbf{X}\}\,\mathbf{D}^{-\frac{1}{2}} = \\ &= \frac{1}{n}\mathbf{D}^{-\frac{1}{2}}\tilde{\mathbf{X}}^{\top}\tilde{\mathbf{X}}\,\mathbf{D}^{-\frac{1}{2}} = \\ &= \frac{1}{n}\mathbf{Z}^{\top}\mathbf{Z}\end{aligned}$$

where $\mathbf{D}$ is defined as in Equation 11.6. The correlation matrix is:

1. **Squared**, i.e. $k \times k$, and **symmetric**.
2. **Positive semi-definited** matrix.
3. **Trace** (Equation 32.8), i.e. $trace(\mathbb{C}r\{\mathbf{X}\}) = \sum_{j=1}^{k} 1 = k$.

The elements of the correlation matrix are:

$$\underset{k\times k}{\mathbb{C}r\{\mathbf{X}\}} = \begin{pmatrix} 1 & \dots & \mathbb{C}r\{\mathbf{x}_1, \mathbf{x}_j\} & \dots & \mathbb{C}r\{\mathbf{x}_1, \mathbf{x}_k\} \\ \vdots & & \vdots & & \vdots \\ \mathbb{C}r\{\mathbf{x}_j, \mathbf{x}_1\} & \dots & 1 & \dots & \mathbb{C}r\{\mathbf{x}_j, \mathbf{x}_k\} \\ \vdots & & \vdots & & \vdots \\ \mathbb{C}r\{\mathbf{x}_k, \mathbf{x}_1\} & \dots & \mathbb{C}r\{\mathbf{x}_k, \mathbf{x}_j\} & \dots & 1 \end{pmatrix}$$

> **!** Correlation matrix and standardized variables
>
> The correlation matrix can be seen as the variance-covariance matrix of the standardized variables $\mathbf{Z}$ (Section 11.4). In fact, a generic standardized j-variable has $\mathbb{V}\{\mathbf{z}_j\} = 1$ and $\mathbb{C}v\{\mathbf{z}_j, \mathbf{z}_k\} = \mathbb{C}r\{\mathbf{z}_j, \mathbf{z}_k\}$.

Part III

Statistical models

12 Statistical models

Statistical modeling applies statistical methods to real-world data to give empirical content to relationships. It aims to quantify phenomena and develop models and test hypotheses, making it a crucial field for economic research, policy analysis, and decision-making. The aim of the statistical modeling is to study the (unknown) mechanism that generates the data, i.e., the Data Generating Process (DGP). The **statistical model** is a function that approximates the DGP.

12.1 The matrix of data

Let's consider n realizations defining a sample for $i = 1, 2, \dots, n$. Suppose we have p dependent variables and k explanatory variables (also known as predictors). The data matrix for $\mathbf{X}$, the **exogenous** (regressors), is then composed as:

$$\underset{n\times k}{\mathbf{X}} = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,j} & \dots & x_{1,k} \\ x_{2,1} & x_{2,2} & \dots & x_{2,j} & \dots & x_{2,k} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,j} & \dots & x_{n,k} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_k^T \end{pmatrix}$$

where

- The **i**-th **row** contains the variables related to the i-th statistical unit (e.g., an individual, a firm, or a country).
- The **j**-th **column** contains all the observations related to the j-th variable.

The matrix $\mathbf{Y}$ represent the **endogenous** (dependent), i.e.

$$\underset{n\times p}{\mathbf{Y}} = \begin{pmatrix} y_{1,1} & y_{1,2} & \dots & y_{1,j} & \dots & y_{1,p} \\ y_{2,1} & y_{2,2} & \dots & y_{2,j} & \dots & y_{2,p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{n,1} & y_{n,2} & \dots & y_{n,j} & \dots & y_{n,p} \end{pmatrix} = \begin{pmatrix} \mathbf{y}_1^T \\ \mathbf{y}_2^T \\ \vdots \\ \mathbf{y}_p^T \end{pmatrix}$$

Hence, the complete matrix of data is given by:

$$\underset{n\times(k+p)}{\mathbf{W}} = (\mathbf{Y}\ \mathbf{X}) = \begin{pmatrix} y_{1,1} & \dots & y_{1,p} & x_{1,1} & \dots & x_{1,k} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ y_{n,1} & \dots & y_{n,p} & x_{n,1} & \dots & x_{n,k} \end{pmatrix} \tag{12.1}$$

In general, if $p = 1$ then the model has only one equation to satisfy for $i = 1, \dots, n$. For example, a linear model with one equation reads:

$$y_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \cdots + b_k x_{i,k} + e_i. \tag{12.2}$$

Otherwise, when $p > 1$ there are more than one dependent variable and the model is composed by p-equations for $i = 1, \dots, n$, i.e. the same linear model with p equations reads:

$$\begin{cases} y_{i,1} = b_{0,1} + b_{1,1} x_{i,1} + b_{1,2} x_{i,2} + \cdots + b_{1,k} x_{i,k} + e_{i,1} \\ y_{i,2} = b_{0,2} + b_{2,1} x_{i,1} + b_{2,2} x_{i,2} + \cdots + b_{2,k} x_{i,k} + e_{i,2} \\ \vdots \\ y_{i,p} = b_{0,p} + b_{p,1} x_{i,1} + b_{p,2} x_{i,2} + \cdots + b_{p,k} x_{i,k} + e_{i,p} \end{cases} \tag{12.3}$$

12.2 Joint, conditional and marginals

Let's consider the bi-dimensional random vector $\mathbf{W}$ in Equation 12.1 and let's write the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$, i.e.

$$\underbrace{\mathbb{P}(\mathbf{Y} \leq \mathbf{y}, \mathbf{X} \leq \mathbf{x})}_{\text{joint probability}} = \underbrace{F_{\mathbf{Y},\mathbf{X}}(\mathbf{y}, \mathbf{x})}_{\text{distribution function}} \tag{12.4}$$

In the continuous case, there exists a joint density $f_{\mathbf{Y},\mathbf{X}}(\mathbf{y}, \mathbf{x})$ such that:

$$F_{\mathbf{Y},\mathbf{X}}(\mathbf{y}, \mathbf{x}) = \int_{-\infty}^{\mathbf{x}} \int_{-\infty}^{\mathbf{y}} f_{\mathbf{Y},\mathbf{X}}(\mathbf{y}, \mathbf{x}) d\mathbf{y} d\mathbf{x} \tag{12.5}$$

Moreover, from the joint distribution (Equation 12.4) it is possible to recover the marginals distributions, i.e.

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \partial_{\mathbf{y}} F_{\mathbf{Y},\mathbf{X}}(\mathbf{y}, \mathbf{x}) = \int_{-\infty}^{\infty} f_{\mathbf{Y},\mathbf{X}}(\mathbf{y}, \mathbf{x}) d\mathbf{x} \\ f_{\mathbf{X}}(\mathbf{x}) &= \partial_{\mathbf{x}} F_{\mathbf{Y},\mathbf{X}}(\mathbf{y}, \mathbf{x}) = \int_{-\infty}^{\infty} f_{\mathbf{Y},\mathbf{X}}(\mathbf{y}, \mathbf{x}) d\mathbf{y} \end{aligned} \tag{12.6}$$

Given the marginals (Equation 12.6), it is possible to compute the unconditional moments, i.e.

1. **First moment**: $\mathbb{E}\{\mathbf{Y}\} = \int_{-\infty}^{\infty} \mathbf{y} f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y}$.
2. **Second moment**: $\mathbb{E}\{\mathbf{Y}^2\} = \int_{-\infty}^{\infty} \mathbf{y}^2 f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y}$.
3. **Variance**: $\mathbb{V}\{\mathbf{Y}\} = \mathbb{E}\{\mathbf{Y}^2\} - \mathbb{E}\{\mathbf{Y}\}^2$.

Using the Bayes theorem, from the joint distribution (Equation 12.4) it is possible to recover the conditional distribution, i.e

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \frac{f_{\mathbf{Y},\mathbf{X}}(\mathbf{y},\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} \tag{12.7}$$

Given the conditional distributions, it is possible to compute the conditional moments, i.e.

1. **First moment**: $\mathbb{E}\{\mathbf{Y}|\mathbf{X}\} = \int_{-\infty}^{\infty} \mathbf{y} f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) d\mathbf{y}$.
2. **Second moment**: $\mathbb{E}\{\mathbf{Y}^2|\mathbf{X}\} = \int_{-\infty}^{\infty} \mathbf{y}^2 f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) d\mathbf{y}$.

Hence, from Equation 12.7 the joint density can be represented as the product of the conditional and the marginal, i.e.

$$\underset{\text{joint}}{f_{\mathbf{Y},\mathbf{X}}(\mathbf{y},\mathbf{x})} = \underset{\text{conditional}}{f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})} \cdot \underset{\text{marginal}}{f_{\mathbf{X}}(\mathbf{x})} \tag{12.8}$$

Inference in a multivariate Gaussian model

Let's consider a Gaussian setup, i.e.

$$\mathbf{W}^\top = \begin{pmatrix} \mathbf{Y} \\ \mathbf{X} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_{\mathbf{Y}} \\ \mu_{\mathbf{X}} \end{pmatrix}, \begin{pmatrix} \Sigma_{\mathbf{YY}} & \Sigma_{\mathbf{YX}} \\ \mu_{\mathbf{XY}} & \Sigma_{\mathbf{XX}} \end{pmatrix} \right)$$

For a Gaussian setup if $(\mathbf{X}\ \mathbf{Y})$ are jointly normal, then the marginals are normal, i.e.

$$\mathbf{Y} \sim \mathcal{N}(\mu_{\mathbf{Y}}, \Sigma_{\mathbf{YY}}), \quad \mathbf{X} \sim \mathcal{N}(\mu_{\mathbf{X}}, \Sigma_{\mathbf{XX}}).$$

and also the conditionals distributions are normal, i.e.

$$\mathbf{Y}|\mathbf{X} \sim \mathcal{N}(\mu_{\mathbf{Y}|\mathbf{X}}, \Sigma_{\mathbf{YY}|\mathbf{X}}), \quad \mathbf{X}|\mathbf{Y} \sim \mathcal{N}(\mu_{\mathbf{X}|\mathbf{Y}}, \Sigma_{\mathbf{XX}|\mathbf{Y}}).$$

and the conditional moments reads explicitly as:

$$\begin{aligned} \mathbb{E}\{\mathbf{Y} \mid \mathbf{X}\} = \mu_{\mathbf{Y}|\mathbf{X}} = \\ &= \mu_{\mathbf{Y}} + \Sigma_{\mathbf{YX}} \cdot \Sigma_{\mathbf{XX}}^{-1}(\mathbf{X} - \mu_{\mathbf{X}}) = \\ &= \mu_{\mathbf{Y}} - \Sigma_{\mathbf{YX}} \cdot \Sigma_{\mathbf{XX}}^{-1} \mu_{\mathbf{X}} + \Sigma_{\mathbf{YX}} \cdot \Sigma_{\mathbf{XX}}^{-1} \mathbf{X} = \\ &= \mu_{\mathbf{Y}} - \mathbf{b}_{\mathbf{Y}|\mathbf{X}} \mu_{\mathbf{X}} + \mathbf{b}_{\mathbf{Y}|\mathbf{X}} \mathbf{X} = \\ &= \mathbf{a}_{\mathbf{Y}|\mathbf{X}} + \mathbf{b}_{\mathbf{Y}|\mathbf{X}} \mathbf{X} \end{aligned}$$

and

$$\mathbb{V}\{\mathbf{Y}|\mathbf{X}\} = \Sigma_{\mathbf{YY}} - \Sigma_{\mathbf{YX}} \cdot \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{YX}}$$

In this setup the parameters are:

- **Joint distribution**, $\theta = \{\mu_{\mathbf{Y}}, \mu_{\mathbf{X}}, \Sigma_{\mathbf{XX}}, \Sigma_{\mathbf{XY}}, \Sigma_{\mathbf{YY}}\}$.

- **Conditional distribution**, $\lambda_1 = \{\mathbf{a}_{\mathbf{Y}|\mathbf{X}}, \mathbf{b}_{\mathbf{Y}|\mathbf{X}}, \Sigma_{\mathbf{YY}|\mathbf{X}}\}$.
- **Marginal distribution**, $\lambda_2 = \{\mu_{\mathbf{X}}, \Sigma_{\mathbf{XX}}\}$.

Noting that λ_1 is a function of θ, i.e. $\tau = f(\lambda_1)$ in the Gaussian case it is possible to prove that λ_1 and λ_2 are free to vary. Hence, imposing restrictions on λ_1 do not impose restrictions on λ_2. In general, if the parameters of interest are a function of the conditional distribution and λ_1 and λ_2 are free to vary, then the inference can be done without losing of information considering the conditional model. In this case we say that $\mathbf{X}$ is **weakly exogenous** for $\tau = f(\lambda_1)$.

12.3 Conditional expectation model

Let's consider a very general conditional expectation model with $p = 1$, of which the linear models are a special case. In matrix notation it can be written as:

$$\mathbf{y} = \mathbb{E}\{\mathbf{y} \mid \mathbf{X}\} + \mathbf{e} \tag{12.9}$$

where the conditional expectation errors are defined as:

$$\mathbf{e} = \mathbf{y} - \mathbb{E}\{\mathbf{y} \mid \mathbf{X}\} \tag{12.10}$$

Then, in general the unconditional expectation of the residuals $\mathbf{e}$ and the covariance between the residuals and the regressors are zero, i.e.

$$\mathbb{E}\{\mathbf{e}\} = 0, \quad \mathbb{E}\{\mathbf{e}\mathbf{X}\} = 0.$$

Moreover, the conditional expectation error is orthogonal to any transformation of the conditioning variables. Consider a more general setup, i.e.

$$\mathbf{y} = \mathbb{E}\{\mathbf{y} \mid \mathbf{X}\} + \mathbf{e}, \quad \mathbb{E}\{\mathbf{y} \mid \mathbf{X}\} = g(\mathbf{X}) \tag{12.11}$$

we have that

$$\mathbb{E}\{\mathbf{e}\, g(\mathbf{X})\} = 0.$$

In a conditional expectation model the residuals and the regressors are uncorrelated

Proof. Let's start the unconditional expectation of the residuals defined in Equation 12.9,

i.e.

$$\begin{aligned}\mathbb{E}\{\mathbf{e}\} &= \mathbb{E}\{\mathbf{y} - \mathbb{E}\{\mathbf{y} \mid \mathbf{X}\}\} = \\ &= \mathbb{E}\{\mathbf{y}\} - \mathbb{E}\{\mathbb{E}\{\mathbf{y} \mid \mathbf{X}\}\} = \\ &= \mathbb{E}\{\mathbf{y}\} - \mathbb{E}\{\mathbf{y}\} = 0\end{aligned}$$

Then, let's compute the expected value of between the residuals and the regressors, i.e.

$$\mathbb{E}\{\mathbf{e}\} = 0 \implies \mathbb{C}v\{\mathbf{e}, \mathbf{X}\} = \mathbb{E}\{\mathbf{e}\mathbf{X}\}$$

For simplicity let's assume that $\mathbf{X}$ can takes only values in $\{0, 1\}$. Applying the tower property of conditional expectation one obtain:

$$\begin{aligned}\mathbb{E}\{\mathbf{e}\mathbf{X}\} &= \mathbb{E}\{\mathbb{E}\{\mathbf{e}\mathbf{X} \mid \mathbf{X}\}\} = \\ &= \mathbb{E}\{\mathbf{e}\mathbf{X} \mid \mathbf{X} = 0\}\mathbb{P}(\mathbf{X} = 0) + \mathbb{E}\{\mathbf{e}\mathbf{X} \mid \mathbf{X} = 1\}\mathbb{P}(\mathbf{X} = 1) = \\ &= \mathbb{E}\{\mathbf{e}\mathbf{X} \mid \mathbf{X} = 1\}\mathbb{P}(\mathbf{X} = 1)\end{aligned}$$

Then, let's substitute $\mathbf{e}$ from Equation 12.9 and $\mathbf{X}$ with 1, i.e.

$$\begin{aligned}\mathbb{E}\{\mathbf{e}\mathbf{X}\} &= \mathbb{E}\{(\mathbf{y} - \mathbb{E}\{\mathbf{y}|\mathbf{X}\})\mathbf{X} \mid \mathbf{X} = 1\}\mathbb{P}(\mathbf{X} = 1) = \\ &= \mathbb{E}\{\mathbf{y} \mid \mathbf{X} = 1\}\mathbb{P}(\mathbf{X} = 1) - \mathbb{E}\{\mathbb{E}\{\mathbf{y}|\mathbf{X}\} \mid \mathbf{X} = 1\}\mathbb{P}(\mathbf{X} = 1) = \\ &= \mathbb{E}\{\mathbf{y} \mid \mathbf{X} = 1\}\mathbb{P}(\mathbf{X} = 1) - \mathbb{E}\{\mathbf{y} \mid \mathbf{X} = 1\}\mathbb{P}(\mathbf{X} = 1) = 0\end{aligned}$$

For a general transformation of the regressors as in Equation 12.11, the covariance is computed as:

$$\begin{aligned}\mathbb{E}\{\mathbf{e}g(\mathbf{X})\} &= \mathbb{E}\{\mathbb{E}\{\mathbf{e}g(\mathbf{X}) \mid \mathbf{X}\}\} = \\ &= \mathbb{E}\{g(\mathbf{X})\,\mathbb{E}\{\mathbf{e} \mid \mathbf{X}\}\} = \\ &= \mathbb{E}\{g(\mathbf{X})\,\mathbb{E}\{\mathbf{y} - \mathbb{E}\{\mathbf{y}|\mathbf{X}\} \mid \mathbf{X}\}\} = \\ &= \mathbb{E}\{g(\mathbf{X})\,[\mathbb{E}\{\mathbf{y} \mid \mathbf{X}\} - \mathbb{E}\{\mathbf{y} \mid \mathbf{X}\}]\} = 0\end{aligned}$$

□

13 Introduction to linear models

Let's consider an uni-equational linear model, i.e. with $p = 1$ in (Equation 12.2), is expressed in compact matrix notation as:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}, \tag{13.1}$$

where $\mathbf{b}$ and $\mathbf{e}$ represent the true parameters and residuals in population. Let's consider a sample of n-observations extracted from a population, then the matrix of the regressors $\mathbf{X}$ reads

$$\underset{n\times k}{\mathbf{X}} = \begin{pmatrix} x_{1,1} & \dots & x_{1,k} \\ \vdots & & \vdots \\ x_{n,1} & \dots & x_{n,k} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_k^\top \end{pmatrix},$$

while the vectors of dependent variable and of the residuals reads

$$\underset{n\times 1}{\mathbf{y}} = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix}, \quad \underset{n\times 1}{\mathbf{e}} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

Hence, the matrix of data is composed by:

$$\underset{n\times(k+1)}{\mathbf{W}} = (\mathbf{y}\ \mathbf{X}) = \begin{pmatrix} y_1 & x_{1,1} & \dots & x_{1,k} \\ \vdots & \vdots & \ddots & \vdots \\ y_n & x_{n,1} & \dots & x_{n,k} \end{pmatrix} \tag{13.2}$$

Depending on the assumptions made on the variance of the residuals the linear models can be distinguished in 3 classes as shown in Figure 13.1.

For a generalized linear model the variance-covariance matrix of the residuals in matrix notation is written as:

$$\underset{n\times n}{\Sigma} = \mathbb{V}\{\mathbf{e}\mathbf{e}^\top|\mathbf{X}\} = \mathbb{E}\{\mathbf{e}\mathbf{e}^\top|\mathbf{X}\}$$

where the $n \times n$ elements are

$$\underset{n\times n}{\Sigma} = \begin{pmatrix} e_1^2 & e_1e_2 & \dots & e_1e_n \\ e_2e_1 & e_2^2 & \dots & e_2e_n \\ \vdots & \vdots & & \vdots \\ e_ne_1 & e_ne_2 & \dots & e_n^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \dots & \sigma_{1,n} \\ \sigma_{2,1} & \sigma_2^2 & \dots & \sigma_{2,n} \\ \vdots & \vdots & & \vdots \\ \sigma_{n,1} & \sigma_{n,2} & \dots & \sigma_n^2 \end{pmatrix} \tag{13.3}$$

Since the matrix Σ is symmetric the number of unique values (free elements) are given by n variances and $\frac{n(n-1)}{2}$ covariances. Hence, the total number of free elements is given by:

$$n + \frac{n(n-1)}{2} = \frac{2n + n^2 - n}{2} = \frac{n + n^2}{2} = \frac{n(n+1)}{2}$$

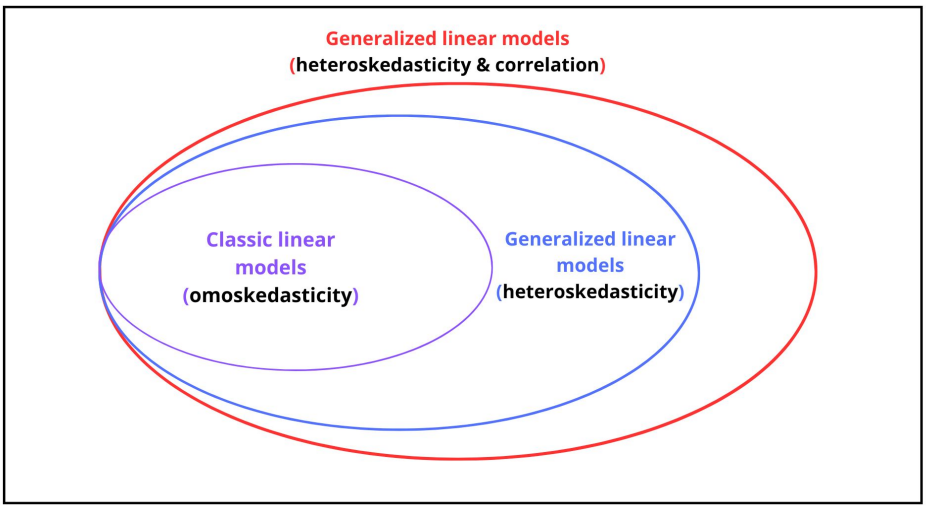


Figure 13.1: Different classes of linear models.

13.0.1 Estimators of b

Let's denote with $\Theta_{\mathbf{b}}$ the parameter space, i.e. $\Theta_{\mathbf{b}} \subset \mathbb{R}^k$, and with Q an estimator function of the unknown true parameter $\mathbf{b} \in \Theta_{\mathbf{b}}$. Then, the function Q defines an **estimator** of $\mathbf{b}$, meaning it is a function that takes the matrix of data as input and returns a vector of parameters within $\Theta_{\mathbf{b}}$ as output:

$$Q : \mathbf{W} \longrightarrow \Theta_{\mathbf{b}}, \quad \text{such that} \quad Q(\mathbf{W}) = \hat{\mathbf{b}} \in \Theta_{\mathbf{b}}$$

where $\hat{\mathbf{b}}$ is an **estimates** of the true population's parameter $\mathbf{b}$. Then, the fitted values $\hat{\mathbf{y}}$ are a function of the estimate and are defined as:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}} \tag{13.4}$$

Consequently, the fitted residuals, which measure the discrepancies between the observed and the fitted values, are also a function of $\hat{\mathbf{b}}$, i.e.

$$\mathbf{e}(\hat{\mathbf{b}}) = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} \tag{13.5}$$

13.1 Variance decomposition

In a linear model, the deviance (or total variance) of the dependent variable $\mathbf{y}$ can be decomposed into the sum of the regression variance and the dispersion variance. This decomposition helps us understand how much of the total variability in the data is explained by the model and how much is due to unexplained variability (residuals).

- **Total Deviance** ($\mathbb{D}ev\{\mathbf{y}\}$): represents the total variability of the dependent variable $\mathbf{y}$. It is calculated as the sum of the squared difference of y_i from its mean $\bar{y}$.
- **Regression Deviance** ($\mathbb{D}ev\mathbb{R}eg\{\mathbf{y}\}$): represents the portion of variability that is explained by the regression model. It is computed as the sum of the squared differences between the fitted values $\hat{y}_i$ and $\bar{y}$.
- **Dispersion Deviance** ($\mathbb{D}ev\mathbb{D}isp\{\mathbf{y}\}$): represents the portion of variability that is not explained by the model. It is computed as the sum of the squared differences between the observed values y_i and the fitted values $\hat{y}_i$ (Equation 13.4).

Hence, the total deviance of $\mathbf{y}$ can be decomposed as follows:

$$\begin{aligned} \mathbb{D}ev\{\mathbf{y}\} &= \mathbb{D}ev\mathbb{R}eg\{\mathbf{y}\} + \mathbb{D}ev\mathbb{D}isp\{\mathbf{y}\} \\ \sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \sum_{i=1}^{n}(\hat{y}_i-y_i)^2 \\ \mathbf{y}^\top\mathbf{y}-n\bar{y}^2 &= \mathbf{b}^\top\mathbf{X}^\top\mathbf{X}\mathbf{b}-n\bar{y}^2 + \mathbf{e}^\top\mathbf{e} \end{aligned} \tag{13.6}$$

Regression deviance

Proof. Let's prove the expression for the regression deviance $\mathbb{D}ev\mathbb{R}eg\{\mathbf{y}\}$, i.e.

$$\begin{aligned} \mathbb{D}ev\mathbb{R}eg\{\mathbf{y}\} &= \mathbb{D}ev\{\mathbf{y}\} - \mathbb{D}ev\mathbb{D}isp\{\mathbf{y}\} = \\ &= \mathbf{y}^\top\mathbf{y} - n\bar{y}^2 - \mathbf{e}^\top\mathbf{e} = \\ &= \mathbf{y}^\top\mathbf{y} - n\bar{y}^2 - (\mathbf{y}-\mathbf{X}\mathbf{b})^\top(\mathbf{y}-\mathbf{X}\mathbf{b}) = \\ &= \mathbf{y}^\top\mathbf{y} - n\bar{y}^2 + \mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{X}\mathbf{b} - \mathbf{y}\mathbf{b}^\top\mathbf{X}^\top + \mathbf{b}^\top\mathbf{X}^\top\mathbf{X}\mathbf{b} = \\ &= 2\mathbf{y}^\top\mathbf{y} - n\bar{y}^2 - 2\mathbf{y}^\top(\mathbf{X}\mathbf{b}) + \mathbf{b}^\top\mathbf{X}^\top\mathbf{X}\mathbf{b} = \\ &= \mathbf{b}^\top\mathbf{X}^\top\mathbf{X}\mathbf{b} - n\bar{y}^2 \end{aligned}$$

□

The decomposition of the deviance of $\mathbf{y}$ holds true also with respect to the correspondents degrees of freedom,

Table 13.1: Deviance and variance decomposition in a multivariate linear model

Deviance	**Degrees of freedom**	**Variance**
$\mathbb{D}ev\{\mathbf{y}\} = \sum_{i=1}^{n}(y_i-\bar{y})^2$	$n-1$	$\hat{s}_y^2 = \frac{\mathbb{D}ev\{\mathbf{y}\}}{n-1}$
$\mathbb{D}ev\mathbb{R}eg\{\mathbf{y}\} = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	k	$\hat{s}_r^2 = \frac{\mathbb{D}ev\mathbb{R}eg\{\mathbf{y}\}}{k}$
$\mathbb{D}ev\mathbb{D}isp\{\mathbf{y}\} = \sum_{i=1}^{n}(\hat{y}_i-y_i)^2$	$n-k-1$	$\hat{s}_e^2 = \frac{\mathbb{D}ev\mathbb{D}isp\{\mathbf{y}\}}{n-k-1}$

13.2 Multivariate R Squared

The R^2 statistic, also known as the coefficient of determination, is a measure used to assess the goodness of fit of a regression model. In a multivariate context, it evaluates how well the independent variables explain the variability of the dependent variable.

Definition 13.1. (Multivariate R^2)
The R^2 is defined as the ratio of the deviance explained by the model ($\mathbb{D}ev\mathbb{R}eg\{\mathbf{y}\}$) to the total deviance ($\mathbb{D}ev\{\mathbf{y}\}$). It can also be expressed as one minus the ratio of the residual deviance ($\mathbb{D}ev\mathbb{D}isp\{\mathbf{y}\}$) to the total deviance, i.e.

$$R^2 = \frac{\mathbb{D}ev\mathbb{R}eg\{\mathbf{y}\}}{\mathbb{D}ev\{\mathbf{y}\}} = 1 - \frac{\mathbb{D}ev\mathbb{D}isp\{\mathbf{y}\}}{\mathbb{D}ev\{\mathbf{y}\}} \tag{13.7}$$

Using the variance decomposition (Equation 13.6), it is possible to write the R^2 as:

$$R^2 = \frac{\mathbf{b}^\top\mathbf{X}^\top\mathbf{X}\mathbf{b} - n\bar{y}^2}{\mathbf{y}^\top\mathbf{y} - n\bar{y}^2} = 1 - \frac{\mathbf{e}^\top\mathbf{e}}{\mathbf{y}^\top\mathbf{y} - n\bar{y}^2}$$

The numerator represents the variance explained by the regression model, while the denominator the total variance in the dependent variable. The term $\mathbf{e}^\top\mathbf{e}$ in the second expression represents the variance of the residuals, or the variance not explained by the model. An R^2 value close to 1 indicates that a large proportion of the variability in the dependent variable has been accounted for by the regression model, while a value close to 0 indicates that the model explains very little of the variability.

> ⚠ Limitations of R^2
>
> The R^2 metric has some limitations. Firstly, it can be close to 1 even if the relationship between the variables is not linear. Additionally, R^2 increases whenever a new regressor is added to the model, making it unsuitable for comparing models with different numbers of regressors.

A more robust indicator that does not always increase with the addition of a new regressor is the **adjusted** R^2, which is computed as:

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{\mathbb{D}ev\mathbb{D}isp\{\mathbf{y}\}}{\mathbb{D}ev\{\mathbf{y}\}} = 1 - \frac{\hat{s}_e^2}{\hat{s}_y^2}$$

The adjusted R^2 can be negative, and its value will always be less than or equal to that of R^2. Unlike R^2, the adjusted R^2 increases only when the new explanatory variable improves the model more than would be expected simply by adding another variable.

Adjusted R^2

Proof. To arrive at the formulation of the adjusted R^2 let's consider that under the null hypothesis $H_0 : b_1 = b_2 = \cdots = b_k$ the variance of regression $\hat{s}_r^2$ (Table 13.1) is a correct estimate of the variance of the residuals σ_e^2. Hence, under H_0:

$$\frac{n-1}{k}\mathbb{E}\left\{\frac{\mathbb{D}ev\mathbb{R}eg\{\mathbf{y}\}}{\mathbb{D}ev\{\mathbf{y}\}}\right\} \tilde{=} 1$$

This implies that the expectation of the R^2 is not zero (as it should be under H_0) but:

$$\mathbb{E}\{R^2\} \tilde{=} \frac{k}{n-1}$$

Let's rescale the R^2 such that when H_0 holds true it is equal to zero, i.e.

$$R_c^2 = R^2 - \frac{k}{n-1}$$

However, the specification of R_c^2 implies that when $R^2 = 1$ (perfect linear relation between $\mathbf{X}$ and $\mathbf{y}$) the value of $R_c^2 < 1$, i.e. $R_c^2 = \frac{n-k-1}{n-1} < 1$. Hence, let's correct again the indicator such that it takes values in $[0, 1]$, i.e.

$$\begin{aligned}
\bar{R}^2 &= \left(R^2 - \frac{k}{n-1}\right)\frac{n-1}{n-k-1} = \\
&= \left(\frac{R^2(n-1)-k}{n-1}\right)\frac{n-1}{n-k-1} = \\
&= \frac{n-1}{n-k-1}R^2 - \frac{k}{n-k-1}
\end{aligned}$$

Remembering that R^2 can be rewritten as in Equation 13.7 one obtain:

$$\begin{aligned}
\bar{R}^2 &= \frac{n-1}{n-k-1}\left(1 - \frac{\mathbb{D}ev\mathbb{D}isp\{\mathbf{y}\}}{\mathbb{D}ev\{\mathbf{y}\}}\right) - \frac{k}{n-k-1} = \\
&= \frac{(n-1)\mathbb{D}ev\{\mathbf{y}\} - (n-1)\mathbb{D}ev\mathbb{D}isp\{\mathbf{y}\}}{\mathbb{D}ev\{\mathbf{y}\}(n-k-1)} - \frac{k}{n-k-1} = \\
&= \frac{n-1}{n-k-1} - \frac{n-1}{n-k-1}\frac{\mathbb{D}ev\mathbb{D}isp\{\mathbf{y}\}}{\mathbb{D}ev\{\mathbf{y}\}} - \frac{k}{n-k-1} = \\
&= 1 - \frac{n-1}{n-k-1}\frac{\mathbb{D}ev\mathbb{D}isp\{\mathbf{y}\}}{\mathbb{D}ev\{\mathbf{y}\}} = \\
&= 1 - \frac{\hat{s}_e^2}{\hat{s}_y^2}
\end{aligned}$$

□

14 Classic linear models

14.1 Working hypothesis

Let's start from the *classic* assumptions for a linear model, i.e. the Gauss-Markov ones. The working hypothesis for such kind of models are:

1. $\mathbb{E}\{y_i|\mathbf{x}_1, \dots, \mathbf{x}_n\} = \mathbb{E}\{y_i|\mathbf{X}\} = \mathbf{x}_i^\top \mathbf{b}$ for $i = 1, \dots n$.
2. $\mathbb{V}\{y_i|\mathbf{x}_1, \dots, \mathbf{x}_n\} = \mathbb{V}\{y_i|\mathbf{X}\} = \sigma_e^2$ with $0 < \sigma_e^2 < \infty$.
3. $\mathbb{C}v\{y_i, y_j|\mathbf{x}_1, \dots, \mathbf{x}_n\} = \mathbb{C}v\{y_i, y_j|\mathbf{X}\} = 0$ with $i \neq j$ and $i, j \in \{1, \dots, n\}$.

Equivalently the formulation in terms of the stochastic component reads

1. $y_i = \mathbf{x}_i^\top \mathbf{b} + e_i$ for $i = 1, \dots n$.
2. $\mathbb{E}\{e_i|\mathbf{x}_1, \dots, \mathbf{x}_n\} = \mathbb{E}\{e_i|\mathbf{X}\} = \mathbf{0}$.
3. $\mathbb{V}\{e_i|\mathbf{x}_1, \dots, \mathbf{x}_n\} = \mathbb{V}\{e_i|\mathbf{X}\} = \sigma_e^2$ with $0 < \sigma_e^2 < \infty$.
4. $\mathbb{C}v\{e_i, e_j|\mathbf{x}_1, \dots, \mathbf{x}_n\} = \mathbb{C}v\{e_i, e_j|\mathbf{X}\} = 0$ with $i \neq j$ and $i, j \in \{1, \dots, n\}$.

Hence, the error terms is assumed to be IID with constant variance and the variance covariance matrix in Equation 13.3 reduces to $\Sigma = \sigma_e^2 \mathbf{I}_n$.

14.2 Ordinary least squares (OLS)

Proposition 14.1. *(**Ordinary Least Squares (OLS)**)*
The ordinary least squares estimator (OLS) is the function Q that minimize the sum of the squared residuals and return an estimate $\mathbf{b}^{OLS}$ of the true parameter $\mathbf{b}$. The OLS optimization problem reads:

$$\underset{\mathbf{b}^{OLS} \in \Theta_{\mathbf{b}}}{argmin}\ Q(\mathbf{b}^{OLS}) = \underset{\mathbf{b}^{OLS} \in \Theta_{\mathbf{b}}}{argmin}\ \{\mathbf{e}(\mathbf{b}^{OLS})^\top \mathbf{e}(\mathbf{b}^{OLS})\} \tag{14.1}$$

*Notably, if $\mathbf{X}$ is **non-singular** it is possible to recover an analytic solution, i.e.*

$$\mathbf{b}^{OLS} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \tag{14.2}$$

Equivalently, it is possible to express the solution in terms of the covariance matrix of the $\mathbf{X}$ and between the $\mathbf{X}$ and the $\mathbf{y}$, i.e.

$$\mathbf{b}^{OLS} = \mathbb{C}v\{\mathbf{X}\}^{-1} \mathbb{C}v\{\mathbf{X}, \mathbf{Y}\} \tag{14.3}$$

Singularity of $\mathbf{X}$

Note that the solution is available if and only if $\mathbf{X}$ is **non-singular**. Hence, the columns should not be linearly dependent. In fact, one of the k-variables can be written as a linear combination of the others, then the determinant of the matrix $\mathbf{X}^\top\mathbf{X}$ is zero and the inversion is not possible. Moreover, to have that $rank(\mathbf{X}^\top\mathbf{X}) = k$ it is necessary that the number of observations have to be greater or equal than the number of regressors, i.e. $n \geq k$.

Ordinary Least Square (OLS)

Proof. Let's prove the optimal solution in Equation 14.2. Developing the optimization problem in Equation 14.1:

$$\begin{aligned} Q(\mathbf{b}^{\text{OLS}}) &= \mathbf{e}(\mathbf{b}^{\text{OLS}})^\top\mathbf{e}(\mathbf{b}^{\text{OLS}}) = \\ &= (\mathbf{y} - \mathbf{X}\mathbf{b}^{\text{OLS}})^\top(\mathbf{y} - \mathbf{X}\mathbf{b}^{\text{OLS}}) = \\ &= \mathbf{y}^\top\mathbf{y} - (\mathbf{b}^{\text{OLS}})^\top\mathbf{X}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{X}\mathbf{b}^{\text{OLS}} + (\mathbf{b}^{\text{OLS}})^\top\mathbf{X}^\top\mathbf{X}\mathbf{b}^{\text{OLS}} = \\ &= \mathbf{y}^\top\mathbf{y} - 2(\mathbf{b}^{\text{OLS}})^\top\mathbf{X}^\top\mathbf{y} + (\mathbf{b}^{\text{OLS}})^\top\mathbf{X}^\top\mathbf{X}\mathbf{b}^{\text{OLS}} \end{aligned} \tag{14.4}$$

In order to find the minimum, let's compute the derivative with respect to $\mathbf{b}^{\text{OLS}}$ of Q and setting it equal to zero, i.e.

$$\begin{aligned} &\frac{d\mathrm{Q}(\mathbf{b}^{\text{OLS}})}{d\mathbf{b}^{\text{OLS}}} = -2\mathbf{X}^\top\mathbf{y} + 2\mathbf{X}^\top\mathbf{X}\mathbf{b}^{\text{OLS}} = \mathbf{0} \\ &\implies \mathbf{X}^\top\mathbf{y} = \mathbf{X}^\top\mathbf{X}\mathbf{b}^{\text{OLS}} \\ &\implies \mathbf{b}^{\text{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \end{aligned}$$

The second derivatives is **positive**, hence the solution corresponds to a minimum for the function Q, i.e.

$$\frac{d^2\mathrm{Q}(\mathbf{b}^{\text{OLS}})}{d\mathbf{b}^{\text{OLS}}d(\mathbf{b}^{\text{OLS}})^\top} = 2\mathbf{X}^\top\mathbf{X} > 0$$

Let's now consider the alternative expression in Equation 14.3. Considering the same optimization problem, let's denote as:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top = \mathbb{C}v\{\mathbf{X}\} \quad \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i y_i^\top = \mathbb{C}v\{\mathbf{X}, \mathbf{Y}\}$$

Then, substituting such values in Equation 14.2 it is straightforward to prove that $\mathbf{b}^{\text{OLS}}$ can be written as in Equation 14.3. □

⚠ Intercept estimate

If in the data matrix $\mathbf{X}$ was included a column with ones, then the intercept parameter is obtained from Equation 14.2 or Equation 14.3. However, if it was not included, it is computed as:

$$\alpha^{\text{OLS}} = \mathbb{E}\{\mathbf{Y}\} - \mathbf{b}^{\text{OLS}}\mathbb{E}\{\mathbf{X}\}$$

14.2.1 Projection matrices

Substituting the OLS solution (Equation 14.2) in Equation 13.1 we obtain the matrix $\mathbf{H}$, that project the vector $\mathbf{y}$ on the sub space of $\mathbb{R}^n$ generated by the matrix of the regressors $\mathbf{X}$, i.e.

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \tag{14.5}$$

As properties we have that:

1. $\mathbf{H}$ is an $n \times n$ **symmetric** matrix.
2. $\mathbf{H}\ \mathbf{H} = \mathbf{H}$ is **idempotent**.
3. $\mathbf{H}\ \mathbf{X} = \mathbf{X}$.

Instead, substituting the OLS solution (Equation 14.2) in the residuals (Equation 13.5) we obtain the projection matrix $\mathbf{M}$ that project the vector $\mathbf{y}$ on the orthogonal sub-space with respect to the sub-space generated by the matrix of the regressors $\mathbf{X}$, i.e.

$$\mathbf{M} = \mathbf{I}_n - \mathbf{H} \tag{14.6}$$

As properties we have that:

1. $\mathbf{M}$ is an $n \times n$ **symmetric** matrix.
2. $\mathbf{M}\ \mathbf{M} = \mathbf{H}$ is **idempotent**.
3. $\mathbf{M}\ \mathbf{X} = \mathbf{0}$.

By definition $\mathbf{M}$ and $\mathbf{H}$ are orthogonal, i.e. $\mathbf{H}\ \mathbf{M} = \mathbf{0}$. Hence, the fitted values defined as $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ are the projection of the empiric values on the sub-space generated by $\mathbf{X}$. Symmetrically, the fitted residuals $\hat{\mathbf{e}} = \mathbf{M}\mathbf{y}$ are the projection of the empiric values on the sub-space orthogonal to the sub-space generated by $\mathbf{X}$.

Projection matrices

Proof. Let's consider the property 2 of $\mathbf{H}$, i.e.

$$\begin{aligned}\mathbf{H}\,\mathbf{H} &= (\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\,(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top) = \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \\ &= \mathbf{H}\end{aligned}$$

Let's consider the property 3 of $\mathbf{H}$, i.e.

$$\mathbf{H}\,\mathbf{X} = (\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\,\mathbf{X} = \mathbf{X}$$

Let's consider the property 2 of $\mathbf{M}$, i.e.

$$\begin{aligned}\mathbf{M}\,\mathbf{M} &= (\mathbf{I}_n - \mathbf{H})\,(\mathbf{I}_n - \mathbf{H}) = \\ &= \mathbf{I}_n - \mathbf{H} = \\ &= \mathbf{M}\end{aligned}$$

Let's consider the property 3 of $\mathbf{M}$, i.e.

$$\begin{aligned}\mathbf{M}\,\mathbf{X} &= (\mathbf{I}_n - \mathbf{H})\mathbf{X} = \\ &= (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\,\mathbf{X} = \\ &= \mathbf{X} - \mathbf{X} = \mathbf{0}\end{aligned}$$

Finally, let's prove the orthogonality between $\mathbf{M}$ and $\mathbf{H}$, i.e.

$$\mathbf{H}\,\mathbf{M} = \mathbf{H}\,(\mathbf{I}_n - \mathbf{H}) = \mathbf{H} - \mathbf{H} = \mathbf{0}$$

□

14.3 Properties OLS

Theorem 14.1. *(Gauss-Markov theorem)*
Under the Gauss-Markov hypothesis the Ordinary Least Square (OLS) estimate is ***BLUE*** *(**Best Linear Unbiased Estimator**), where "*best*" stands for the estimator with* minimum variance *in the class of linear unbiased estimators of* $\mathbf{b}$. *The Gauss-Markov hypothesis are:*

1. $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$.
2. $\mathbb{E}\{\mathbf{e}\} = 0$.
3. $\mathbb{E}\{\mathbf{e}\mathbf{e}^\top\} = \sigma_e^2 \boldsymbol{I}_n$, *i.e.* ***omoskedasticity***.
4. $\mathbf{X}$ *is non-stochastic and independent from the errors for all* n*'s.*

Proposition 14.2. *(Properties OLS estimator)*
1. ***Unbiased****:* $\mathbf{b}^{OLS}$ *is correct and it's conditional expectation is equal to true parameter in population, i.e.*

$$\mathbb{E}\{\mathbf{b}^{OLS} \mid \mathbf{X}\} = \mathbf{b} \tag{14.7}$$

2. ***Linear*** *in the sense that it can be written as a linear combination of* $\mathbf{y}$ *and* $\mathbf{X}$*, i.e.* $\mathbf{b}^{OLS} = \mathbf{A}_x\mathbf{y}$*, where* $\mathbf{A}_x$ *do not depend on* $\mathbf{y}$*, i.e.*

$$\mathbf{b}^{OLS} = \mathbf{A}_x\mathbf{y}, \quad \mathbf{A}_x = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \tag{14.8}$$

3. Under the Gauss-Markov hypothesis (Theorem 14.1) it has ***minimum variance*** *in the class of the unbiased linear estimators and it reads:*

$$\mathbb{V}\{\mathbf{b}^{OLS} \mid \mathbf{X}\} = \sigma_e^2(\mathbf{X}^\top\mathbf{X})^{-1} \tag{14.9}$$

⚠ Variance Inflation Factor (VIF)

The elements on the diagonal of the matrix $(\mathbf{X}^\top\mathbf{X})^{-1}$ determine the variances while the other elements the covariances. In general the variance of the j-th regressor is denoted as $\mathbb{V}\{b_j\} = \sigma_e^2 c_{jj}$ where c_{jj} is the j-th element on the diagonal of $(\mathbf{X}^\top\mathbf{X})^{-1}$. An alternative expression for the variance is:

$$\mathbb{V}\{b_j\} = \frac{\sigma_e^2}{\mathbb{D}ev\{\mathbf{X}_j\}}\frac{1}{1-R_{j0}^2}$$

where R_{j0}^2 is the multivariate coefficient of determination on the regression of $\mathbf{X_j}$ on the other regressors. The term $\frac{1}{1-R_{j0}^2}$ is also denoted as VIF_j standing for **Variance Inflation Factor**.

ℹ Properties of the OLS estimator

Proof.

- The OLS estimator is **correct**: it's expected value is computed from Equation 14.2 and substituting Equation 13.1, is equal to the true parameter in population, i.e.

$$\begin{aligned}\mathbb{E}\{\mathbf{b}^{\text{OLS}} \mid \mathbf{X}\} &= \mathbb{E}\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \mid \mathbf{X}\} = \\ &= \mathbb{E}\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\mathbf{b}+\mathbf{e}) \mid \mathbf{X}\} = \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\mathbf{b} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbb{E}\{\mathbf{e} \mid \mathbf{X}\} = \\ &= \mathbf{b}\end{aligned}$$

- In general, applying the properties of the variance operator, the **variance** of $\mathbf{b}^{\text{OLS}}$

is computed as:

$$\begin{aligned}\mathbb{V}\{\mathbf{b}^{\text{OLS}} \mid \mathbf{X}\} &= \mathbb{V}\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \mid \mathbf{X}\} = \\ &= \mathbb{V}\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\mathbf{b}+\mathbf{e}) \mid \mathbf{X}\} = \\ &= \mathbb{V}\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\mathbf{b} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{e} \mid \mathbf{X}\} = \\ &= \mathbb{V}\{\mathbf{b} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{e} \mid \mathbf{X}\} = \\ &= \mathbb{V}\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{e} \mid \mathbf{X}\}\end{aligned}$$

Then, since $\mathbf{X}$ is non-stochastic it is possible to take it outside the variance squaring it and obtaining:

$$\begin{aligned}\mathbb{V}\{\mathbf{b}^{\text{OLS}} \mid \mathbf{X}\} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbb{V}\{\mathbf{e} \mid \mathbf{X}\}\,\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} = \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbb{E}\{\mathbf{e}\mathbf{e}^\top \mid \mathbf{X}\}\,\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\end{aligned} \tag{14.10}$$

Under the Gauss Markov hypothesis (Theorem 14.1) the conditional variance $\mathbb{V}\{\mathbf{e} \mid \mathbf{X}\} = \sigma \cdot \mathbf{I}_n$ and therefore the Equation 14.10 reduces to:

$$\begin{aligned}\mathbb{V}\{\mathbf{b}^{\text{OLS}} \mid \mathbf{X}\} &= \sigma_e^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} = \\ &= \sigma_e^2(\mathbf{X}^\top\mathbf{X})^{-1}\end{aligned}$$

□

14.4 Estimator of σ_e^2

The OLS estimator do not depend on σ_e^2 and it is not possible to obtain in one step both the estimators. As far as we know σ_e^2 is the variance of the residuals of which we know the realized values on the sample $\hat{\mathbf{e}} = \{\hat{e}_1, \hat{e}_2, \dots, \hat{e}_n\}$. Hence, let's define a correct estimator of the population variance σ_e^2 as:

$$\hat{s}_e^2 = \frac{\hat{\mathbf{e}}^\top\,\hat{\mathbf{e}}}{n-k-1}$$

Instead, in general the regression variance overestimate the true variance σ_e^2, i.e.

$$\hat{s}_e^2 = \sigma_e^2 + g(\mathbf{b}, \mathbf{X}, k), \quad g(\mathbf{b}, \mathbf{X}, k) \geq 0$$

Only in the special case where $b_1 = b_2 = \cdots = b_k$ in population, then $g(\mathbf{b}, \mathbf{X}, k) = 0$ and also the regression variance produces a correct estimate of σ_e^2.

i Correct estimator of σ_e^2

Proof. By definition, the residuals can be computed pre multiplying the matrix $\mathbf{M}$ to $\mathbf{y}$,

i.e.

$$\begin{aligned}\hat{\mathbf{e}} &= \mathbf{y} - \hat{\mathbf{y}} = \\ &= \mathbf{y} - \mathbf{X}\mathbf{b}^{\text{OLS}} = \\ &= \mathbf{y} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \\ &= (\mathbf{I}_n - \mathbf{H})\mathbf{y} = \\ &= \mathbf{M}\mathbf{y}\end{aligned}$$

Substituting $\mathbf{y} = \mathbf{X}\mathbf{b}^{\text{OLS}} + \mathbf{e}$:

$$\begin{aligned}\hat{\mathbf{e}} &= \mathbf{M}(\mathbf{X}\mathbf{b}^{\text{OLS}} + \mathbf{e}) = \\ &= \mathbf{M}\mathbf{X}\mathbf{b}^{\text{OLS}} + \mathbf{M}\mathbf{e} = \\ &= \mathbf{M}\mathbf{e}\end{aligned}$$

since $\mathbf{M}\mathbf{X} = 0$. Being $\mathbf{M}$ symmetric and idempotent:

$$\begin{aligned}\hat{\mathbf{e}}^\top\hat{\mathbf{e}} &= (\mathbf{M}\mathbf{e})^\top(\mathbf{M}\mathbf{e}) = \\ &= \mathbf{e}^\top\mathbf{M}^\top\mathbf{M}\mathbf{e} = \\ &= \mathbf{e}^\top\mathbf{M}\mathbf{e}\end{aligned}$$

The expected value of the deviance of dispersion is:

$$\begin{aligned}\mathbb{E}\{\hat{\mathbf{e}}^\top\hat{\mathbf{e}}\} &= \mathbb{E}\{\mathbf{e}^\top\mathbf{M}\mathbf{e}\} = \\ &= \mathbb{E}\{trace(\mathbf{e}^\top\mathbf{M}\mathbf{e})\} = \\ &= \mathbb{E}\{trace(\mathbf{M}\mathbf{e}\mathbf{e}^\top)\} = \\ &= trace(\mathbf{M}\mathbb{E}\{\mathbf{e}\mathbf{e}^\top\}) = \\ &= \mathbb{E}\{\mathbf{e}\mathbf{e}^\top\} \cdot trace(\mathbf{M}\mathbf{I}_n) = \\ &= \sigma_e^2 \cdot trace(\mathbf{M}\,\mathbf{I}_n) = \\ &= \sigma_e^2 \cdot trace(\mathbf{M})\end{aligned}$$

since $\hat{\mathbf{e}}^\top\mathbf{M}\hat{\mathbf{e}}$ is a scalar. The trace of the matrix $\mathbf{M}$ is:

$$\begin{aligned}trace(\mathbf{M}) &= trace(\mathbf{I}_n - \mathbf{H}) = \\ &= trace(\mathbf{I}_n) - trace(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top) = \\ &= trace(\mathbf{I}_n) - trace(\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}) = \\ &= trace(\mathbf{I}_n) - trace(\mathbf{J}_{k+1}) = \\ &= n - k - 1\end{aligned}$$

Hence the expectation of the deviance of dispersion is:

$$\mathbb{E}\{\mathbb{D}ev\mathbb{D}isp\{\mathbf{y}\}\} = \sigma_e^2 \cdot (n - k - 1) \implies \hat{s}_e^2 = \frac{\hat{\mathbf{e}}^\top\hat{\mathbf{e}}}{n - k - 1}$$

Instead, in general the expected value of the deviance of regression is greater than σ_e^2, i.e.

$$\mathbb{E}\{\mathbb{D}ev\mathbb{R}eg\{\mathbf{y}\}\} = k \cdot \sigma_e^2 + g(\mathbf{b}, \mathbf{X}), \quad g(\mathbf{b}, \mathbf{X}) \geq 0$$

In the special case in which $b_1 = b_2 = \cdots = b_k$ in population, then $g(\mathbf{b}, \mathbf{X})$ is zero and also the variance of regression produces a correct estimate of σ_e^2. □

14.5 Test on the parameters

Let's consider a linear model where the residuals $\mathbf{e}$ are IID normally distributed random variables. Hence, the working hypothesis of the Gauss Markov theorem holds true.

14.5.1 F-test

The F-test evaluates the significance of the entire regression model by testing the null hypothesis of **linear independence between y and X**, i.e.

$$H_0 : b_1 = b_2 = \cdots = b_k = 0$$

where the only coefficient different from zero is the intercept. The test statistic is given by

$$F = \frac{\hat{s}_r^2}{\hat{s}_e^2} = \frac{\mathbb{D}ev\mathbb{R}eg(\mathbf{y}) \cdot (n-k-1)}{k \cdot \mathbb{D}ev\mathbb{D}isp(\mathbf{y})} \sim \mathrm{F}_{k,n-k-1}$$

where $\hat{s}_r^2$ is the regression variance, $\hat{s}_e^2$ is the dispersion variance. By fixing a significance level α, the null hypothesis H_0 is rejected if $F > F_\alpha$.

⚠ Interpretation F-test

If the null hypothesis H_0 is **rejected** then:

- The variability of Y explained by the model is significantly greater than the residual variability.
- At least one of the k regressors has a coefficient b_k that is significantly different from zero in the population.

On contrary if H_0 is **not rejected**, then the model is not adequate and there is no evidence of a linear relation between $\mathbf{y}$ and $\mathbf{X}$.

Remembering the relation between the deviance and the R^2, i.e.

- $\mathbb{D}ev\mathbb{R}eg(\mathbf{y}) = R^2\mathbb{D}ev(\mathbf{y})$.
- $\mathbb{D}ev\mathbb{D}isp(\mathbf{y}) = (1-R^2)\mathbb{D}ev(\mathbf{y})$.

it is possible to express the F-test in terms of the multivariate R^2 as:

$$F = \frac{R^2}{1-R^2}\frac{n-k-1}{k} \sim \mathrm{F}_{k,n-k-1}$$

14.5.2 t-test

The t-test evaluates the significance of the one regression parameter by testing the null hypothesis of **linear independence between y and $\mathbf{X}_j$ given the effect of the others $k-1$ regressors**, i.e.

$$H_0 : \hat{b}_j = b_j$$

If the normality of the residuals holds true, then $\hat{\mathbf{b}}$ is a multivariate normal and so $\hat{b}_j$ is normally distributed. Standardizing:

$$t = \frac{\hat{b}_j - b_j}{\sqrt{\sigma_e^2 \cdot c_{jj}}} \sim \mathcal{N}(0,1) \tag{14.11}$$

Under H_0 and substituting σ_e^2 with it's correct estimate $\hat{\sigma}_e^2$, then

$$t \overset{H_0}{=} \frac{\hat{b}_j}{\sqrt{\hat{\sigma}_e^2 \cdot c_{jj}}} \sim t_{n-k-1}$$

14.5.3 Confidence intervals

From Equation 14.11 it is possible to build an interval with confidence level α for b_j as:

$$\begin{aligned} b_j &= \hat{b}_j \pm t_{\alpha/2,n-k-1}\sqrt{\sigma_e^2 \cdot c_{jj}} \\ &= \hat{b}_j \pm t_{\alpha/2,n-k-1}\sqrt{\mathbb{V}\{\hat{b}_j|\mathbf{X}\}} \end{aligned}$$

where $t_{\alpha/2,n-k-1}$ is the quantile at level $\alpha/2$ of a Student-t distribution with $n-k-1$ degrees of freedom and c_{jj} is the j-th element on the diagonal of $(\mathbf{X}^\top\mathbf{X})^{-1}$.

15 Generalized least square

15.1 Working hypothesis

The **assumptions** of the generalized least square regression are:

1. $\mathbb{E}\{y_i|\mathbf{x}_1, \dots, \mathbf{x}_n\} = \mathbb{E}\{y_i|\mathbf{X}\} = \mathbf{x}_i^\top \mathbf{b}$.
2. $\mathbb{V}\{y_i|\mathbf{x}_1, \dots, \mathbf{x}_n\} = \mathbb{V}\{y_i|\mathbf{X}\} = \sigma_i^2$ with $0 < \sigma_i^2 < \infty$.
3. $\mathbb{C}v\{y_i, y_j|\mathbf{x}_1, \dots, \mathbf{x}_n\} = \mathbb{C}v\{y_i, y_j|\mathbf{X}\} = \sigma_{ij}$

equivalently the formulation in terms of the stochastic component $\mathbf{u}$ reads

1. $y_i = \mathbf{x}_i^\top \mathbf{b} + e_i$ for $i = 1, \dots n$.
2. $\mathbb{E}\{e_i|\mathbf{x}_1, \dots, \mathbf{x}_n\} = \mathbb{E}\{e_i|\mathbf{X}\} = 0$.
3. $\mathbb{V}\{e_i|\mathbf{x}_1, \dots, \mathbf{x}_n\} = \mathbb{V}\{e_i|\mathbf{X}\} = \sigma_i^2$ with $0 < \sigma_i^2 < \infty$.
4. $\mathbb{C}v\{e_i, e_j|\mathbf{x}_1, \dots, \mathbf{x}_n\} = \mathbb{C}v\{e_i, e_j|\mathbf{X}\} = \sigma_{ij}$

In this case the variance covariance matrix Σ is defined as in Equation 13.3 and contains the variances and the covariances between the observations.

15.2 Generalized least squares estimator

Proposition 15.1. *(Generalized Least Squares (GLS))*
The generalized least squares estimator (GLS) is the function Q that minimize the weighted sum of the squared residuals and return an estimate of the true parameter $\mathbf{b}$, i.e. $\hat{\mathbf{b}} = \mathbf{b}^{GLS}$. The GLS optimization problem reads:

$$\underset{\mathbf{b}^{GLS} \in \Theta_{\mathbf{b}}}{argmin}\, Q(\mathbf{b}^{GLS}) = \underset{\mathbf{b}^{GLS} \in \Theta_{\mathbf{b}}}{argmin} \left\{ \mathbf{e}(\mathbf{b}^{GLS})^\top \Sigma^{-1} \mathbf{e}(\mathbf{b}^{GLS}) \right\} \tag{15.1}$$

*Notably, if $\mathbf{X}$ and Σ are **non-singular** it is possible to recover an analytic solution, i.e.*

$$\mathbf{b}^{GLS} = (\mathbf{X}^\top \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \Sigma^{-1} \mathbf{y} \tag{15.2}$$

⚠ Singularity of $\mathbf{X}$ or Σ

The solution is available if and only if $\mathbf{X}$ and Σ are **non-singular**. In practice the conditions are:

1. $rank(\Sigma) = \max = n$ for the inversion of Σ.
2. $rank(\mathbf{X}) = \max = k$ and condition 1. for the inversion of $\mathbf{X}^\top\Sigma^{-1}\mathbf{X}$.

ℹ Generalized Least Square (GLS)

Proof. Let's prove the optimal solution in Proposition 15.1. Developing the optimization problem in Equation 15.1:

$$\begin{aligned} Q(\mathbf{b}^{\text{GLS}}) &= \mathbf{e}(\mathbf{b}^{\text{GLS}})^\top\Sigma^{-1}\mathbf{e}(\mathbf{b}^{\text{GLS}}) = \\ &= (\mathbf{y} - \mathbf{X}\mathbf{b}^{\text{GLS}})^\top\Sigma^{-1}(\mathbf{y} - \mathbf{X}\mathbf{b}^{\text{GLS}}) = \\ &= \mathbf{y}^\top\Sigma^{-1}\mathbf{y} - 2(\mathbf{b}^{\text{GLS}})^\top\mathbf{X}^\top\Sigma^{-1}\mathbf{y} + (\mathbf{b}^{\text{GLS}})^\top\mathbf{X}^\top\Sigma^{-1}\mathbf{X}\mathbf{b}^{\text{GLS}} \end{aligned}$$

In order to find the minimum, let's compute the derivative with respect to $\mathbf{b}^{\text{GLS}}$ of Q and setting it equal to zero, i.e.

$$\begin{aligned} &\frac{d\text{Q}(\mathbf{b}^{\text{GLS}})}{d\mathbf{b}^{\text{OLS}}} = -2\mathbf{X}^\top\Sigma^{-1}\mathbf{y} + 2\mathbf{X}^\top\Sigma^{-1}\mathbf{X}\mathbf{b}^{\text{GLS}} = 0 \\ &\implies \mathbf{X}^\top\Sigma^{-1}\mathbf{y} = \mathbf{X}^\top\Sigma^{-1}\mathbf{X}\mathbf{b}^{\text{GLS}} \\ &\implies \mathbf{b}^{\text{GLS}} = (\mathbf{X}^\top\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\Sigma^{-1}\mathbf{y} \end{aligned}$$

□

15.3 Properties GLS

Theorem 15.1. *(**Aikten theorem**)*
*Under the hypothesis of the Generalized linear models the Generalized Least Square (GLS) estimate is **BLUE** (**Best Linear Unbiased Estimator**), where "*best*" stands for the estimator with* minimum variance *in the class of linear unbiased estimators of* $\mathbf{b}$*. The Aikten hypothesis are:*

1. $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$.
2. $\mathbb{E}\{\mathbf{e}\} = 0$.
3. $\mathbb{E}\{\mathbf{e}\mathbf{e}^\top\} = \Sigma$, *i.e. heteroskedastic and correlated errors.*
4. $\mathbf{X}$ *is non-stochastic and independent from the errors* $\mathbf{e}$ *for all n's.*

Proposition 15.2. *(Properties GLS estimator)*

1. ***Unbiased:*** $\mathbf{b}^{GLS}$ *is correct and it's conditional expectation is equal to true parameter in population, i.e.*
$$\mathbb{E}\{\mathbf{b}^{GLS}|\mathbf{X}\} = \mathbf{b} \tag{15.3}$$
2. ***Linear*** *in the sense that it can be written as a linear combination of* $\mathbf{y}$ *and* $\mathbf{X}$*, i.e.* $\mathbf{b}^{GLS} = \mathbf{A}_x\mathbf{y}$*, where* $\mathbf{A}_x$ *do not depend on* $\mathbf{y}$*, i.e.*
$$\mathbf{b}^{GLS} = \mathbf{A}_x\mathbf{y} \quad \mathbf{A}_x = (\mathbf{X}^\top\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\Sigma^{-1} \tag{15.4}$$
3. *Under the Aikten hypothesis (Theorem 15.1) it has* ***minimum variance*** *in the class of the unbiased linear estimators and it reads:*
$$\mathbb{V}\{\mathbf{b}^{GLS}|\mathbf{X}\} = (\mathbf{X}^\top\Sigma^{-1}\mathbf{X})^{-1} \tag{15.5}$$

Properties of the GLS estimator

Proof.

- The GLS estimator is **correct**. It's expected value is computed from Equation 15.2 and substituting Equation 13.1, is equal to the true parameter in population, i.e.
$$\begin{aligned}\mathbb{E}\{\mathbf{b}^{\text{GLS}}|\mathbf{X}\} &= \mathbb{E}\{(\mathbf{X}^\top\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\Sigma^{-1}\mathbf{y}\} = \\ &= \mathbb{E}\{(\mathbf{X}^\top\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\Sigma^{-1}(\mathbf{X}\mathbf{b}+\mathbf{e})\} = \\ &= (\mathbf{X}^\top\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\Sigma^{-1}\mathbf{X}\mathbf{b} + (\mathbf{X}^\top\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\Sigma^{-1}\mathbb{E}\{\mathbf{e}|\mathbf{X}\} = \\ &= \mathbf{b}\end{aligned} \tag{15.6}$$
- Under the assumption of heteroskedastic and correlated observations the conditional variance of $\mathbf{b}^{\text{GLS}}$ follows similarly as for the OLS case (Equation 14.10) but with $\mathbb{V}\{\mathbf{e}|\mathbf{X}\} = \Sigma$, i.e.
$$\begin{aligned}\mathbb{V}\{\mathbf{b}^{\text{GLS}}|\mathbf{X}\} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Sigma^{-1}\mathbb{V}\{\mathbf{e}|\mathbf{X}\}\,\Sigma^{-1}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} = \\ &= (\mathbf{X}^\top\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\Sigma^{-1}\Sigma\Sigma^{-1}\mathbf{X}(\mathbf{X}^\top\Sigma^{-1}\mathbf{X})^{-1} = \\ &= (\mathbf{X}^\top\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\Sigma^{-1}\mathbf{X}(\mathbf{X}^\top\Sigma^{-1}\mathbf{X})^{-1} = \\ &= (\mathbf{X}^\top\Sigma^{-1}\mathbf{X})^{-1}\end{aligned} \tag{15.7}$$
where Equation 14.9 become a special case of Equation 15.7 where $\Sigma = \sigma_e^2\mathbf{I}_n$.

□

15.4 Alternative derivation

Let's consider a linear model of the form $\mathbf{y} = \mathbf{Xb} + \varepsilon$ and a transformation matrix $\mathbf{T}_{n\times n}$. Multiplying on both sides by $\mathbf{T}$, the model can be rewritten as follows:

$$\begin{aligned} \mathbf{T}\,\mathbf{y} &= \mathbf{TX}\,\mathbf{b} + \mathbf{T}\varepsilon \\ \Downarrow \quad & \quad \Downarrow \quad \Downarrow \\ \tilde{\mathbf{y}} \quad &= \quad \tilde{\mathbf{X}}\,\mathbf{b} \; + \; \tilde{\varepsilon} \end{aligned}$$

The conditional mean of the transformed models reads as:

$$\mathbb{E}\{\tilde{\mathbf{y}}|\tilde{\mathbf{X}}\} = \tilde{\mathbf{X}}\,\mathbf{b}$$

while it's conditional variance

$$\mathbb{V}\{\tilde{\mathbf{y}}|\tilde{\mathbf{X}}\} = \mathbb{V}\{\tilde{\varepsilon}|\tilde{\mathbf{X}}\} = \mathbf{T}\,\Sigma\,\mathbf{T}^{\top}$$

The idea is to identify a transformation matrix $\mathbf{T}$ such that the conditional variance became equal to the identity matrix, i.e. $\mathbb{V}\{\tilde{\varepsilon}|\tilde{\mathbf{X}}\} = \mathbf{I}_n$. In this way it is possible to work under the Gauss-Markov assumptions obtaining an estimator with minimum variance. Let's decompose the variance-covariance matrix (Equation 13.3) as

$$\Sigma = \mathbf{e}\,\Lambda\,\mathbf{e}^{T}$$

where

- Λ is the diagonal matrix containing the eigenvalues.
- $\mathbf{e}$ is the matrix with the eigenvectors that satisfy the following relation, i.e. $\mathbf{e}^{\top}\mathbf{e} = \mathbf{e}\mathbf{e}^{\top} = \mathbf{J}_n$.

Setting the transformation matrix as $\mathbf{T} = \Lambda^{-1/2}\,\mathbf{e}^{\top}$ gives that the conditional variance is equal to 1 for all the observations, i.e.

$$\begin{aligned} \mathbb{V}\{\tilde{\varepsilon}|\tilde{\mathbf{X}}\} &= \mathbf{T}\,\Sigma\,\mathbf{T}^{\top} = \\ &= (\Lambda^{-1/2}\,\mathbf{e}^{\top})\,\mathbf{e}\,\Lambda\,\mathbf{e}^{\top}\,(\mathbf{e}\,\Lambda^{-1/2}) = \\ &= \Lambda^{-1/2}\Lambda\,\Lambda^{-1/2} = \mathbf{J}_n \end{aligned}$$

Moreover, the matrix $\mathbf{T} = \Lambda^{-1/2}\,\mathbf{e}^{\top}$ satisfies the product:

$$\mathbf{T}^{\top}\mathbf{T} = \mathbf{e}\,\Lambda^{-1/2}\Lambda^{-1/2}\mathbf{e}^{\top} = \mathbf{e}\,\Lambda^{-1}\mathbf{e}^{\top} = \Sigma^{-1} \tag{15.8}$$

Finally, substituting $\tilde{\mathbf{X}} = \mathbf{TX}$ in the OLS formula (Equation 14.2) and using the result Equation 15.8 one obtain exactly the GLS estimator in Equation 15.2, i.e.

$$\begin{aligned} \tilde{\mathbf{b}} &= (\tilde{\mathbf{X}}^{\top}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^{\top}\tilde{\mathbf{y}} = \\ &= (\mathbf{X}^{\top}\mathbf{T}^{\top}\mathbf{TX})^{-1}\mathbf{X}^{\top}\mathbf{T}^{\top}\mathbf{Ty} = \\ &= (\mathbf{X}^{\top}\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^{\top}\Sigma^{-1}\mathbf{y} \end{aligned}$$

15.5 Models with heteroskedasticity

15.5.1 Working hypothesis

The **assumptions** of the generalized linear model with heteroskedastic errors are:

1. $\mathbb{E}\{y_i|\mathbf{x}_1, ..., \mathbf{x}_n\} = \mathbb{E}\{y_i|\mathbf{X}\} = \mathbf{x}_i^\top \mathbf{b}$.
2. $\mathbb{V}\{y_i|\mathbf{x}_1, ..., \mathbf{x}_n\} = \mathbb{V}\{y_i|\mathbf{X}\} = \sigma_i^2$ with $0 < \sigma_i^2 < \infty$.
3. $\mathbb{C}v\{y_i, y_j|\mathbf{x}_1, ..., \mathbf{x}_n\} = \mathbb{C}v\{y_i, y_j|\mathbf{X}\} = 0$

equivalently the formulation in terms of the stochastic component

1. $y_i = \mathbf{x}_i^\top \beta + e_i$ for $i = 1, ... n$.
2. $\mathbb{E}\{e_i|\mathbf{x}_1, ..., \mathbf{x}_n\} = \mathbb{E}\{e_i|\mathbf{X}\} = \mathbf{0}$.
3. $\mathbb{V}\{e_i|\mathbf{x}_1, ..., \mathbf{x}_n\} = \mathbb{V}\{e_i|\mathbf{X}\} = \sigma_i^2$ with $0 < \sigma_i^2 < \infty$.
4. $\mathbb{C}v\{e_i, e_j|\mathbf{x}_1, ..., \mathbf{x}_n\} = \mathbb{C}v\{e_i, e_j|\mathbf{X}\} = 0$

For an heteroskedastic linear model the variance-covariance matrix of the residuals in matrix notation is written as:

$$\underset{n\times n}{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & ... & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & ... & \sigma_n^2 \end{pmatrix} = \text{diag}\left(\sigma_1^2,\ \sigma_2^2,\ ...,\ \sigma_n^2\right)$$

16 Restricted linear models

16.1 A general framework for linear restrictions

Let's consider a generic uni-variate linear model with k-regressors, namely

$$\mathbf{y} = b_1\mathbf{X}_1 + \cdots + b_j\mathbf{X}_j + \cdots + b_k\mathbf{X}_k + \mathbf{e} = \mathbf{bX} + \mathbf{e}$$

and suppose that we are interest in testing if the b_j coefficient is statistically different from a certain known value r. In this case the null hypothesis, that is $H_0 : b_j = r$, can be equivalently represented using a more flexible matrix notation, i.e.

$$H_0 : b_j = r \iff H_0 : \mathbf{R}^\top\mathbf{b} - \mathbf{r} = \mathbf{0}$$

where

$$\underset{k\times 1}{\mathbf{R}}^\top = \underbrace{(0 \ \ldots \ 1 \ \ldots \ 0)}_{j\text{-th position}}$$

Hence, the linear restriction in matrix form reads explicitly as

$$H_0 : \underset{k\times 1}{\mathbf{R}}^\top \underset{k\times 1}{\mathbf{b}} - \underset{1\times 1}{\mathbf{r}} = \underset{1\times 1}{\mathbf{0}} \iff \underbrace{(0 \ \ldots \ 1 \ \ldots \ 0)}_{j\text{-th position}} \begin{pmatrix} b_1 \\ \vdots \\ b_j \\ \vdots \\ b_k \end{pmatrix} - (r) = (0)$$

16.2 Multiple restrictions

Let's consider a linear model of the form

$$\mathbf{y} = b_1\mathbf{X}_1 + b_2\mathbf{X}_2 + b_3\mathbf{X}_3 + b_4\mathbf{X}_4$$

and suppose that the aim is to test at the same time the following null hypothesis, i.e.

$$\begin{aligned} H_0 : \quad & (1) \quad b_1 - b_2 = 0 \quad b_1 \text{ and } b_2 \text{ has same effect} \\ & (2) \quad b_3 + b_4 = 1 \quad b_3 \text{ plus } b_4 \text{ unitary root} \end{aligned}$$

Let's construct the vector for **(1)** (first column of R) and **(2)** (first column of R), i.e.

$$\underset{2\times 4}{\mathbf{R}}^\top \underset{4\times 1}{\mathbf{b}} - \underset{2\times 1}{\mathbf{r}} = \underset{2\times 1}{\mathbf{0}} \iff \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

16.3 Restricted least squares

Proposition 16.1. *Let's consider a set of m linear hypothesis on the parameters of the model taking the form*

$$H_0 : \underset{m\times k}{\mathbf{R}}^{\top} \underset{k\times 1}{\mathbf{b}} - \underset{m\times 1}{\mathbf{r}} = \underset{m\times 1}{\mathbf{0}}$$

Then the parameters that satisfies the condition are no more in $\Theta_{\mathbf{b}}$ but in a subset $\tilde{\Theta}_{\mathbf{b}}$ where the linear constraint holds true, i.e.

$$\tilde{\Theta}_{\mathbf{b}} = \{\mathbf{b} \in \mathbb{R}^k : \mathbf{R}^{\top}\mathbf{b} - \mathbf{r} = \mathbf{0}\}$$

Hence, the optimization problem is restricted to search only the parameters that satisfy the constraint, i.e.

$$\underset{\mathbf{b}^{RLS}\in\tilde{\Theta}_{\mathbf{b}}}{argmin}\, Q(\mathbf{b}^{RLS}) = \underset{\mathbf{b}^{RLS}\in\tilde{\Theta}_{\mathbf{b}}}{argmin} \{\mathbf{e}(\mathbf{b}^{RLS})^T\mathbf{e}(\mathbf{b}^{RLS})\} \tag{16.1}$$

Notably, it is available an analytic solution for $\mathbf{b}^{RLS}$, i.e.

$$\mathbf{b}^{RLS} = \mathbf{b}^{OLS} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{R}\left[\mathbf{R}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{R}\right]^{-1}(\mathbf{R}^T\mathbf{b}^{OLS} - \mathbf{r}) \tag{16.2}$$

i Restricted Least Square (RLS)

Proof. In order to solve the minimization problem in Equation 16.1, let's construct the Lagrangian $L(x, \lambda)$, i.e.

$$L(x, \lambda) = f(x) - \lambda^{\top} g(x),$$

where λ is the vector of the **Lagrange multipliers**. Minimizing $L(x, \lambda)$ is equivalent to find the value of x that minimize $f(x)$ under the constraint $g(x) = 0$. In fact, it is possible to prove that the minimum is found as:

$$\underset{x\in\chi}{\operatorname{argmin}}\, L(x, \lambda) \implies \begin{cases} (A) & \partial_x L(x, \lambda) = 0 \\ (B) & \partial_\lambda L(x, \lambda) = 0 \end{cases} \tag{16.3}$$

In the case of RLS estimate the Lagrangian reads:

$$L(\mathbf{b}^{\text{RLS}}, \lambda) = Q(\mathbf{b}^{\text{RLS}}) - 2\lambda^{\top}(\mathbf{R}^{\top}\mathbf{b}^{\text{RLS}} - \mathbf{r}),$$

where Q is the same loss function defined for the OLS case (Equation 14.4) and 2 is a constant. Then, from Equation 16.3 one obtain the following system of equation, i.e.

$$\begin{cases} (A) & \partial_{\mathbf{b}^{\text{RLS}}} L(\mathbf{b}^{\text{RLS}}, \lambda) = -2\mathbf{X}^{\top}\mathbf{y} + 2\mathbf{X}^{\top}\mathbf{X}\mathbf{b}^{\text{RLS}} - 2\mathbf{R}\lambda = \mathbf{0} \\ (B) & \partial_\lambda L(\mathbf{b}^{\text{RLS}}, \lambda) = -2(\mathbf{R}^{\top}\mathbf{b}^{\text{RLS}} - \mathbf{r}) = \mathbf{0} \end{cases}$$

Let's explicit $\mathbf{b}^{\text{RLS}}$ from **(A)**, i.e.

$$\begin{aligned}\mathbf{b}^{\text{RLS}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}\lambda = \\ &= \mathbf{b}^{\text{OLS}} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}\lambda\end{aligned} \tag{16.4}$$

Let's now substitute Equation 16.4 in **(B)**, i.e.

$$\begin{aligned}&\mathbf{R}^\top\mathbf{b}^{\text{RLS}} - \mathbf{r} = \mathbf{0} \\ \implies &\mathbf{R}^\top\left[\mathbf{b}^{\text{OLS}} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}\lambda\right] - \mathbf{r} = \mathbf{0} \\ \implies &\mathbf{R}^\top\mathbf{b}^{\text{OLS}} - \mathbf{R}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}\lambda - \mathbf{r} = \mathbf{0} \\ \implies &\mathbf{R}^\top\mathbf{b}^{\text{OLS}} - \mathbf{r} = \left[\mathbf{R}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}\right]\lambda\end{aligned}$$

Hence, it is possible to explicit the Lagrange multipliers λ as:

$$\lambda = \left[\mathbf{R}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}\right]^{-1}(\mathbf{R}^\top\mathbf{b}^{\text{OLS}} - \mathbf{r}) \tag{16.5}$$

Finally, substituting Equation 16.5 in Equation 16.4 gives the optimal solution, i.e.

$$\begin{aligned}\mathbf{b}^{\text{RLS}} &= \mathbf{b}^{\text{OLS}} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}\lambda = \\ &= \mathbf{b}^{\text{OLS}} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}\left[\mathbf{R}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}\right]^{-1}(\mathbf{R}^\top\mathbf{b}^{\text{OLS}} - \mathbf{r})\end{aligned}$$

Note that if constraints hold true in the OLS estimate, H_0 is true and therefore $\mathbf{R}^\top\mathbf{b}^{\text{OLS}} - \mathbf{r} = \mathbf{0}$. Hence the RLS and OLS parameters are the same, i.e. $\mathbf{b}^{\text{RLS}} = \mathbf{b}^{\text{OLS}}$. □

16.4 Properties RLS

1. The RLS estimator is **correct** if and only if the restriction imposed by H_0 is true in population. In fact, it's expected value is computed as:

$$\mathbb{E}\{\mathbf{b}^{\text{RLS}} \mid \mathbf{X}\} = \mathbf{b} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}\left[\mathbf{R}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}\right]^{-1}(\mathbf{R}^\top\mathbf{b} - \mathbf{r}) \tag{16.6}$$

   and it is correct if and only if the second component is zero, i.e. if H_0 holds true.

i Correctness of RLS estimator

Proof. Let's apply the expected value on Equation 16.2 remembering that $\mathbf{X}$, $\mathbf{R}$ and $\mathbf{r}$ are non-stochastic and that $\mathbf{b}^{\text{OLS}}$ is correct (Equation 14.7). Developing the computations

gives:

$$\begin{aligned}\mathbb{E}\{\mathbf{b}^{\text{RLS}} \mid \mathbf{X}\} &= \mathbb{E}\{\mathbf{b}^{\text{OLS}} \mid \mathbf{X}\} - \mathbb{E}\left\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}\left[\mathbf{R}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}\right]^{-1}(\mathbf{R}^\top\mathbf{b}^{\text{OLS}} - \mathbf{r}) \mid \mathbf{X}\right\} = \\ &= \mathbf{b} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}\left[\mathbf{R}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}\right]^{-1}(\mathbf{R}^\top\mathbb{E}\{\mathbf{b}^{\text{OLS}} \mid \mathbf{X}\} - \mathbf{r}) = \\ &= \mathbf{b} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}\left[\mathbf{R}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}\right]^{-1}(\mathbf{R}^\top\mathbf{b} - \mathbf{r})\end{aligned}$$

Hence $\mathbf{b}^{\text{RLS}}$ is correct if and only if the restriction holds in population.

$$\mathbb{E}\{\mathbf{b}^{\text{RLS}} \mid \mathbf{X}\} = \mathbf{b} \iff \mathbf{R}^\top\mathbf{b} - \mathbf{r} = \mathbf{0}$$

□

16.5 A test for linear restrictions

In order to build a test for the linear restriction imposed by $\mathbf{R}^\top\mathbf{b} - \mathbf{r} = \mathbf{0}$, it is necessary that the stochastic component $\mathbf{e}$ is normally distributed. Under normality:

$$\mathbf{b}^{\text{OLS}} \sim \mathcal{N}(\mathbf{b}, \sigma_e^2(\mathbf{X}^\top\mathbf{X})^{-1})$$

Then, let's consider the hull hypothesis H_0 and the alternative H_1, i.e.

$$H_0 : \mathbf{R}^\top\mathbf{b}^{\text{OLS}} - \mathbf{r} = \mathbf{0} \quad \text{vs} \quad H_1 : \mathbf{R}^\top\mathbf{b}^{\text{OLS}} - \mathbf{r} \neq \mathbf{0}$$

Using the scaling property of the multivariate normal:

$$\mathbf{R}^\top\mathbf{b}^{\text{OLS}} - \mathbf{r} \sim \mathcal{N}(\mathbf{R}^\top\mathbf{b} - \mathbf{r},\ \sigma_e^2\mathbf{R}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R})$$

Remembering the connection between the distribution of the quadratic form of a multivariate normal and the χ^2 distribution in in property 3. (Section 33.1.2) one obtain the following statistic:

$$T_q = (\mathbf{R}^\top\mathbf{b} - \mathbf{r})^\top(\sigma_e^2\mathbf{R}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R})^{-1}(\mathbf{R}^\top\mathbf{b} - \mathbf{r}) \tag{16.7}$$

that under H_0 is distributed as a χ^2_q where q is the number of linear restrictions, i.e.

$$T_q \overset{H_0}{\sim} \chi^2(q) \tag{16.8}$$

Instead, under H_1 it is possible to use the result in property 4. (Section 33.1.2) to show that the statistic T_q is distributed as a non central $\chi^2(q, \delta)$, i.e.

$$T_q \overset{H_1}{\sim} \chi^2(q, \delta) \tag{16.9}$$

where the non centrality parameter δ is computed as:

$$\delta = (\mathbf{R}^\top\mathbf{b} - \mathbf{r})^\top(\sigma_e^2\mathbf{R}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R})^{-1}(\mathbf{R}^\top\mathbf{b} - \mathbf{r})$$

As general decision rule H_0 is rejected if the statistic in Equation 16.8 is greater than the quantile with confidence level α of a $\chi^2(q)$ random variable. Such critic value, denoted with $\chi^2_\alpha(q)$ represents the value for which the probability that a $\chi^2(q)$ is greater than the value $\chi^2_\alpha(q)$ is exactly α, i.e.

$$\mathbb{P}(\chi^2_q > x_\alpha) = \alpha$$

In this case the probability to have an **error of type I**, i.e. rejecting H_0 when H_0 is true is exactly α.

17 Multiequationals linear models

Let's consider a multivariate linear model, i.e. with $p > 1$ in (Equation 12.3), then the model in matrix notation reads:

$$\underset{n\times p}{Y} = \underset{n\times 1}{\mathbf{J}_{n,1}}\ \underset{1\times p}{a^\top} + \underset{n\times k}{X}\ \underset{k\times p}{b^\top} + \underset{n\times p}{e}$$

17.1 OLS estimate

As in the uni-variate case the optimal parameters are computed as:

$$\mathbf{b}^{\text{OLS}} = \mathbb{C}v(Y, X)\,\mathbb{C}v(X)^{-1}$$
$$\alpha^{\text{OLS}} = \mathbb{E}\{Y\} - \mathbf{b}^{\text{OLS}}\,\mathbb{E}\{X\}$$

And the variance covariance matrix of the residuals is computed as:

$$\Sigma = \mathbb{C}v(e) = \mathbb{C}v(Y) - \mathbf{b}^{\text{OLS}}\,\mathbb{C}v(Y, X)$$

17.1.1 Example

Let's consider n-simulated observations of the explicative variables $\mathbf{X}$ drown from a multivariate normal, i.e. $\mathbf{X} \sim \mathcal{N}(\mathbb{E}\{X\}, \mathbb{C}v\{X\})$, with parameters

$$\mathbb{E}\{X\} = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} \quad \mathbb{C}v\{X\} = \begin{pmatrix} 0.5 & 0.2 & 0.1 \\ 0.2 & 1.2 & 0.1 \\ 0.1 & 0.1 & 0.3 \end{pmatrix}$$

Let's consider two dependent variables, hence $p = 2$ and $k = 3$. Let's now simulate the $p \times k = 6$ slopes parameter drown from a standard normal, i.e. for $j = 1, \dots, 6$, $b_j \sim \mathcal{N}(0, 1)$. The intercept parameters $\mathbf{a}$ are simulated drown from a uniform distribution in [0,1]. In the multivariate case $\mathbf{a}$ and $\mathbf{b}$ became matrices, i.e.

$$\underset{p\times k}{\mathbf{b}} = \begin{pmatrix} b_{1,1} & b_{1,2} & b_{1,3} \\ b_{2,1} & b_{2,2} & b_{2,3} \end{pmatrix} \quad \underset{p\times 1}{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$

Table 17.1: Fitted parameters

Type	β_1	β_2	β_3
True	0.8500	-0.9253	0.8936
True	-0.9410	0.5390	-0.1820
Fitted	0.8457	-0.8699	0.9396
Fitted	-0.9532	0.5518	-0.1804

Type	α_1	α_2
True	0.8137	0.8068
Fitted	0.7942	0.7423

For $i = 1, \dots, n$, let's consider a model of the form:

$$\begin{cases} Y_{i,1} = \beta_{0,1} + \beta_{1,1} X_{i,1} + \beta_{1,2} X_{i,2} + \beta_{1,k} X_{i,3} + u_{i,1} \\ Y_{i,2} = \beta_{0,2} + \beta_{2,1} X_{i,1} + \beta_{2,2} X_{i,2} + \beta_{2,k} X_{i,3} + u_{i,2} \end{cases}$$

where $u_{i,1}$ and $u_{i,2}$ are simulated from a multivariate normal random variables with true covariance matrix equal to:

$$\mathbb{C}v\{\mathbf{u}\} = \begin{pmatrix} 0.55 & 0.3 \\ 0.3 & 0.70 \end{pmatrix}$$

Hence, the procedure is structured as:

1. Simulate of the explanatory variables, the regression parameters and the residuals.
2. Simulate the perturbed $\tilde{\mathbf{Y}}$ (regression with errors).
3. Fit the regression parameters on the $\tilde{\mathbf{Y}}$.

4. Compute the fitted residuals from the prediction obtained with the parameters in step 3. and compute their variance covariance matrix.
5. Compare the results with the true parameters.

Part IV

Time Series

18 Time series

Let $\{y_t\}_{t\in\mathcal{T}}$ be a time series or stochastic process, i.e., a collection of random variables indexed by a set of time indices $\mathcal{T}$. For each $t \in \mathcal{T}$, define the filtration $\mathcal{F}_t$ as the information available up to time t, i.e.

$$\mathcal{F}_t = \sigma(y_0, y_1, \dots, y_t).$$

That is, $\mathcal{F}_t$ is the smallest σ-algebra containing all events observable up to time t. In more informal settings (when avoiding measure-theoretic details), we can write:

$$\mathcal{F}_t = \{y_0, y_1, \dots, y_t\}.$$

The filtration $\{\mathcal{F}_t\}_{t\in\mathcal{T}}$ is formally defined as an increasing sequence of σ-algebras:

$$\begin{aligned}
&\mathcal{F}_0 = \{y_0\} \\
&\mathcal{F}_1 = \mathcal{F}_0 \cup \{y_1\} = \{y_0, y_1\} \\
&\mathcal{F}_2 = \mathcal{F}_1 \cup \{y_2\} = \{y_0, y_1, y_2\} \\
&\vdots \\
&\mathcal{F}_t = \mathcal{F}_{t-1} \cup \{y_t\} = \{y_0, y_1, \dots, y_t\}
\end{aligned}$$

and represent the *information set*.

18.1 Stationarity

Definition 18.1. (**Strongly stationary**)
A process $\{y_t\}_{t\in\mathcal{T}}$ is **strongly stationary** if and only if for all set of index $\{t_1, t_2, \dots, t_n\} \in \mathcal{T}$ and for every $h > 0$

$$\mathbb{P}(y_{t_1}, y_{t_2}, \dots, y_{t_n}) = \mathbb{P}(y_{t_{1+h}}, y_{t_{2+h}}, \dots, y_{t_{n+h}}).$$

Hence, the joint distribution of a strongly stationary process is invariant with respect to a shift h in time.

Definition 18.2. (**Weakly stationary**)
A process $\{y_t\}_{t\in\mathcal{T}}$ is called **weakly stationary** or **covariance stationary** if and only if for every t and k:

1. $\mathbb{E}\{y_t\} = \mu$ and $|\mu| < \infty$.
2. $\mathbb{C}v\{y_t, y_{t+k}\} = \gamma(k)$ and $|\gamma(k)| < \infty$.

Hence, for a weakly stationary process, the expectation, variance are finite and constant and the covariance $\gamma(k)$ do not depends on time t, but only on the lag k between two observations.

> ⚠ Strong does not imply weakly and viceversa
>
> In general if a process is **strong stationary** (Definition 18.1) does not implies automatically that it is also weakly stationary. For example, an independent and identically Cauchy distributed process is strongly stationary, but since its expectation and variance are not finite the process is not weakly stationary.

18.2 Notable processes

Definition 18.3. (**IID process**)
A time series, $\{u_t\}_{t\in\mathcal{T}}$ where each u_t is independent from the others and all u_t has the same distribution for all t is called *independent and identically distributed* process (IID). Such kind of process, usually denoted as $u_t \sim \text{IID}(0, \sigma_u^2)$, is *strongly stationary* (Definition 18.1). Moreover, if the mean and variance are finite, the covariance is zero and the process is also *weakly stationary* (Definition 18.2), i.e.

$$\gamma_t(k) = \mathbb{C}v\{y_t, y_{t+k}\} = \mathbb{E}\{y_t y_{t+k}\} = \mathbb{E}\{y_t\}\mathbb{E}\{y_{t+k}\} = 0.$$

Definition 18.4. (**White noise**)
A time series $u_{t_{t\in\mathcal{T}}}$, commonly denoted as

$$u_t \sim \text{WN}(0, \sigma_u^2). \tag{18.1}$$

is called *White Noise* if satisfies the following properties:

1. The expectation is equal to zero, i.e. $\mathbb{E}\{u_t\} = 0$ for all $t \in \mathcal{T}$.
2. The variance is finite and constant for all $t \in \mathcal{T}$, i.e. $\mathbb{V}\{u_t\} = \sigma_u^2 < \infty$.
3. The process is uncorrelated over time for all $t \neq s$, i.e. $\mathbb{C}v\{u_t, u_s\} = 0$.

A White Noise process is weakly stationary (Definition 18.2). In fact, the autocovariance function of the process depends on the lag, but not on time, i.e. it is equal to the variance for $t = s$ and is zero otherwise. This process is more general than an IID process (Definition 18.3), since it does not requires the stochastic independence of the time series for all t.

Definition 18.5. (Martingale difference sequence) Let $u_t t \in \mathcal{T}$ be a stochastic process and let $\mathcal{F}tt \in \mathcal{T}$ be a filtration such that $\mathcal{F}t-1$ represents the information available up to time $t-1$. Then u_t is said to be a martingale difference sequence (MDS) with respect to the filtration $\mathcal{F}t$ if

$$\mathbb{E}\{u_t \mid \mathcal{F}_{t-1}\} = 0 \quad \forall t \in \mathcal{T}. \tag{18.2}$$

This implies that u_t is a mean-zero process uncorrelated with any information contained $\mathcal{F}_{t-1}$. The definition can be extended to a case where the filtration $\mathcal{F}_{t-1}$ includes also other processes X. In this case, u_t is said to be an *MDS conditionally to* X if the same condition in Equation 18.2 holds.

18.3 Lag operator

The lag operator L is a function that allows to translate a time series in time. In general, the lag operator associate at y_t it's lagged value y_{t-1}, i.e.

$$L(y_t) = y_{t-1}. \tag{18.3}$$

More formally, L is the operator that takes one whole time series and produces another; the second time series is the same as the first, but moved backwards or forward one point in time. From the definition, we list some properties related to the Lag operator, i.e.

1. **Backward** $L^k(y_t) = y_{t-k}$.
2. **Forward** $L^{-k}(y_t = y_{t+k}$.
3. $L(ay_t + bx_t) = ay_{t-1} + bx_{t-1}$.

18.3.1 Polynomial of Lag operator

Given a time series y_t, it is possible to define *polynomials of the Lag operator*, i.e.

$$\phi(L)y_t = y_t + \phi_1 y_{t-1} + \cdots + \phi_k y_{t-p} = \sum_{i=0}^{p} \phi_i y_{t-i}.$$

where in general

$$\phi(L) = 1 + \phi_1 L + \phi_2 L^2 + \cdots + \phi_p L^p = \sum_{j=0}^{p} \phi_j L^j. \tag{18.4}$$

For the polynomial $\phi(L)$ holds the factorization

$$\phi(L) = \left(1 - \frac{1}{z_1}L\right)\left(1 - \frac{1}{z_2}L\right)...\left(1 - \frac{1}{z_k}L\right) = \prod_{i=1}^{k}\left(1 - \frac{1}{z_i}L\right),$$

where $z_1, \dots, z_k$ are the complex solutions of the **characteristic equation**, i.e.

$$\phi(z) = 1 + \phi_1 z + \phi_2 z^2 + \cdots + \phi_k z^k = 0.$$

Hence the factorization holds true if and only if:

$$|z_i| > 1 \ \forall i \iff \frac{1}{|z_i|} < 1.$$

In other words, the modulus of the solutions must outside the unit circle, otherwise the geometric series is not convergent and the factorization do not holds true anymore. The factorization of the lag polynomial allows us to define its inverse, i.e.

$$\phi^{-1}(L) = \prod_{i=1}^{p} \left(1 - \frac{1}{z_i} L\right)^{-1},$$

In fact, the inverse of the i-th term can be expressed with a Taylor expansion as infinite sum if and only if $|\phi_i| < 1$, i.e.

$$(1 - \phi_i L)^{-1} = 1 + \phi_i L + (\phi_i L)^2 + \cdots = \sum_{j=0}^{\infty} \phi_i^j L^j \iff |\phi_i| < 1.$$

that is equivalent to $|z_i| > 1$ for all i since $\phi_i = \frac{1}{z_i}$.

AR(1) and geometric series

For example, let's consider an Autoregressive process of order 1, i.e.

$$y_t = \phi_1 y_{t-1} + e_t \iff \phi(L) y_t = e_t \iff y_t = \phi^{-1}(L) e_t$$

In fact,

$$\begin{aligned} \phi(L) y_t &= y_t - \phi_1 y_{t-1} = \\ &= y_t - \phi_1 y_t L = \\ &= y_t (1 - \phi_1 L) \end{aligned}$$

Considering such polynomial, its inverse polynomial $\phi(L)^{-1}$, defined such that $\phi(L)\phi^{-1}(L) = 1$, is defined as geometric series, i.e.

$$\phi^{-1}(L) = 1 + \phi_1 L + (\phi_1 L)^2 + \cdots = \sum_{j=0}^{\infty} \phi_1^j L^j = \frac{1}{1 - \phi_1 L} \iff |\phi_1| < 1,$$

that converges if and only if $|\phi_1| < 1$. Moreover, if $|\phi_1| < 1$ it is possible to prove that

$\phi^{-1}(L)$ is indeed the inverse polynomial of $\phi(L)$, in fact:

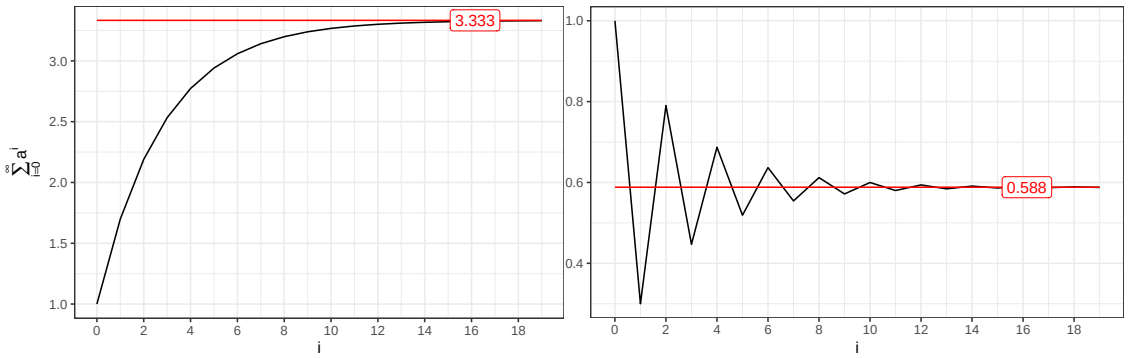
$$\begin{aligned}\phi(L)\phi^{-1}(L) &= (1-\phi_1 L)\cdot\sum_{j=0}^{\infty}(\phi_1 L)^j = \\ &= \sum_{j=0}^{\infty}(\phi_1 L)^j - \phi_1 L\sum_{j=0}^{\infty}(\phi_1 L)^j = \\ &= \sum_{j=0}^{\infty}(\phi_1 L)^j - \sum_{j=0}^{\infty}(\phi_1 L)^{j+1} = \\ &= \sum_{j=0}^{\infty}(\phi_1 L)^j - \sum_{j=0}^{\infty}(\phi_1 L)^j + 1 = 1\end{aligned}$$

Therefore, the process y_t can be equivalently expressed as:

$$y_t = \phi^{-1}(L)e_t = \sum_{j=0}^{\infty}\phi_1^j e_{t-j}$$

The factorization of any polynomial of the form of $\phi(L)$ is connected to the convergence of the following geometric series, i.e.

$$\sum_{j=0}^{\infty}\phi^j = \frac{1}{1-\phi} \iff |\phi| < 1.$$



(a) $0 < a < 1$. (b) $-1 < a < 0$.

Figure 18.1: Convergent series for AR(1) parameter (I).

Another important series that is convergent only if and only if $|a| < 1$, i.e.

$$\sum_{i=0}^{\infty}a^{2i} = \frac{1}{1-a^2} \iff |a| < 1.$$

Due to the square, in this case we do not distinguish between $0 < a < 1$ and $-1 < a < 0$ since they lead to the same result.

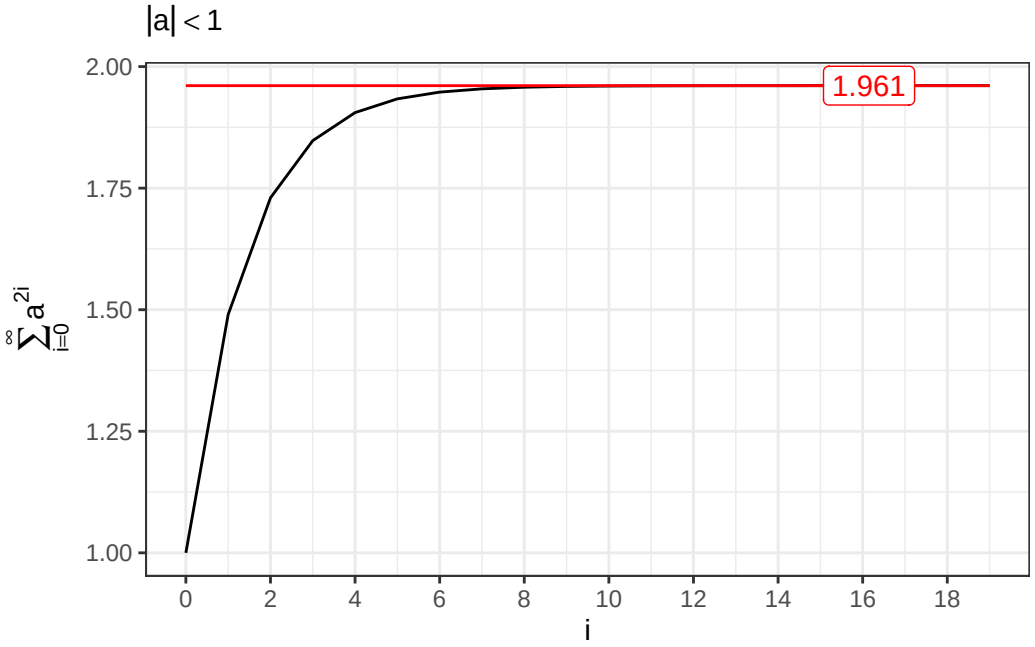


Figure 18.2: Convergent series for AR(1) parameter (II).

19 MA and AR processes

19.1 MA(q)

In time series analysis an univariate time series y_t is defined as Moving Average process with order q (MA(q)), when it satisfy the equations of the differences, i.e.

$$y_t = u_t + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q}, \tag{19.1}$$

where $u_t \sim \text{WN}(0, \sigma_u^2)$. An MA(q) process can be equivalently expressed as a polynomial in θ of the Lag operator (Equation 18.4), i.e.

$$y_t = \Theta(L)u_t, \quad u_t \sim \text{WN}(0, \sigma_u^2),$$

where $\Theta(L)$ is a polynomial of the form $\Theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q$. Given this representation it is clear that an MA(q) process is stationary independently on the value of the parameters. Moreover, the stationary process y_t admits has a **infinite moving average** or MA(∞) representation (see Wold (1939)) if it satisfies the equations of differences, i.e.

$$y_t = \Psi(L)u_t = u_t + \psi_1 u_{t-1} + \cdots = \sum_{i=1}^{\infty} \psi_j u_{t-j},$$

under the following condition the process is stationary and ergodic, i.e.

$$\sum_{i=1}^{\infty} |\psi_j| < \infty.$$

If the above condition holds, then the process MA(∞) can be written in compact form as:

$$y_t = \Psi(L)u_t, \quad u_t \sim \text{WN}(0, \sigma_u^2), \ \Psi(L) = \sum_{i=1}^{\infty} \psi_j L^j$$

19.1.1 Expectation

Proposition 19.1 (Expectation of an MA(q))**.** *In general, the expected value of an MA(q) process depends on the distribution of u_t. Under the standard assumption that u_t is a White Noise (Equation 18.1), then it's expected value is equal to zero for every t, i.e.*

$$\mathbb{E}\{y_t\} = \sum_{i=0}^{q} \theta_i \mathbb{E}\{u_{t-i}\} = 0.$$

Proof: Proposition 19.1

Proof. Given a process y_t such that $\mathbb{E}\{y_t\} = \mu$, it is always possible to simply reparametrize the Equation 19.1 as:

$$y_t = \mu + u_t + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q},$$

or rescale the process, i.e $\tilde{y}_t = y_t - \mu$, and work under a process with zero mean. Then, let's consider a process an MA process of order q, then the expectation of the process is computed as

$$\mathbb{E}\{y_t\} = \mathbb{E}\left\{u_t + \sum_{i=1}^{q} \theta_i u_{t-i}\right\} = \sum_{i=0}^{q} \theta_i \mathbb{E}\{u_{t-i}\} = 0 \quad .$$

Hence, the expected value of y_t depends on the expected value of the residuals u_t, that under the White Noise assumption is zero for every t. □

19.1.2 Autocovariance function

For every $k > 0$, the autocovariance function, denoted as γ_k, is defined as:

$$\gamma_k = \mathbb{C}v\{y_t, y_{t-k}\} = \begin{cases} \sigma_u^2 \sum_{i=1}^{q-k} \theta_i \theta_{i+k} & k \leq q \\ 0 & k > q \end{cases} \quad .$$

The covariance is different from zero only when the lag k is lower than the order of the process q. Setting $k = 0$ one obtain the variance, i.e.

$$\gamma_0 = \mathbb{V}\{y_t\} = \sigma_u^2 \sum_{i=1}^{q} \theta_i^2 \quad .$$

It follows that, the autocorrelation function is bounded up to the lag q, i.e.

$$\rho_k = \mathbb{C}r\{y_t, y_{t-k}\} = \begin{cases} 1 & k = 0 \\ \frac{\sigma_u^2 \sum_{i=1}^{q-k} \theta_i \theta_{i+k}}{\sigma_u^2 (1+\theta_1^2+\cdots+\theta_q^2)} & 0 < k \leq q \\ 0 & k > q \end{cases}$$

Proposition 19.2 (Moments of an MA(1))**.** *Let's consider a process $y_t \sim MA(1)$, i.e.*

$$y_t = \mu + \theta_1 u_{t-1} + u_t.$$

Independently, from the specific distribution of u_t, the process has to be a White Noise, hence with an expected value equal to zero. Therefore, the expectation of an MA(1) process is equal to μ, i.e. $\mathbb{E}\{y_t\} = \mu$. The variance instead is equal to

$$\gamma_0 = \mathbb{V}\{y_t\} = \sigma_u^2(1 + \theta_1^2).$$

In general, the auto covariance function for the order k is defined as

$$\gamma_k = \mathbb{C}v\{y_t, y_{t-k}\} = \begin{cases} \theta_1\sigma_u^2 & k \leq 1 \\ 0 & k > 1 \end{cases}.$$

It follows that, the auto covariance function is bounded up to the first lag, i.e.

$$\sum_{j=0}^{\infty} |\gamma_j| = \sigma_u^2(1 + \theta_1 + \theta_1^2),$$

and therefore the process is always stationary without requiring any condition on the parameter θ_1. Also the autocorrelation is different from zero only between the first two lags, i.e. the process is said to have a short memory

$$\rho_k = \mathbb{C}r\{y_t, y_{t-k}\} = \begin{cases} \frac{\theta_1}{1+\theta_1^2} & k \leq 1 \\ 0 & k > 1 \end{cases}.$$

Proof: Proposition 19.2

Proof. Let's consider an MA(1) process $y_t = \mu + \theta_1 u_{t-1} + u_t$, where u_t is a White Noise process (Equation 18.1). The expected value of y_t depends on the intercept μ, i.e.

$$\mathbb{E}\{y_t\} = \mu + \theta_1\mathbb{E}\{u_{t-1}\} + \mathbb{E}\{u_t\} = \mu \quad .$$

Under the White Noise assumption the residuals are uncorrelated, hence the variance is computed as

$$\begin{aligned} \gamma_0 &= \mathbb{V}\{\mu + u_t + \theta_1^2 u_{t-1}\} = \\ &= \mathbb{V}\{u_t + \theta_1^2 u_{t-1}\} = \\ &= \mathbb{V}\{u_t\} + \theta_1^2\mathbb{V}\{u_{t-1}\} = \\ &= \sigma_u^2 + \theta_1\sigma_u^2 = \\ &= \sigma_u^2(1 + \theta_1^2) \end{aligned}$$

By definition, the autocovariance function between time t and a generic lagged value $t-k$

reads

$$\begin{aligned}
\gamma_k &= \mathbb{C}v\{y_t, y_{t-k}\} \\
&= \mathbb{E}\{y_t y_{t-k}\} - \mathbb{E}\{y_t\}\mathbb{E}\{y_{t-k}\} = \\
&= \mathbb{E}\{(\mu + u_t + \theta_1 u_{t-1})(\mu + u_{t-k} + \theta_1 u_{t-k-1})\} - \mu^2 = \\
&= \mathbb{E}\{u_t u_{t-k}\} + \theta_1 \mathbb{E}\{u_{t-k} u_{t-1}\} + \theta_1 \mathbb{E}\{u_t u_{t-k-1}\} + \theta_1^2 \mathbb{E}\{u_{t-1} u_{t-k-1}\} + \\
&\quad + \mu^2 + \mu\mathbb{E}\{u_{t-k}\} + \mu\theta_1 \mathbb{E}\{u_{t-k-1}\} + \mu\mathbb{E}\{u_t\} + \mu\theta_1 \mathbb{E}\{u_{t-1}\} = \\
&= \theta_1 \mathbb{E}\{u_{t-k} u_{t-1}\} = \\
&= \begin{cases} \theta_1 \mathbb{V}\{u_{t-1}\} & k \leq 1 \\ 0 & k > 1 \end{cases}
\end{aligned}$$

This is a consequence of u_t being a White Noise and so uncorrelated in time, i.e. $\mathbb{E}\{u_t u_{t-k}\} = 0$ for every t. This implies that, also the correlation between two lags is zero if $k > 1$. □

Example: stationary MA(1)

Example 19.1. Under the assumption that the residuals are Gaussian, i.e. $u_t \sim \mathcal{N}(0, \sigma^2)$, we can simulate scenarios of a moving-average process of order 1 of the form

$$y_t \sim \text{MA}(1) \iff y_t = \mu + \theta_1 u_{t-1} + u_t. \tag{19.2}$$

1. Next step dynamics from Equation 19.2.

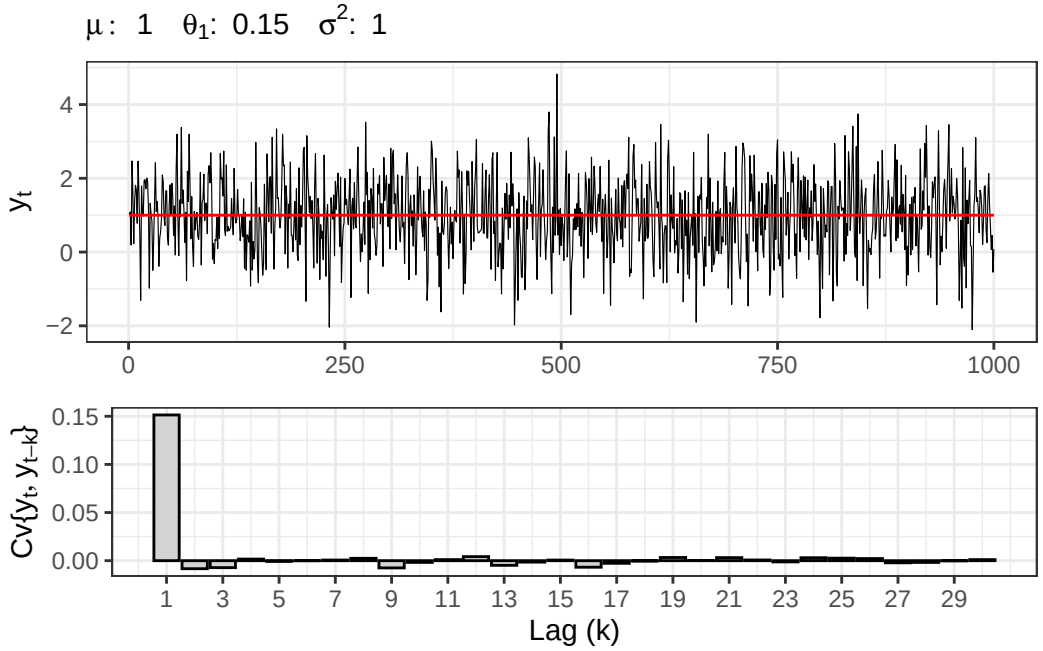


Figure 19.1: Simulation of an MA(1) process with long term expected value (red, top) and empiric autocovariance for the first 30 lag.(bottom).

Let's now compute the expectation, variance and covariance on simulated values and with the formulas.

Table 19.1: Empiric and theoric expectation, variance, covariance and correlation (first lag) for a stationary MA(1) process.

Statistic	Formula	Monte Carlo	|Error|
$\mathbb{E}\{y_t\}$	1.0000000	0.9974245	0.258%
$\mathbb{V}\{y_t\}$	1.0225000	1.0301274	-0.74%
$\mathbb{C}v\{y_t, y_{t-1}\}$	0.1500000	0.1513608	-0.899%
$\mathbb{C}r\{y_t, y_{t-1}\}$	0.1466993	0.1469331	-0.159%

19.2 AR(P)

In time series analysis an univariate time series y_t is defined as Autoregressive process of order p (AR(p)), when it satisfy the equations of the differences, i.e.

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + u_t, \tag{19.3}$$

where p defines the order of the process and $u_t \sim \text{WN}(0, \sigma_u^2)$. In compact form:

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + u_t.$$

An Autoregressive process can be equivalently expressed in terms of the polynomial operator, i.e.

$$\Phi(L) y_t = u_t, \quad \Phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p.$$

From Section 18.3.1 it follows that it exists a stationary AR(p) process if and only if all the solutions of the characteristic equations, i.e. $\Phi(z) = 0$, are greater than 1 in absolute value. In such case the AR(p) process admits an equivalent representation in terms of MA(∞), i.e.

$$\begin{aligned} y_t = \Phi(L)^{-1} u_t &= \frac{1}{1 - \phi_1 L - \cdots - \phi_p L^p} u_t \\ &= (1 + \psi_1 L + \psi_2 L^2 + \dots) = \sum_{i=1}^{\infty} \psi_i u_{t-i} \end{aligned}$$

19.2.1 Stationary AR(1)

Let's consider an AR(1) process, i.e.

$$y_t = \mu + \phi_1 y_{t-1} + u_t.$$

Through recursion up to time 0 it is possible to express an AR(1) model as an MA(∞), i.e.

$$y_t = \phi_1^t y_0 + \mu \sum_{i=0}^{t-1} \phi_1^i + \sum_{i=0}^{t-1} \phi_1^i u_{t-i},$$

where the process is stationary if and only if $|\phi_1| < 1$. In fact, independently from the specific distribution of the residuals u_t, the unconditional expectation of an AR(1) converges if and only if $|\phi_1| < 1$, i.e.

$$\mathbb{E}\{y_t\} = \phi_1^t y_0 + \mu \sum_{i=0}^{t-1} \phi_1^i = \frac{\mu}{1 - \phi_1}.$$

The variance instead is computed as:

$$\gamma_0 = \mathbb{V}\{y_t\} = \sum_{i=0}^{t-1} \phi_1^{2i} \sigma_u^2 = \frac{\sigma_u^2}{1 - \phi_1^2}.$$

The auto covariance decays exponentially fast depending on the parameter ϕ_1, i.e.

$$\gamma_1 = \mathbb{C}v\{y_t, y_{t-1}\} = \phi_1 \cdot \frac{\sigma_u^2}{1 - \phi_1^2} = \phi_1 \cdot \gamma_0,$$

where in general for the lag k

$$\gamma_k = \mathbb{C}v\{y_t, y_{t-k}\} = \phi_1^{|k|} \cdot \gamma_0.$$

Finally, the autocorrelation function

$$\rho_1 = \mathbb{C}r\{y_t, y_{t-1}\} = \frac{\gamma_1}{\gamma_0} = \phi_1,$$

where in general for the lag k

$$\rho_k = \mathbb{C}r\{y_t, y_{t-k}\} = \phi_1^{|k|}.$$

An example of a simulated AR(1) process ($\phi_1 = 0.95$, $\mu = 0.5$ and $\sigma_u^2 = 1$ and Normally distributed residuals) with its covariance function is shown in Figure 19.2.

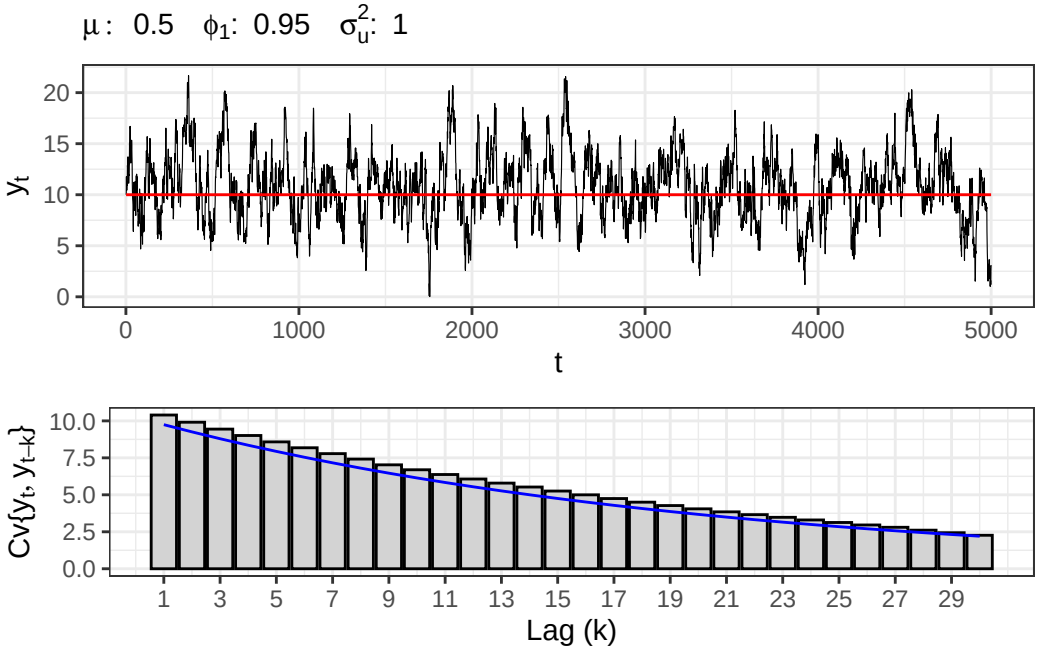


Figure 19.2: AR(1) simulation and expected value (red) on the top. Empirical autocovariance (gray) and fitted exponential decay (blue) at the bottom.

Example: sampling from a stationary AR(1)

Example 19.2. Sampling the process for different t we expect that, on a large number of simulations, the distribution will be normal with stationary moments, i.e. for all t

$$y_t \sim \mathcal{N}\left(\frac{\mu}{1-\phi_1}, \frac{\sigma^2}{1-\phi_1^2}\right).$$

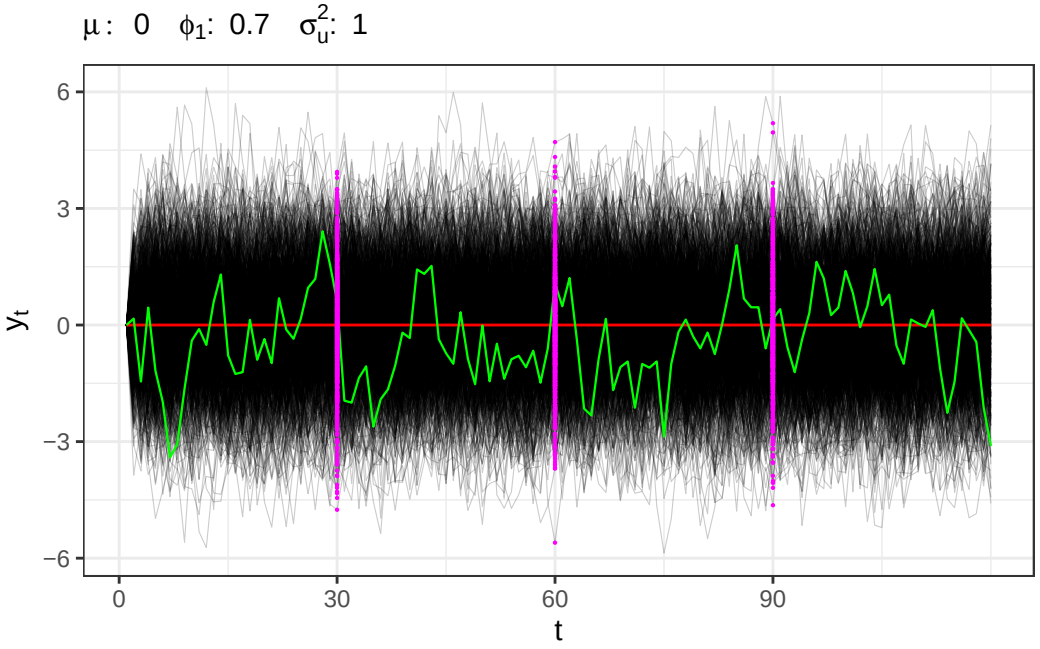


Figure 19.3: Stationary AR(1) simulation with expected value, one possible trajectory and samples at different times.

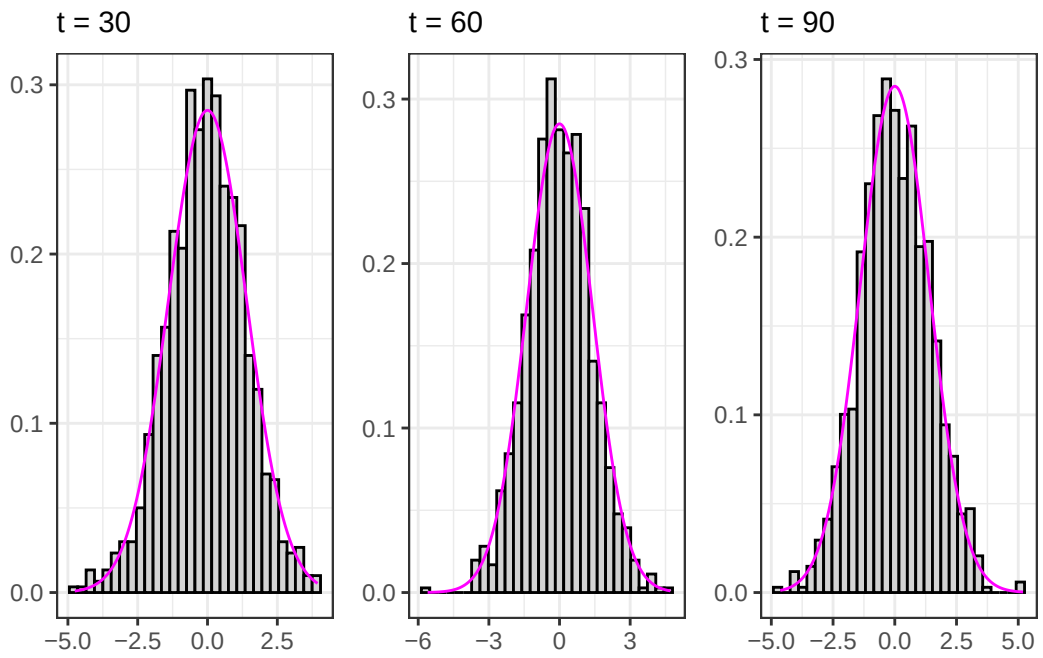


Figure 19.4: Stationary AR(1) histograms for different sampled times with normal pdf from empiric moments and normal pdf with theoric moments.

Table 19.2: Empiric and theoric expectation, variance, covariance and correlation (first lag) for a stationary AR(1) process.

Statistic	Theoric	Empiric
$\mathbb{E}\{y_t\}$	0.000000	-0.0027287
$\mathbb{V}\{y_t\}$	1.960784	1.9530614
$\mathbb{C}v\{y_t, y_{t-1}\}$	1.372549	1.3586670
$\mathbb{C}r\{y_t, y_{t-1}\}$	0.700000	0.6956660

19.2.2 Expectation

Proposition 19.3 (Expectation of an AR(p))**.** *The unconditional expected value of a stationary AR(p) process reads*

$$\mathbb{E}\{y_t\} = \frac{\mu}{1 - \sum_{i=1}^{p} \phi_i}$$

Proof: Proposition 19.3

Proof. Let's consider an AR(p) process y_t, then the unconditional expectation of the process is computed as

$$\begin{aligned}\mathbb{E}\{y_t\} &= \mathbb{E}\left\{\mu + \sum_{i=1}^{p}\phi_i y_{t-i} + u_t\right\} = \\ &= \mu + \sum_{i=1}^{q}\phi_i \mathbb{E}\{y_{t-i}\} + \mathbb{E}\{u_t\} = \\ &= \mu + \sum_{i=1}^{q}\phi_i \mathbb{E}\{y_t\}\end{aligned}$$

Since, under the assumption of stationarity the long term expectation of $\mathbb{E}\{y_{t-i}\}$ is the same as the long term expectation of $\mathbb{E}\{y_t\}$. Hence, solving for the expected value one obtain:

$$\mathbb{E}\{y_t\}\left(1 - \sum_{i=1}^{p}\phi_i\right) = \mu \implies \mathbb{E}\{y_t\} = \frac{\mu}{1 - \sum_{i=1}^{p}\phi_i}$$

□

19.2.3 Yule-Walker equations

If the AR(p) process is stationary, the covariance function satisfies the recursive relation, i.e.

$$\begin{cases}\gamma_k - \phi_1\gamma_{k-1} - \cdots - \phi_{k-p}\gamma_{k-p} = 0 \\ \vdots \\ \gamma_0 = \phi_1\gamma_1 + \cdots + \phi_p\gamma_p + \sigma_u^2\end{cases}$$

where $\gamma_{-k} = \gamma_k$. For $k = 0, \dots, p$ the above equations forms a system of $p+1$ linear equations in $p+1$ unknowns $\gamma_0, \dots, \gamma_p$, also known as **Yule-Walker equations**.

Proposition 19.4 (Variance for an AR(1))**.** *Let's consider an AR(1) model without intercept so that* $\mathbb{E}\{y_t\} = 0$*. Then, its covariance function reads:*

$$\gamma_k = \frac{\sigma^2}{1 - \phi_1^2}\phi_1^k = \gamma_0\phi_1^k$$

Proof: Proposition 19.4

Proof. Let's consider an AR(1) model without intercept so that $\mathbb{E}\{y_t\} = 0$, i.e.

$$y_t = \phi_1 y_{t-1} + u_t.$$

The proof of the covariance function is divided in two parts. Firstly, we compute the variance and covariances of the AR(1) model. Then, we set the system and we solve it. Notably, the variance of y_t reads:

$$\begin{aligned} \gamma_0 = \mathbb{V}\{y_t\} = \mathbb{E}\{y_t^2\} - \mathbb{E}\{y_t\}^2 = \mathbb{E}\{y_t^2\} = \\ = \mathbb{E}\{y_t(\phi_1 y_{t-1} + \varepsilon_t)\} = \\ = \phi_1 \mathbb{E}\{y_t y_{t-1}\} + \mathbb{E}\{y_t \varepsilon_t\} = \\ = \phi_1 \gamma_1 + \sigma^2 \end{aligned} \tag{19.4}$$

remembering that $\mathbb{E}\{y_t \varepsilon_t\} = \mathbb{E}\{\varepsilon_t^2\}$. The covariance with first lag, namely γ_1 is computed as:

$$\begin{aligned} \gamma_1 = \mathbb{C}v\{y_t, y_{t-1}\} = \mathbb{E}\{(\phi_1 y_{t-1} + \varepsilon_t) y_{t-1}\} = \\ = \phi_1 \mathbb{E}\{y_{t-1}^2\} = \phi_1 \gamma_0 \end{aligned} \tag{19.5}$$

The Yule-Walker system is given by Equation 19.4, Equation 19.5, i.e.

$$\begin{cases} \gamma_0 = \phi_1 \gamma_1 + \sigma^2 & (L_0) \\ \gamma_1 = \phi_1 \gamma_0 & (L_1) \end{cases}$$

In order to solve the system, let's substitute γ_1 (Equation 19.5) in γ_0 (Equation 19.4) and solve for γ_0, i.e.

$$\begin{cases} \gamma_0 = \phi_1^2 \gamma_0 + \sigma^2 = \frac{\sigma^2}{1-\phi_1^2} & (L_0) \\ \gamma_1 = \phi_1 \frac{\sigma^2}{1-\phi_1^2} & (L_1) \end{cases}$$

Hence, by the relation $\gamma_k = \phi_1 \gamma_{k-1}$ the covariance reads explicitely:

$$\gamma_k = \frac{\sigma^2}{1-\phi_1^2} \phi_1^k = \gamma_0 \phi_1^k$$

□

Proposition 19.5 (Variance for an AR(2))**.** *Let's consider an AR(2) model without intercept*

so that $\mathbb{E}\{y_t\} = 0$. *Then, its covariance function reads:*

$$\gamma_k = \begin{cases} \gamma_0 = \psi_0 \sigma^2 & k = 0 \\ \gamma_1 = \psi_1 \gamma_0 & k = 1 \\ \gamma_2 = \psi_2 \gamma_0 & k = 2 \\ \gamma_k = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2} & k \geq 3 \end{cases}$$

where

$$\begin{aligned} \psi_0 &= (1 - \phi_1 \psi_1 - \phi_2 \psi_2)^{-1} \\ \psi_1 &= \phi_1 (1 - \phi_2)^{-1} \\ \psi_2 &= \psi_1 \phi_1 + \phi_2 \end{aligned}$$

Proof: Proposition 19.5

Proof. Let's consider an AR(2) model without intercept so that $\mathbb{E}\{y_t\} = 0$, i.e.

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + u_t.$$

The proof of the covariance function is divided in two parts. Firstly, we compute the variance and covariances of the AR(2) model. Then, we set the system and we solve it. Notably, the variance of y_t reads:

$$\begin{aligned} \gamma_0 = \mathbb{V}\{y_t\} = \mathbb{E}\{y_t^2\} - \mathbb{E}\{y_t\}^2 = \mathbb{E}\{y_t^2\} = \\ = \mathbb{E}\{y_t(\phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t)\} = \\ = \phi_1 \mathbb{E}\{y_t y_{t-1}\} + \phi_2 \mathbb{E}\{y_t y_{t-2}\} + \mathbb{E}\{y_t \varepsilon_t\} = \\ = \phi_1 \gamma_1 + \phi_2 \gamma_2 + \sigma^2 \end{aligned} \tag{19.6}$$

remembering that $\mathbb{E}\{y_t \varepsilon_t\} = \mathbb{E}\{\varepsilon_t^2\}$. The covariance with first lag, namely γ_1 is computed as:

$$\begin{aligned} \gamma_1 = \mathbb{C}v\{y_t, y_{t-1}\} = \mathbb{E}\{(\phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t) y_{t-1}\} = \\ = \phi_1 \mathbb{E}\{y_{t-1}^2\} + \phi_2 \mathbb{E}\{y_{t-2} y_{t-1}\} = \\ = \phi_1 \gamma_0 + \phi_2 \gamma_1 \end{aligned} \tag{19.7}$$

The covariance with second lag, namely γ_2 is computed as:

$$\begin{aligned} \gamma_2 = \mathbb{C}v\{y_t, y_{t-1}\} = \mathbb{E}\{(\phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t) y_{t-2}\} = \\ = \phi_1 \mathbb{E}\{y_{t-1} y_{t-2}\} + \phi_2 \mathbb{E}\{y_{t-2}^2\} = \\ = \phi_1 \gamma_1 + \phi_2 \gamma_0 \end{aligned} \tag{19.8}$$

The Yule-Walker system is given by Equation 19.6, Equation 19.7 and Equation 19.8, i.e.

$$\begin{cases} \gamma_0 = \phi_1 \gamma_1 + \phi_2 \gamma_2 + \sigma^2 & (L_0) \\ \gamma_1 = \phi_1 \gamma_0 + \phi_2 \gamma_1 & (L_1) \\ \gamma_2 = \phi_1 \gamma_1 + \phi_2 \gamma_0 & (L_2) \end{cases}$$

In order to solve the system, let's start by solving for γ_1 (Equation 19.7) in terms of γ_0 in L_1, i.e.

$$\gamma_1 = \underbrace{\left(\frac{\phi_1}{1-\phi_2}\right)}_{\psi_1}\gamma_0 = \psi_1\gamma_0. \tag{19.9}$$

Substituting γ_1 from Equation 19.9 into γ_2 (Equation 19.8) in L_2 gives:

$$\gamma_2 = \underbrace{(\psi_1\phi_1 + \phi_2)}_{\psi_2}\gamma_0 = \psi_2\gamma_0. \tag{19.10}$$

Finally, substituting γ_1 (Equation 19.9) and γ_2 (Equation 19.10) into γ_0 (Equation 19.6) in L_0 gives an explicit expression of the variance of y_t, i.e.

$$\gamma_0 = \underbrace{\frac{1}{1-\phi_1\psi_1-\phi_2\psi_2}}_{\psi_0}\sigma^2 = \psi_0\sigma^2.$$

Table 19.3: Theoric long term variance and variance computed on 500 Monte Carlo simulations (t = 100000).

Covariance	Formula	MonteCarlo
$\mathbb{V}\{y_t\}$	1.1363636	1.1362110
$\mathbb{C}v\{y_t, y_{t-1}\}$	0.3787879	0.3785980
$\mathbb{C}v\{y_t, y_{t-2}\}$	0.2272727	0.2271252
$\mathbb{C}v\{y_t, y_{t-3}\}$	0.1060606	0.1060733
$\mathbb{C}v\{y_t, y_{t-4}\}$	0.0545455	0.0544696
$\mathbb{C}v\{y_t, y_{t-5}\}$	0.0269697	0.0268185

□

Proposition 19.6 (Variance for an AR(3))**.** *Let's consider an AR(3) model without intercept so that* $\mathbb{E}\{y_t\} = 0$*. Then, its covariance function reads:*

$$\gamma_k = \begin{cases} \gamma_0 = \psi_0\sigma^2 & k = 0 \\ \gamma_1 = \psi_1\gamma_0 & k = 1 \\ \gamma_2 = \psi_2\gamma_0 & k = 2 \\ \gamma_3 = \psi_3\gamma_0 & k = 3 \\ \gamma_k = \phi_1\gamma_{k-1} + \phi_2\gamma_{k-2} + \phi_3\gamma_{k-3} & k \geq 4 \end{cases}$$

where

$$\begin{aligned}\psi_0 &= (1-\phi_1\psi_1-\phi_2\psi_2-\phi_3\psi_3)^{-1}\\ \psi_1 &= (\phi_1+\phi_2\phi_3)(1-\phi_2-\phi_3^2-\phi_1\phi_3)^{-1}\\ \psi_2 &= \phi_1\psi_1+\phi_3\psi_1+\phi_2\\ \psi_3 &= \phi_1\psi_2+\phi_2\psi_1+\phi_3\end{aligned}$$

Proof: Proposition 19.6

Proof. Let's consider an AR(3) model without intercept so that $\mathbb{E}\{y_t\}=0$, i.e.

$$y_t = \phi_1 y_{t-1}+\phi_2 y_{t-2}+\phi_3 y_{t-3}+u_t$$

where $\varepsilon_t \sim \text{WN}(0,\sigma^2)$. Notably, the variance is computed as:

$$\begin{aligned}\gamma_0 = \mathbb{V}\{y_t\} &= \mathbb{E}\{y_t^2\}-\mathbb{E}\{y_t\}^2 =\\ &= \mathbb{E}\{y_t^2\} =\\ &= \mathbb{E}\{y_t(\phi_1 y_{t-1}+\phi_2 y_{t-2}+\phi_3 y_{t-3}+\varepsilon_t)\} =\\ &= \phi_1\mathbb{E}\{y_t y_{t-1}\}+\phi_2\mathbb{E}\{y_t y_{t-2}\}+\phi_3\mathbb{E}\{y_t y_{t-3}\}+\mathbb{E}\{y_t\varepsilon_t\} =\\ &= \phi_1\gamma_1+\phi_2\gamma_2+\phi_3\gamma_3+\sigma^2\end{aligned} \tag{19.11}$$

remembering that $\mathbb{E}\{y_t\varepsilon_t\}=\mathbb{E}\{\varepsilon_t^2\}$. The covariance with first lag, namely γ_1 is computed as:

$$\begin{aligned}\gamma_1 = \mathbb{C}v\{y_t,y_{t-1}\} &= \mathbb{E}\{(\phi_1 y_{t-1}+\phi_2 y_{t-2}+\phi_3 y_{t-3}+\varepsilon_t)y_{t-1}\} =\\ &= \phi_1\mathbb{E}\{y_{t-1}^2\}+\phi_2\mathbb{E}\{y_{t-2}y_{t-1}\}+\phi_3\mathbb{E}\{y_{t-3}y_{t-1}\} =\\ &= \phi_1\gamma_0+\phi_2\gamma_1+\phi_2\gamma_2\end{aligned} \tag{19.12}$$

The covariance with second lag, namely γ_2 is computed as:

$$\begin{aligned}\gamma_2 = \mathbb{C}v\{y_t,y_{t-2}\} &= \mathbb{E}\{(\phi_1 y_{t-1}+\phi_2 y_{t-2}+\phi_3 y_{t-3}+\varepsilon_t)y_{t-2}\} =\\ &= \phi_1\mathbb{E}\{y_{t-1}y_{t-2}\}+\phi_2\mathbb{E}\{y_{t-2}^2\}+\phi_3\mathbb{E}\{y_{t-3}y_{t-2}\} =\\ &= \phi_1\gamma_1+\phi_2\gamma_0+\phi_3\gamma_1\end{aligned} \tag{19.13}$$

The covariance with third lag, namely γ_3 is computed as:

$$\begin{aligned}\gamma_3 = \mathbb{C}v\{y_t,y_{t-3}\} &= \mathbb{E}\{(\phi_1 y_{t-1}+\phi_2 y_{t-2}+\phi_3 y_{t-3}+\varepsilon_t)y_{t-3}\} =\\ &= \phi_1\mathbb{E}\{y_{t-1}y_{t-3}\}+\phi_2\mathbb{E}\{y_{t-2}y_{t-3}\}+\phi_3\mathbb{E}\{y_{t-3}^2\} =\\ &= \phi_1\gamma_2+\phi_2\gamma_1+\phi_3\gamma_0\end{aligned} \tag{19.14}$$

The Yule-Walker system is given by Equation 19.11, Equation 19.12, Equation 19.13 and Equation 19.14 reads

$$\begin{cases}\gamma_0 = \phi_1\gamma_1+\phi_2\gamma_2+\phi_3\gamma_3+\sigma^2 & (L_0)\\ \gamma_1 = \phi_1\gamma_0+\phi_2\gamma_1+\phi_3\gamma_2 & (L_1)\\ \gamma_2 = \phi_1\gamma_1+\phi_2\gamma_0+\phi_3\gamma_1 & (L_2)\\ \gamma_3 = \phi_1\gamma_2+\phi_2\gamma_1+\phi_3\gamma_0 & (L_3)\end{cases}$$

Let's start by expressing γ_2 (Equation 19.13) in terms of γ_1 and γ_0 from L_2, i.e.

$$\gamma_2 = (\phi_1 + \phi_3)\gamma_1 + \phi_2\gamma_0 \tag{19.15}$$

Then, let's substitute the above expression of γ_2 (Equation 19.15) in γ_1 (Equation 19.12) from L_1, i.e.

$$\gamma_1 = \phi_1\gamma_0 + \phi_2\gamma_1 + \phi_3(\phi_1 + \phi_3)\gamma_1 + \phi_3\phi_2\gamma_0$$

At this point γ_1 depends only on γ_0, hence we can solve it:

$$\gamma_1 = \underbrace{\frac{\phi_1 + \phi_2\phi_3}{1 - \phi_2 - \phi_3^2 - \phi_1\phi_3}}_{\psi_1}\gamma_0 = \psi_1\gamma_0 \tag{19.16}$$

With γ_1 solved, one can come back to the expression of γ_2 (Equation 19.15) and substitute the result in γ_1 (Equation 19.16) obtaining an explicit expression for γ_2, i.e.

$$\gamma_2 = \underbrace{(\phi_1\psi_1 + \phi_3\psi_1 + \phi_2)}_{\psi_2}\gamma_0 = \psi_2\gamma_0 \tag{19.17}$$

Substituting the explicit expressions of γ_2 (Equation 19.17) and γ_1 (Equation 19.16) into γ_3 (Equation 19.14) completes the system, i.e.

$$\gamma_3 = \underbrace{(\phi_1\psi_2 + \phi_2\psi_1 + \phi_3)}_{\psi_3}\gamma_0 = \psi_3\gamma_0 \tag{19.18}$$

Finally, substituting γ_1 (Equation 19.16), γ_2 (Equation 19.17) and γ_3 (Equation 19.18) in γ_0 (Equation 19.11) gives the variance, i.e.

$$\gamma_0 = \underbrace{\frac{1}{1 - \phi_1\psi_1 - \phi_2\psi_2 - \phi_3\psi_3}}_{\psi_0}\sigma^2 \tag{19.19}$$

$$\begin{aligned}
\psi_0 &= (1 - \phi_1\psi_1 - \phi_2\psi_2 - \phi_3\psi_3)^{-1} \\
\psi_1 &= (\phi_1 + \phi_2\phi_3)(1 - \phi_2 - \phi_3^2 - \phi_1\phi_3)^{-1} \\
\psi_2 &= \phi_1\psi_1 + \phi_3\psi_1 + \phi_2 \\
\psi_3 &= \phi_1\psi_2 + \phi_2\psi_1 + \phi_3
\end{aligned}$$

Table 19.4: Theoric long term variance and variance computed on 500 Monte Carlo simulations (t = 100000).

Covariance	Formula	MonteCarlo
$\mathbb{V}\{y_t\}$	1.1538304	1.1535204

$\mathbb{C}v\{y_t, y_{t-1}\}$	0.3987743	0.3985179
$\mathbb{C}v\{y_t, y_{t-2}\}$	0.2549540	0.2548731
$\mathbb{C}v\{y_t, y_{t-3}\}$	0.1740552	0.1738990
$\mathbb{C}v\{y_t, y_{t-4}\}$	0.0976507	0.0975948
$\mathbb{C}v\{y_t, y_{t-5}\}$	0.0594484	0.0596297

□

Proposition 19.7 (Variance for an AR(4)). *Let's consider an AR(4) model without intercept so that* $\mathbb{E}\{y_t\} = 0$. *Then, its covariance function reads:*

$$\gamma_k = \begin{cases} \gamma_0 = \psi_0 \sigma^2 & k = 0 \\ \gamma_1 = \psi_1 \gamma_0 & k = 1 \\ \gamma_2 = \psi_2 \gamma_0 & k = 2 \\ \gamma_3 = \psi_3 \gamma_0 & k = 3 \\ \gamma_4 = \psi_4 \gamma_0 & k = 4 \\ \gamma_k = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2} + \phi_3 \gamma_{k-3} + \phi_4 \gamma_{k-4} & k \geq 5 \end{cases}$$

where

$$\psi_1 = \left(\frac{\phi_3 \phi_2}{1-\phi_4} + \frac{\phi_1 \phi_4 \phi_2}{1-\phi_4} + \phi_1 + \phi_3 \phi_4 \right) \left(1 - \frac{\phi_3(\phi_1+\phi_3)}{1-\phi_4} - \frac{\phi_1 \phi_4 (\phi_1+\phi_3)}{1-\phi_4} - \phi_2 - \phi_2 \phi_4 - \phi_4^2 \right)^{-1}$$
$$\psi_2 = \frac{\psi_1(\phi_1+\phi_3)}{1-\phi_4} + \frac{\phi_2}{1-\phi_4}$$
$$\psi_3 = \phi_1 \psi_2 + \phi_2 \psi_1 + \phi_3 + \phi_4 \psi_1$$
$$\psi_4 = \phi_1 \psi_3 + \phi_2 \psi_2 + \phi_3 \psi_1 + \phi_4$$
$$\psi_0 = (1 - \phi_1 \psi_1 - \phi_2 \psi_2 - \phi_3 \psi_3 - \phi_4 \psi_4)^{-1}$$

i Proof: Proposition 19.7

Proof. Let's consider an AR(4) model without intercept so that $\mathbb{E}\{y_t\} = 0$, i.e.

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \phi_4 y_{t-4} + \varepsilon_t$$

where $\varepsilon_t \sim \text{WN}(0, \sigma^2)$. Notably, the variance is computed as:

$$\begin{aligned}\gamma_0 = \mathbb{V}\{y_t\} &= \mathbb{E}\{y_t^2\} - \mathbb{E}\{y_t\}^2 = \\ &= \mathbb{E}\{y_t^2\} = \\ &= \mathbb{E}\{y_t(\phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \phi_4 y_{t-4} + \varepsilon_t)\} = \\ &= \phi_1 \mathbb{E}\{y_t y_{t-1}\} + \phi_2 \mathbb{E}\{y_t y_{t-2}\} + \phi_3 \mathbb{E}\{y_t y_{t-3}\} + \phi_4 \mathbb{E}\{y_t y_{t-4}\} + \mathbb{E}\{y_t \varepsilon_t\} = \\ &= \phi_1 \gamma_1 + \phi_2 \gamma_2 + \phi_3 \gamma_3 + \phi_4 \gamma_4 + \sigma^2\end{aligned} \tag{19.20}$$

remembering that $\mathbb{E}\{y_t \varepsilon_t\} = \mathbb{E}\{\varepsilon_t^2\}$. The covariance with first lag, namely γ_1 is computed as:

$$\begin{aligned}\gamma_1 = \mathbb{C}v\{y_t, y_{t-1}\} &= \mathbb{E}\{(\phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \phi_4 y_{t-4} + \varepsilon_t) y_{t-1}\} = \\ &= \phi_1 \mathbb{E}\{y_{t-1}^2\} + \phi_2 \mathbb{E}\{y_{t-2} y_{t-1}\} + \phi_3 \mathbb{E}\{y_{t-3} y_{t-1}\} + \phi_4 \mathbb{E}\{y_{t-4} y_{t-1}\} = \\ &= \phi_1 \gamma_0 + \phi_2 \gamma_1 + \phi_2 \gamma_2 + \phi_3 \gamma_3\end{aligned} \tag{19.21}$$

The covariance with second lag, namely γ_2 is computed as:

$$\begin{aligned}\gamma_2 = \mathbb{C}v\{y_t, y_{t-2}\} &= \mathbb{E}\{(\phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \phi_4 y_{t-4} + \varepsilon_t) y_{t-2}\} = \\ &= \phi_1 \mathbb{E}\{y_{t-1} y_{t-2}\} + \phi_2 \mathbb{E}\{y_{t-2}^2\} + \phi_3 \mathbb{E}\{y_{t-3} y_{t-2}\} + \phi_4 \mathbb{E}\{y_{t-4} y_{t-2}\} = \\ &= \phi_1 \gamma_1 + \phi_2 \gamma_0 + \phi_3 \gamma_1 + \phi_4 \gamma_2\end{aligned} \tag{19.22}$$

The covariance with third lag, namely γ_3 is computed as:

$$\begin{aligned}\gamma_3 = \mathbb{C}v\{y_t, y_{t-3}\} &= \mathbb{E}\{(\phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \phi_4 y_{t-4} + \varepsilon_t) y_{t-3}\} = \\ &= \phi_1 \mathbb{E}\{y_{t-1} y_{t-3}\} + \phi_2 \mathbb{E}\{y_{t-2} y_{t-3}\} + \phi_3 \mathbb{E}\{y_{t-3}^2\} + \phi_4 \mathbb{E}\{y_{t-4} y_{t-3}\} = \\ &= \phi_1 \gamma_2 + \phi_2 \gamma_1 + \phi_3 \gamma_0 + \phi_4 \gamma_1\end{aligned} \tag{19.23}$$

The covariance with fourth lag, namely γ_4 is computed as:

$$\begin{aligned}\gamma_4 = \mathbb{C}v\{y_t, y_{t-4}\} &= \mathbb{E}\{(\phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \phi_4 y_{t-4} + \varepsilon_t) y_{t-4}\} = \\ &= \phi_1 \mathbb{E}\{y_{t-1} y_{t-4}\} + \phi_2 \mathbb{E}\{y_{t-2} y_{t-4}\} + \phi_3 \mathbb{E}\{y_{t-3} y_{t-4}\} + \phi_4 \mathbb{E}\{y_{t-4}^4\} = \\ &= \phi_1 \gamma_3 + \phi_2 \gamma_2 + \phi_3 \gamma_1 + \phi_4 \gamma_0\end{aligned} \tag{19.24}$$

The Yule-Walker system is given by Equation 19.20, Equation 19.21 and Equation 19.22, Equation 19.23, Equation 19.24 reads

$$\begin{cases} \gamma_0 = \phi_1 \gamma_1 + \phi_2 \gamma_2 + \phi_3 \gamma_3 + \phi_4 \gamma_4 + \sigma^2 & (L_0) \\ \gamma_1 = \phi_1 \gamma_0 + \phi_2 \gamma_1 + \phi_3 \gamma_2 + \phi_4 \gamma_3 & (L_1) \\ \gamma_2 = \phi_1 \gamma_1 + \phi_2 \gamma_0 + \phi_3 \gamma_1 + \phi_4 \gamma_2 & (L_2) \\ \gamma_3 = \phi_1 \gamma_2 + \phi_2 \gamma_1 + \phi_3 \gamma_0 + \phi_4 \gamma_1 & (L_3) \\ \gamma_4 = \phi_1 \gamma_3 + \phi_2 \gamma_2 + \phi_3 \gamma_1 + \phi_4 \gamma_0 & (L_4) \end{cases}$$

To solve the system, let's substitue γ_1 (Equation 19.21) into γ_2 (Equation 19.22) and recover γ_2 in terms of γ_0, i.e.

$$\gamma_2 = \frac{\phi_1 + \phi_3}{1 - \phi_4}\gamma_1 + \frac{\phi_2}{1 - \phi_4}\gamma_0 \tag{19.25}$$

Then, substitue γ_3 (Equation 19.23) into γ_1 (Equation 19.21), i.e.

$$\gamma_1 = (\phi_3 + \phi_1\phi_4)\gamma_2 + (\phi_2 + \phi_2\phi_4 + \phi_4^2)\gamma_1 + (\phi_1 + \phi_3\phi_4)\gamma_0 \tag{19.26}$$

Then, substitue γ_2 (Equation 19.25) into the previous expression for γ_1 (Equation 19.26), i.e.

$$\begin{aligned}\gamma_1 = &\left(\frac{\phi_3(\phi_1 + \phi_3)}{1 - \phi_4} + \frac{\phi_1\phi_4(\phi_1 + \phi_3)}{1 - \phi_4} + \phi_2 + \phi_2\phi_4 + \phi_4^2\right)\gamma_1 + \\ &+ \left(\frac{\phi_3\phi_2}{1 - \phi_4} + \frac{\phi_1\phi_4\phi_2}{1 - \phi_4} + \phi_1 + \phi_3\phi_4\right)\gamma_0\end{aligned}$$

Hence, recovering γ_1 one obtain:

$$\gamma_1 = \underbrace{\left(\frac{\frac{\phi_3\phi_2}{1-\phi_4} + \frac{\phi_1\phi_4\phi_2}{1-\phi_4} + \phi_1 + \phi_3\phi_4}{1 - \frac{\phi_3(\phi_1+\phi_3)}{1-\phi_4} - \frac{\phi_1\phi_4(\phi_1+\phi_3)}{1-\phi_4} - \phi_2 - \phi_2\phi_4 - \phi_4^2}\right)}_{\psi_1}\gamma_0 = \psi_1\gamma_0 \tag{19.27}$$

Substituing γ_1 (Equation 19.27) into γ_2 (Equation 19.25) gives

$$\gamma_2 = \underbrace{\left(\frac{\psi_1(\phi_1 + \phi_3)}{1 - \phi_4} + \frac{\phi_2}{1 - \phi_4}\right)}_{\psi_2}\gamma_0 = \psi_2\gamma_0 \tag{19.28}$$

Then, let's rewrite γ_3 (Equation 19.23) substituing γ_1 (Equation 19.27) and γ_2 (Equation 19.28), i.e.

$$\gamma_3 = \underbrace{(\phi_1\psi_2 + \phi_2\psi_1 + \phi_3 + \phi_4\psi_1)}_{\psi_3}\gamma_0 = \psi_3\gamma_0 \tag{19.29}$$

Finally, substituing γ_1 (Equation 19.27), γ_2 (Equation 19.28) and γ_3 (Equation 19.29) into γ_4 (Equation 19.24) gives:

$$\gamma_4 = \underbrace{(\phi_1\psi_3 + \phi_2\psi_2 + \phi_3\psi_1 + \phi_4)}_{\psi_4}\gamma_0 = \psi_4\gamma_0 \tag{19.30}$$

Finally, substituing γ_1 (Equation 19.27), γ_2 (Equation 19.28) and γ_3 (Equation 19.29) and γ_4 (Equation 19.30) into γ_0 (Equation 19.20) gives the variance, i.e.

$$\gamma_0 = \underbrace{\frac{1}{1 - \phi_1\psi_1 - \phi_2\psi_2 - \phi_3\psi_3 - \phi_4\psi_4}}_{\psi_0}\sigma^2 \tag{19.31}$$

Table 19.5: Theoric long term variance and variance computed on 500 Monte Carlo simulations (t = 100000).

Covariance	Formula	MonteCarlo
$\mathbb{V}\{y_t\}$	1.2543907	1.2509545
$\mathbb{C}v\{y_t, y_{t-1}\}$	0.5084245	0.5003887
$\mathbb{C}v\{y_t, y_{t-2}\}$	0.4036375	0.3997435
$\mathbb{C}v\{y_t, y_{t-3}\}$	0.3380467	0.3351931
$\mathbb{C}v\{y_t, y_{t-4}\}$	0.2504338	0.2481127
$\mathbb{C}v\{y_t, y_{t-5}\}$	0.1814536	0.1797885

□

19.2.4 Non-stationary AR(1): random walk

A non-stationary process has expectation and/or variance that changes over time. Considering the setup of an AR(1), if $\phi_1 = 1$ the process degenerates into a so called **random walk** process. Formally, if $\mu \neq 0$ it is called **random walk with drift**, i.e.

$$y_t = \mu + y_{t-1} + u_t.$$

Considering its MA(∞) representation

$$y_t = y_0 + \mu \cdot t + \sum_{i=0}^{t-1} u_{t-i},$$

it is easy to see that the expectation depends on the starting point and on time t and the shocks u_{t-i} never decays. In fact, computing the expectation and variance of a random walk process it emerges a clear dependence on time, i.e.

$$\mathbb{E}\{y_t \mid \mathcal{F}_0\} = y_0 + \mu t \qquad \mathbb{V}\{y_t \mid \mathcal{F}_0\} = t\sigma_u^2$$
$$\mathbb{C}v\{y_t, y_{t-k}\} = (t-k) \cdot \sigma^2 \quad \mathbb{C}r\{y_t, y_{t-k}\} = \sqrt{\frac{t-k}{t}},$$

and the variance tends to explode to ∞ as $t \to \infty$.

⚠ Stochastic trend of a Random walk

Let's define the **stochastic trend** S_t as $S_t = \sum_{i=0}^{t-1} u_{t-i}$, then

1. The expectation of S_t if u_t are all martingale difference sequences, is zero, i.e.

$$\mathbb{E}\{S_t\} = \sum_{i=0}^{t-1} \mathbb{E}\{u_{t-i}\} = \sum_{i=0}^{t-1} \mathbb{E}\{\mathbb{E}\{u_{t-i} \mid \mathcal{F}_{t-1}\}\} = 0,$$

and therefore

$$\mathbb{E}\{y_t\} = y_0 + \mu t + \mathbb{E}\{S_t\} = y_0 + \mu t.$$

2. The variance of S_t, if u_t are all martingale difference sequences, is time-dependent, i.e.

$$\mathbb{V}\{y_t\} = \mathbb{V}\{S_t\} = \sum_{i=0}^{t-1} \mathbb{V}\{u_{t-i}\} = \sum_{i=0}^{t-1} \sigma^2 = t \cdot \sigma^2.$$

while the covariance between two times t and $t-k$ depends on the lag, i.e.

$$\mathbb{C}v\{S_t, S_{t-k}\} = \mathbb{E}\{S_t S_{t-k}\} = \sum_{i=0}^{t-k-1} \mathbb{V}\{u_{t-i}\} = (t-k) \cdot \sigma^2$$

and so the correlation

$$\mathbb{C}r\{S_t, S_{t-k}\} = \frac{\mathbb{C}v\{S_t, S_{t-k}\}}{\sqrt{\mathbb{V}\{S_t\} \cdot \mathbb{V}\{S_{t-k}\}}} = \frac{(t-k) \cdot \sigma^2}{\sqrt{t(t-k) \cdot \sigma^2}} = \frac{(t-k)}{\sqrt{t(t-k)}}$$

tends to one as $t \to \infty$, in fact:

$$\lim_{t\to\infty} \mathbb{C}r\{S_t, S_{t-k}\} = \lim_{t\to\infty} \frac{(t-k)}{\sqrt{t(t-k)}} = \lim_{t\to\infty} \sqrt{\frac{t-k}{t}} = \lim_{t\to\infty} \sqrt{1 - \frac{k}{t}} = 1$$

Example: sampling from non-stationary AR(1)

Example 19.3. Let's simulate an random walk process with drift with a Gaussian error, namely $u_t \sim \mathcal{N}(0, \sigma_u^2)$.

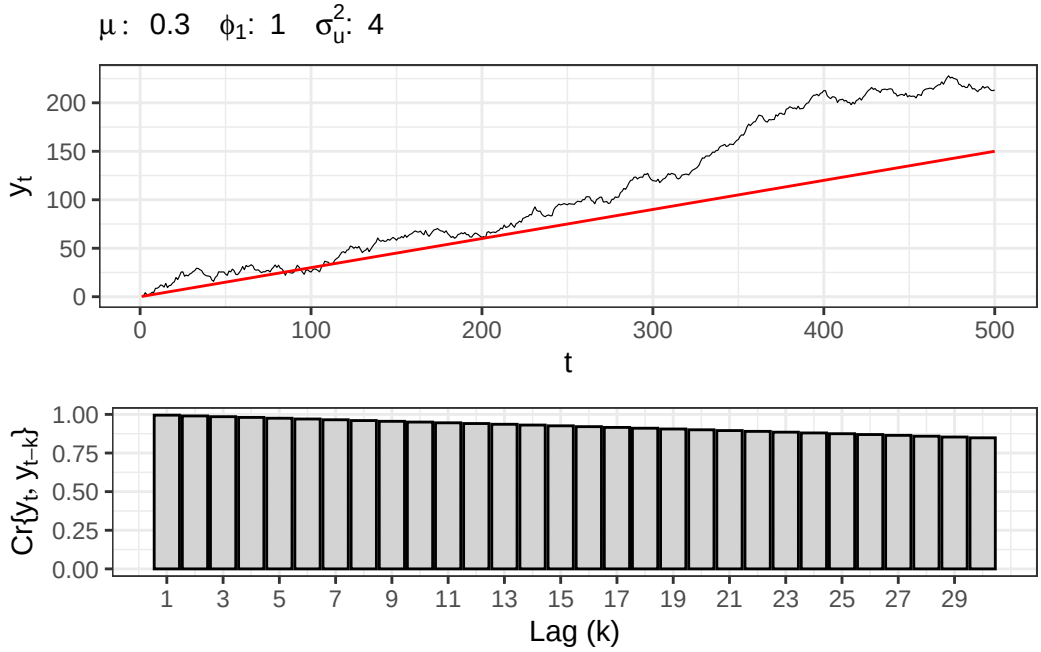


Figure 19.5: Random walk simulation and expected value (red) on the top. Empirical autocorrelation (gray).

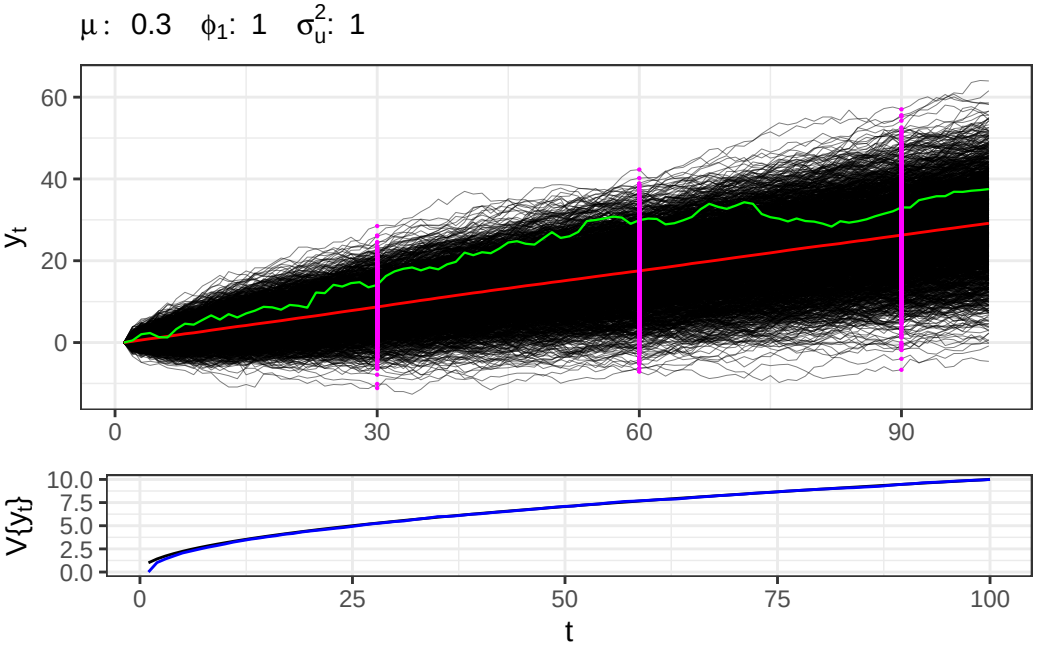


Figure 19.6: Non stationary AR(1) simulation with expected value (red), a possible trajectory (green) and samples for different times (magenta) on the top. Theoretic (blue) and empiric (black) std. deviation at the bottom.

Sampling the process for different t we expect that, on a large number of simulations, the distribution will be still normal but with non-stationary moments, i.e.

$$X_t \sim \mathcal{N}\left(\mu t, \sigma_u^2 t\right).$$

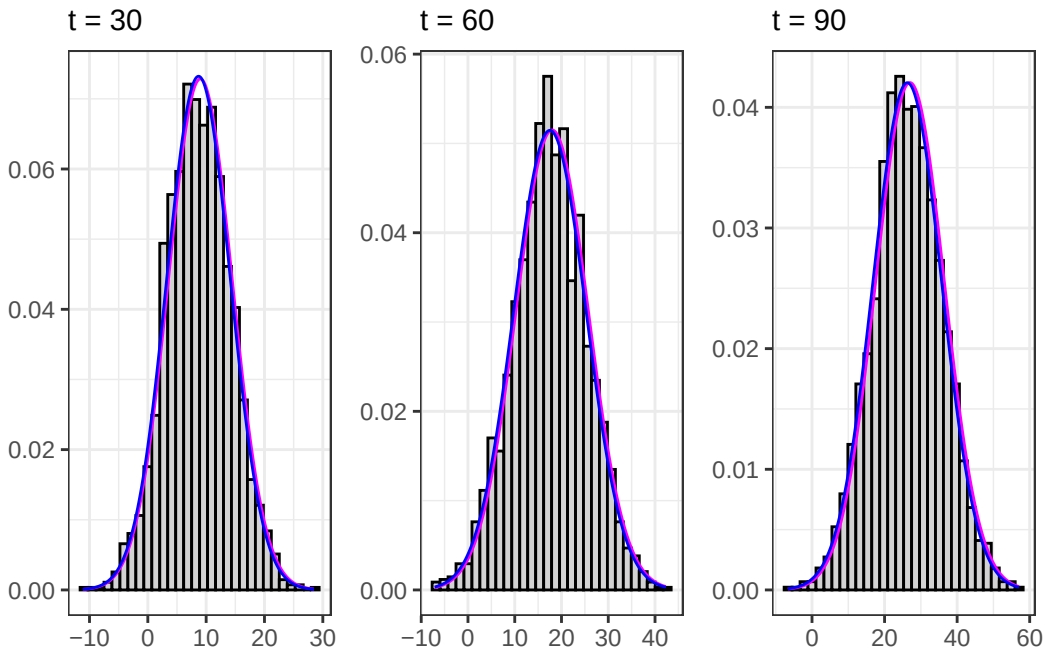


Figure 19.7: Non-stationary AR(1) histograms for different sampled times with normal pdf with empiric moments (blue) and normal pdf with theoretic moments (magenta).

Table 19.6: Empiric and theoric expectation, variance, covariance and correlation (first lag) for a stationary AR(1) process.

t	$\mathbb{E}\{y_t\}$	$\mathbb{E}^{mc}\{y_t\}$	$\mathbb{V}\{y_t\}$	$\mathbb{V}^{mc}\{y_t\}$
30	8.7	8.698311	29	29.69114
60	17.7	17.517805	59	60.09756
90	26.7	26.268452	89	90.05715

20 ARMA processes

20.1 ARMA(p, q)

An Auto Regressive Moving Average processes, for simplicity ARMA(p,q), is a combination of an MA(q) and an AR(p) processes, i.e. an $x_t \sim \text{ARMA}(p, q)$ is formally defined as

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j u_{t-j} + u_t,$$

where in general $u_t \sim \text{WN}(0, \sigma_u^2)$. An example of a simulated ARMA(1,1) process ($\phi_1 = 0.95$, $\theta_1 = 0.45$ $\mu = 0.5$ and $\sigma_u^2 = 1$ and Normally distributed residuals) with its covariance function is shown in Figure 20.1.

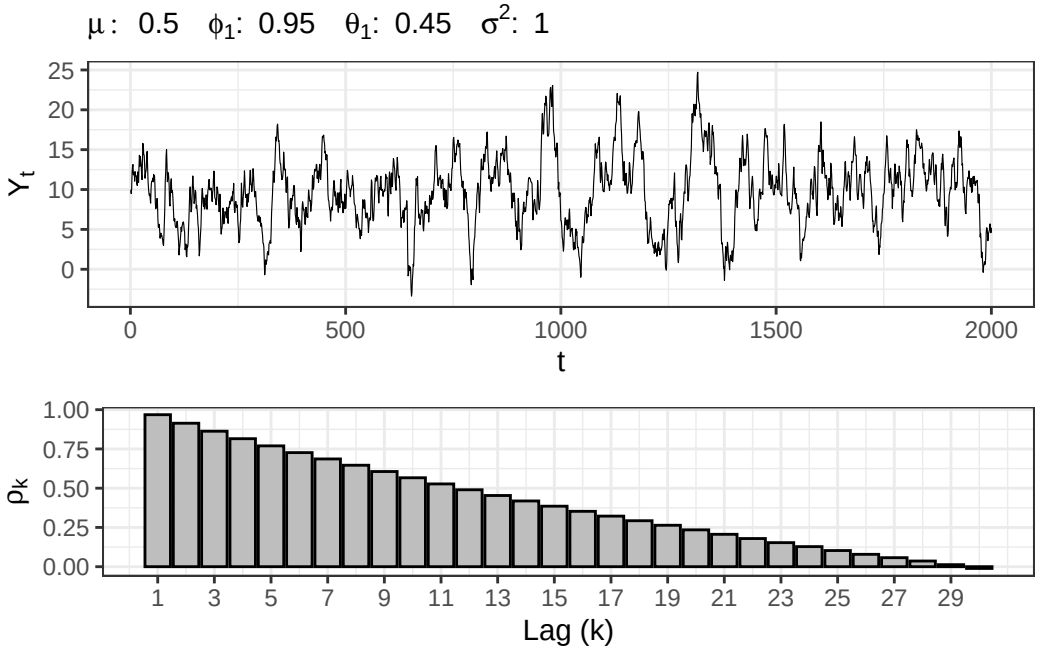


Figure 20.1: ARMA(1,1) simulation on the top and empirical autocorrelation at the bottom.

20.1.1 Matrix form AR(p)

An Autoregressive process of order p AR(p) (Equation 19.3) can be written in matrix form as

$$\mathbf{X}_t = \mathbf{c} + \mathbf{X}_{t-1} + \mathbf{b}\, u_t,$$

where

$$\underbrace{\begin{pmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p} \end{pmatrix}}_{\mathbf{X}_t} = \underbrace{\begin{pmatrix} \mu \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{\mathbf{c}} + \underbrace{\begin{pmatrix} \phi_1 & \phi_2 & \dots & \phi_{p-1} & \phi_p \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix}}_{} \underbrace{\begin{pmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p-1} \end{pmatrix}}_{\mathbf{X}_{t-1}} + \underbrace{\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{\mathbf{b}} \cdot u_t$$

where $\mathbf{X}_t$, $\mathbf{c}$ and $\mathbf{b}$ have dimension $p \times 1$, while is $p \times p$.

> **How to construct the companion matrix ?**
>
> Let's consider the vector containing the coefficients of the model, i.e.
>
> $$\underset{p\times 1}{\gamma} = \begin{pmatrix} \phi_1 & \phi_2 & \dots & \phi_{p-1} & \phi_p \end{pmatrix}.$$
>
> If the order of the Autoregressive process is greater than 1, i.e. $p > 1$, let's consider an identity matrix (Equation 32.3) with dimension $(p-1) \times (p-1)$
>
> $$\underset{(p-1)\times(p-1)}{\mathbf{I}_{p-1}} = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix},$$
>
> then, we combine it with a column of zeros, i.e.
>
> $$\underset{(p-1)\times p}{\mathbf{L}_{p-1}} = \begin{pmatrix} \mathbf{I}_p & \mathbf{0}_{1\times p} \end{pmatrix}.$$
>
> Finally, we combine $\mathbf{L}$ with the AR parameters, i.e.
>
> $$= \begin{pmatrix} \gamma \\ \mathbf{L}_{p-1} \end{pmatrix}.$$

Example: AR(4) in matrix form

Example 20.1. Let's consider for example an AR(4) process. The vector containing the coefficients of the model reads

$$\underset{4\times 1}{\gamma} = \begin{pmatrix} \phi_1 & \phi_2 & \phi_3 & \phi_4 \end{pmatrix}.$$

For an AR(4) we consider a diagonal matrix with dimension 3×3, i.e.

$$\underset{3\times 3}{\mathbf{I}_3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

then, we add a column of zeros, i.e.

$$\underset{3\times 4}{\mathbf{L}_3} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Finally we combine

$$= \begin{pmatrix} \gamma \\ \mathbf{L}_3 \end{pmatrix} = \begin{pmatrix} \phi_1 & \phi_2 & \phi_3 & \phi_4 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Then, let's consider the iteration

$$\begin{pmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ y_{t-3} \end{pmatrix} = \begin{pmatrix} \mu \\ 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \phi_1 & \phi_2 & \phi_3 & \phi_4 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} y_{t-1} \\ y_{t-2} \\ y_{t-3} \\ y_{t-4} \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} u_t.$$

After an explicit computation, one can verify that it lead to the classic AR(4) recursion, i.e.

$$\begin{pmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ y_{t-3} \end{pmatrix} = \begin{pmatrix} \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \phi_4 y_{t-4} \\ y_{t-1} \\ y_{t-2} \\ y_{t-3} \end{pmatrix} + \begin{pmatrix} u_t \\ 0 \\ 0 \\ 0 \end{pmatrix} =$$

$$= \begin{pmatrix} \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \phi_4 y_{t-4} + u_t \\ y_{t-1} \\ y_{t-2} \\ y_{t-3} \end{pmatrix}$$

20.1.2 Matrix for ARMA

In matrix form an ARMA (p,q) process reads,

$$\mathbf{X}_t = \mathbf{c} + \mathbf{A}\mathbf{X}_{t-1} + \mathbf{b}u_t,$$

where, the first component, namely $y_t = \mathbf{e}_{p+q}^\top \mathbf{X}_t$, is extracted as:

$$y_t = \mathbf{e}_{p+q}^\top \mathbf{c} + \mathbf{e}_{p+q}^\top \mathbf{A}\mathbf{X}_{t-1} + \mathbf{e}_{p+q}^\top \mathbf{b}\, u_t.$$

where $\mathbf{e}_{p+q}^\top$ is a basis vector (Equation 32.1) with dimension $p+q$.

⚠ How to construct the companion matrix for an ARMA?

Let's consider the vector γ containing the coefficients of the model, i.e.

$$\underset{1\times(q+p)}{\gamma} = \begin{pmatrix} \phi_1 & \phi_2 & \dots & \phi_{p-1} & \phi_p & \theta_1 & \theta_2 & \dots & \theta_{q-1} & \theta_q \end{pmatrix}.$$

For an ARMA(p,q) we consider two distinct matrices, i.e. a matrix for the AR part we consider the identity matrix (Equation 32.3) with order $p-1$, i.e.

$$\underset{(p-1)\times(p-1)}{\mathbf{I}_{p-1}} = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix},$$

and we combine it with a column of zeros, i.e.

$$\underset{(p-1)\times p}{\mathbf{L}_{p-1}} = \begin{pmatrix} \mathbf{I}_{p-1} & \mathbf{0}_{(p-1)\times 1} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & 0 \\ 0 & 0 & \dots & 1 & 0 & 0 \\ 0 & 0 & \dots & 0 & 1 & 0 \end{pmatrix}.$$

For the MA part we consider the identity matrix with order $q-1$, i.e.

$$\underset{(q-1)\times(q-1)}{\mathbf{I}_{q-1}} = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix},$$

and we combine it with a column of zeros, i.e.

$$\underset{(q-1)\times q}{\mathbf{L}_{q-1}} = \begin{pmatrix}\mathbf{I}_{q-1} & \mathbf{0}_{(q-1)\times 1}\end{pmatrix} = \begin{pmatrix}1 & 0 & ... & 0 & 0 & 0\\ 0 & 1 & ... & 0 & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots & 0\\ 0 & 0 & ... & 1 & 0 & 0\\ 0 & 0 & ... & 0 & 1 & 0\end{pmatrix}.$$

To combine the matrices $\mathbf{L}_{p-1}$ and $\mathbf{L}_{q-1}$, we need add a matrix of zeros. More precisely:

$$\underset{(q+p-1)\times(q+p)}{\mathbf{L}} = \begin{pmatrix}\mathbf{L}_{p-1} & \mathbf{0}_{(p-1)\times q}\\ \mathbf{0}_{1\times p} & \mathbf{0}_{1\times q}\\ \mathbf{0}_{(q-1)\times p} & \mathbf{L}_{q-1}\end{pmatrix}.$$

Finally we combine $\mathbf{L}$ with the parameters, i.e.

$$\underset{(q+p)\times(q+p)}{\mathbf{A}} = \begin{pmatrix}\gamma\\ \mathbf{L}\end{pmatrix}.$$

Then, the vector $\mathbf{b}$ is constructed by combining two basis vectors (Equation 32.1), to ensure it is equal to one in the position 1 and in the position p+1, i.e.

$$\mathbf{b} = \begin{pmatrix}\mathbf{e}_p\\ \mathbf{e}_q\end{pmatrix}.$$

Example: matrix form of an ARMA(2,3)

Example 20.2. For example let's consider an ARMA(2,3):

$$\gamma = \begin{pmatrix}\phi_1 & \phi_2 & \theta_1 & \theta_2 & \theta_3\end{pmatrix}.$$

The matrix for the AR part that is equal to combined with a matrix of zeros, i.e.

$$\underset{1\times 2}{\mathbf{L}_{p-1}} = \begin{pmatrix}1 & 0\end{pmatrix},$$

For the MA part we consider a similar matrix as in the single case but with first row equal to zero, i.e.

$$\underset{2\times 3}{\mathbf{L}_{q-1}} = \begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\end{pmatrix},$$

and we combine

$$\underset{4\times 5}{\mathbf{L}} = \begin{pmatrix}\mathbf{L}_{2\times 2} & \mathbf{0}_{2\times 3}\\ \mathbf{0}_{1\times 2} & \mathbf{0}_{1\times 3}\\ \mathbf{0}_{2\times 2} & \mathbf{L}_{2\times 3}\end{pmatrix} = \begin{pmatrix}1 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 1 & 0\end{pmatrix}.$$

Finally we combine

$$\underset{5\times5}{\mathbf{A}} = \begin{pmatrix} \boldsymbol{\gamma} \\ \mathbf{L} \end{pmatrix} = \begin{pmatrix} \phi_1 & \phi_2 & \theta_1 & \theta_2 & \theta_3 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Then, the vector $\mathbf{b}$ is constructed by combining two vectors

$$\mathbf{b} = \begin{pmatrix} \mathbf{e}_2 \\ \mathbf{e}_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \end{pmatrix}^\top.$$

Let's check it would lead to a classic ARMA(2,3):

$$\begin{pmatrix} y_t \\ y_{t-1} \\ u_t \\ u_{t-1} \\ u_{t-2} \end{pmatrix} = \begin{pmatrix} \phi_1 y_{t-1} + \phi_2 y_{t-2} + \theta_1 u_{t-1} + \theta_2 u_{t-1} + \theta_2 u_{t-2} \\ y_{t-1} \\ 0 \\ u_{t-1} \\ u_{t-2} \end{pmatrix} + \begin{pmatrix} u_t \\ 0 \\ u_t \\ 0 \\ 0 \end{pmatrix}$$

$$= \begin{pmatrix} \phi_1 y_{t-1} + \phi_2 y_{t-2} + \theta_1 u_{t-1} + \theta_2 u_{t-1} + \theta_3 u_{t-2} + u_t \\ y_{t-1} \\ u_t \\ u_{t-1} \\ u_{t-2} \end{pmatrix}$$

20.2 Moments

Proposition 20.1. *Given the information at time t, the forecasted value at time $t+h$ for an AR(p) process reads:*

$$\mathbf{X}_{t+h} = \sum_{j=0}^{h-1} \mathbf{A}^j \mathbf{c} + \mathbf{A}^h \mathbf{X}_t + \sum_{j=0}^{h-1} \mathbf{A}^j \mathbf{b}\, u_{t+h-j}. \tag{20.1}$$

Proof of Proposition 20.1

Proof. Let's start developing the Equation 20.1 with $h = 1$, i.e.

$$\mathbf{X}_{t+1} = \mathbf{c} + \mathbf{A}\mathbf{X}_t + \mathbf{b}\, u_{t+1},$$

with $h = 2$:

$$\begin{aligned}\mathbf{X}_{t+2} &= \mathbf{c} + \mathbf{A}\mathbf{X}_{t+1} + \mathbf{b}\, u_{t+2} = \\ &= \mathbf{A}\left(\mathbf{c} + \mathbf{A}\mathbf{X}_t + \mathbf{b}\, u_{t+1}\right) + \mathbf{b} \cdot u_{t+2} = \\ &= \mathbf{c} + \mathbf{A}\mathbf{c} + \mathbf{A}^2\mathbf{X}_t + \mathbf{b}\, u_{t+2} + \mathbf{A}\mathbf{b}\, u_{t+1}\end{aligned}$$

with $h = 3$:

$$\begin{aligned}\mathbf{X}_{t+3} &= \mathbf{c} + \mathbf{A}\mathbf{X}_{t+2} + \mathbf{b}\; u_{t+3} = \\ &= \mathbf{c} + \mathbf{A}\left(\mathbf{c} + \mathbf{A}\mathbf{c} + \mathbf{A}^2\mathbf{X}_t + \mathbf{b}\, u_{t+2} + \mathbf{A}\mathbf{b}\, u_{t+1}\right) + \mathbf{b}\, u_{t+3} = \\ &= \mathbf{c} + \mathbf{A}\mathbf{c} + \mathbf{A}^2\mathbf{c} + \mathbf{A}^3\mathbf{X}_t + \mathbf{b}\, u_{t+3} + \mathbf{A}\mathbf{b}\, u_{t+2} + \mathbf{A}^2\mathbf{b}\, u_{t+1}\end{aligned}$$

and so on. Hence the impact of the shocks decrease exponentially over time. □

20.2.1 Expectation

Proposition 20.2. *The conditional expectation at time $t+h$ of $\mathbf{X}_{t+h}$ given the information up to time t can be easily computed from Equation 20.1, i.e.*

$$\mathbb{E}\{\mathbf{X}_{t+h} \mid \mathcal{F}_t\} = \mathbf{c}\,\left(\mathbf{I}_p - \mathbf{A}^h\right)\,\left(\mathbf{I}_p - \mathbf{A}\right)^{-1} + \mathbf{A}^h\mathbf{X}_t.$$

i Proof of Proposition 20.2

Proof. The expectation of Equation 20.1 reads:

$$\mathbb{E}\{\mathbf{X}_{t+h} \mid \mathcal{F}_t\} = \sum_{j=0}^{h-1}\mathbf{A}^j\mathbf{c} + \mathbf{A}^h\mathbf{X}_t + \sum_{j=0}^{h-1}\mathbf{A}^j\mathbf{b} \cdot \mathbb{E}\{u_{t+h-j} \mid \mathcal{F}_t\}.$$

Under the assumoption that $\varepsilon_{t+h-j} \sim MDS$, we have that for all t

$$\mathbb{E}\{u_{t+1} \mid \mathcal{F}_t\} = 0.$$

Therefore, we can apply the tower property of the conditional expectation with $\mathcal{F}_t$ an increasing filtration such that $\mathcal{F}_t \subset \mathcal{F}_{t+1} \subset \ldots \mathcal{F}_{t+h}$, i.e. for example with $t+2$,

$$\mathbb{E}\{u_{t+2} \mid \mathcal{F}_t\} = \mathbb{E}\{\mathbb{E}\{u_{t+2} \mid \mathcal{F}_{t+1}\} \mid \mathcal{F}_t\} = 0,$$

with $t+3$,

$$\mathbb{E}\{u_{t+3} \mid \mathcal{F}_t\} = \mathbb{E}\{\mathbb{E}\{u_{t+3} \mid \mathcal{F}_{t+1}\} \mid \mathcal{F}_t\} = \mathbb{E}\{\mathbb{E}\{\mathbb{E}\{u_{t+3} \mid \mathcal{F}_{t+2}\} \mid \mathcal{F}_{t+1}\} \mid \mathcal{F}_t\} = 0.$$

Therefore

$$\mathbb{E}\{\mathbf{X}_{t+h} \mid \mathcal{F}_t\} = \sum_{j=0}^{h-1}\mathbf{A}^j\mathbf{c} + \mathbf{A}^h\mathbf{X}_t,$$

where for constant $\mathbf{c}$, the series further simplifies in

$$\sum_{j=0}^{h-1} \mathbf{A}^j \mathbf{c} = (\mathbf{I}_p - \mathbf{A}^h)\ (\mathbf{I}_p - \mathbf{A})^{-1}\mathbf{c},$$

with $\mathbf{I}_p$ the identity matrix (Equation 32.3). □

20.2.2 Covariance

Proposition 20.3. *Taking the variance of Equation 20.1 on both sides gives*

$$\mathbb{V}\{\mathbf{X}_{t+h} \mid \mathcal{F}_t\} = \sigma_u^2 \sum_{j=0}^{h-1} \mathbf{A}^j \mathbf{b}\mathbf{b}^\top (\mathbf{A}^j)^\top.$$

Moreover, Assuming $u_t \sim WN(0, \sigma_u^2)$, and independence across time, the conditional covariance is:

$$\mathbb{C}v\{\mathbf{X}_{t+h}, \mathbf{X}_{t+k} \mid \mathcal{F}_t\} = \sigma_u^2 \sum_{j=0}^{\min(h,k)-1} \mathbf{A}^{h-1-j} \mathbf{b}\mathbf{b}^\top (\mathbf{A}^{k-1-j})^\top.$$

i Proof of Proposition 20.3

Proof. Taking the conditional variance gives

$$\mathbb{V}\{\mathbf{X}_{t+h} \mid \mathcal{F}_t\} = \sum_{j=0}^{h-1} \mathbb{V}\{\mathbf{A}^j \mathbf{b}\, u_{t+h-j} \mid \mathcal{F}_t\} = \sum_{j=0}^{h-1} \mathbf{A}^j \mathbf{b} \mathbb{V}\{u_{t+h-j} \mid \mathcal{F}_t\} \mathbf{b}^\top (\mathbf{A}^j)^\top.$$

Then, if u_t is White Noise process, then its variance is constant, hence independent from t, and equal to σ_u^2. For the covariance formula, each $\mathbf{X}_{t+h}$ is influenced by past shocks $ut + h - j$. Since the shocks are uncorrelated across time, only shared shocks affect both $\mathbf{X}_{t+h}$ and $\mathbf{X}_{t+k}$. The number of common shocks is $\min(h, k)$, and each contributes:

$$\mathbb{C}v\{\mathbf{A}^{h-1-j}\mathbf{b}u_{t+1+j}, \mathbf{A}^{k-1-j}\mathbf{b}u_{t+1+j}\} = \sigma_u^2 \mathbf{A}^{h-1-j}\mathbf{b}\mathbf{b}^\top (\mathbf{A}^{k-1-j})^\top.$$

□

Example: ARMA(2,3) iterative forecast

Example 20.3. Let's use Monte Carlo simulations to establish if the results are accurate.

Table 20.1: Theoric long term moments and moments computed on 200 Monte Carlo simulations (t = 100000).

Covariance	Formula	MonteCarlo
$\mathbb{E}\{y_t\}$	0.5454545	0.5451258
$\mathbb{V}\{y_t\}$	1.7466575	1.7472192
$\mathbb{C}v\{y_t, y_{t-1}\}$	1.1317615	1.1322255
$\mathbb{C}v\{y_t, y_{t-2}\}$	0.8115271	0.8118823
$\mathbb{C}v\{y_t, y_{t-3}\}$	0.5132223	0.5133506
$\mathbb{C}v\{y_t, y_{t-4}\}$	0.2756958	0.2758341
$\mathbb{C}v\{y_t, y_{t-5}\}$	0.1596921	0.1596755

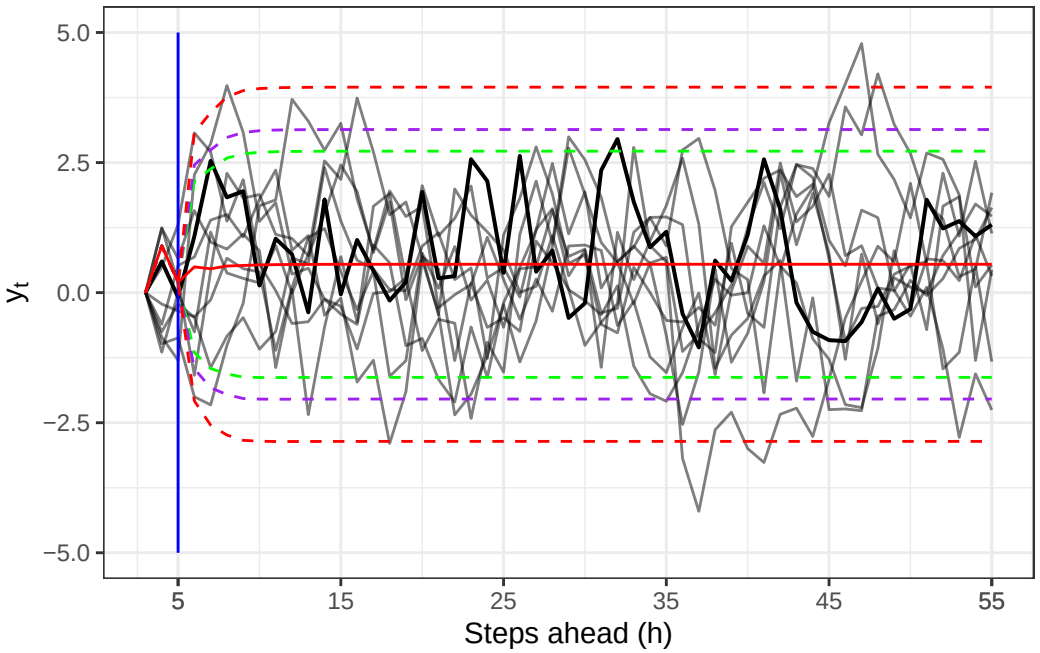


Figure 20.2: ARMA(2,3) simulations with expected value (red) and confidence intervals with $\alpha = 0.1$ (green), $\alpha = 0.05$ (purple) and $\alpha = 0.01$ (red).

21 Conditional variance processes

21.1 ARCH(p) process

The auto regressive and conditionally heteroskedastic process (ARCH) were introduced in 1982 by Robert Engle to model the conditional variance of a time series. It is often an empirical fact in economics that the larger values of time series the larger the variance. Let's define an ARCH process of order p as:

$$\begin{aligned} y_t &= \mathbb{E}\{y_t \mid \mathcal{F}_{t-1}\} + \mathbb{V}\{y_t \mid \mathcal{F}_{t-1}\} u_t \\ \mathbb{E}\{y_t \mid \mathcal{F}_{t-1}\} &= \mu \\ \mathbb{V}\{y_t \mid \mathcal{F}_{t-1}\} &= \sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i (y_{t-i} - \mu)^2 \end{aligned} \tag{21.1}$$

where $u_t \sim \text{MDS}(0, 1)$.

21.1.1 Moments

The conditional mean of an ARCH(p) process (Equation 21.1) is equal to

$$\mathbb{E}\{y_t \mid \mathcal{F}_{t-1}\} = \mu.$$

and the conditional variance is not stochastic given the information at time $t-1$ and for a general $h \geq 1$ reads:

$$\mathbb{V}\{y_{t+h} \mid \mathcal{F}_t\} = \mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_t\}.$$

Proof: Conditional moments ARCH(p)

Proof. Taking the conditional expectation on the process in Equation 21.1 one obtain:

$$\begin{aligned} \mathbb{E}\{y_t \mid \mathcal{F}_{t-1}\} &= \mu + \mathbb{E}\{\sigma_t u_t \mid \mathcal{F}_{t-1}\} = \\ &= \mu + \sigma_t \mathbb{E}\{u_t \mid \mathcal{F}_{t-1}\} = \\ &= \mu \end{aligned}$$

since σ_t is known and not stochastic given the information at $t-1$ in $\mathcal{F}_{t-1}$ and u_t conditionally to $\mathcal{F}_{t-1}$ is a Martingale Difference Sequence with mean zero (Equation 18.2).

The conditional variance of y at time $t+h$ depends on the conditional expectation of σ_t^2, i.e.

$$\begin{aligned}\mathbb{V}\{y_{t+h} \mid \mathcal{F}_t\} &= \mathbb{V}\{\sigma_{t+h}u_{t+h} \mid \mathcal{F}_t\} = \\ &= \mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_t\}\mathbb{E}\{u_{t+h}^2 \mid \mathcal{F}_t\} - \mathbb{E}\{\sigma_{t+h} \mid \mathcal{F}_t\}\mathbb{E}\{u_{t+h} \mid \mathcal{F}_t\} = \\ &= \mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_t\}\end{aligned}$$

since $\mathbb{E}\{u_{t+h} \mid \mathcal{F}_t\} = 0$ and $\mathbb{E}\{u_{t+h}^2 \mid \mathcal{F}_t\} = 1$. □

⚠ Stationarity ARCH(p)

Regarding the variance, the long-term expectation of σ_t^2 reads

$$\mathbb{E}\{\sigma_t^2\} = \frac{\omega}{1 - \sum_{i=1}^{p} \alpha_i}.$$

where $\omega \neq 0$. It is clear that to ensure that the process is stationary the following condition on the ARCH parameters must be satisfied:

$$\sum_{i=1}^{p} \alpha_i < 1, \quad \alpha_i > 0$$

for all $i \in \{1, \dots, p\}$.

21.1.2 Example: ARCH(1) process

Let's simulate an ARCH(1) process with normal residuals, i.e.

$$\begin{aligned}y_t &= \mu + \sigma_t u_t \\ \sigma_t^2 &= \omega + \alpha_1 (y_{t-1} - \mu)^2\end{aligned}$$

where $u_t \sim \mathcal{N}(0, 1)$.

21.1.3 Example: ARCH(3) process

Let's simulate an ARCH(3) process with normal residuals, i.e.

$$\begin{aligned}y_t &= \mu + \sigma_t u_t \\ \sigma_t^2 &= \omega + \alpha_1 (y_{t-1} - \mu)^2 + \alpha_2 (y_{t-2} - \mu)^2 + \alpha_3 (y_{t-3} - \mu)^2\end{aligned}$$

where $u_t \sim \mathcal{N}(0, 1)$.

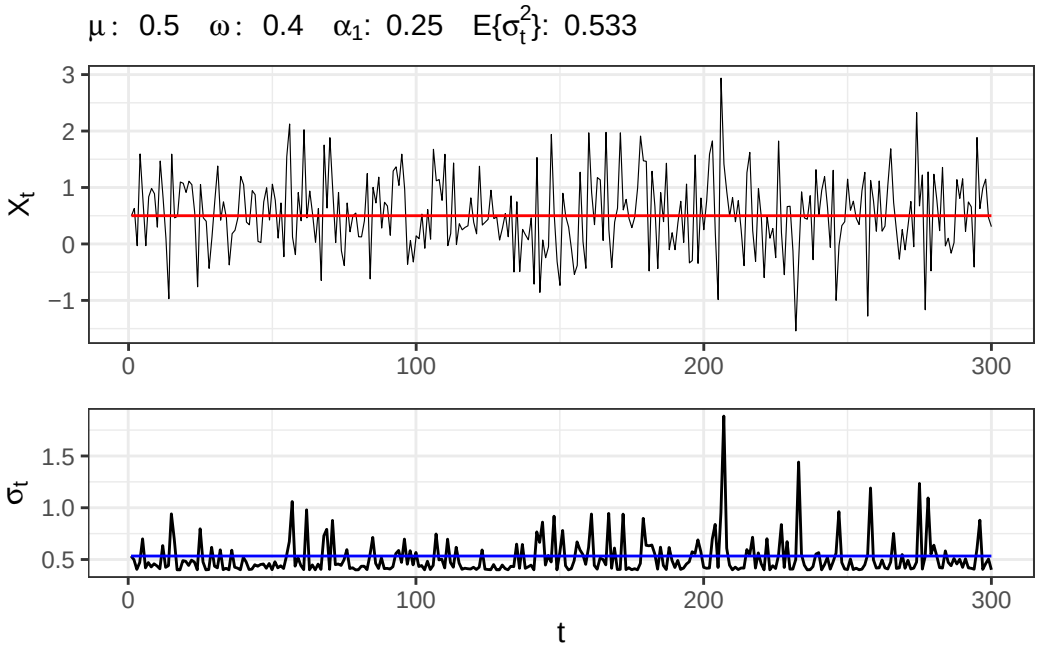


Figure 21.1: On the top an ARCH(1) simulation with its long term mean (red). On the bottom the correspondent stochastic variance with its long term mean (blue).

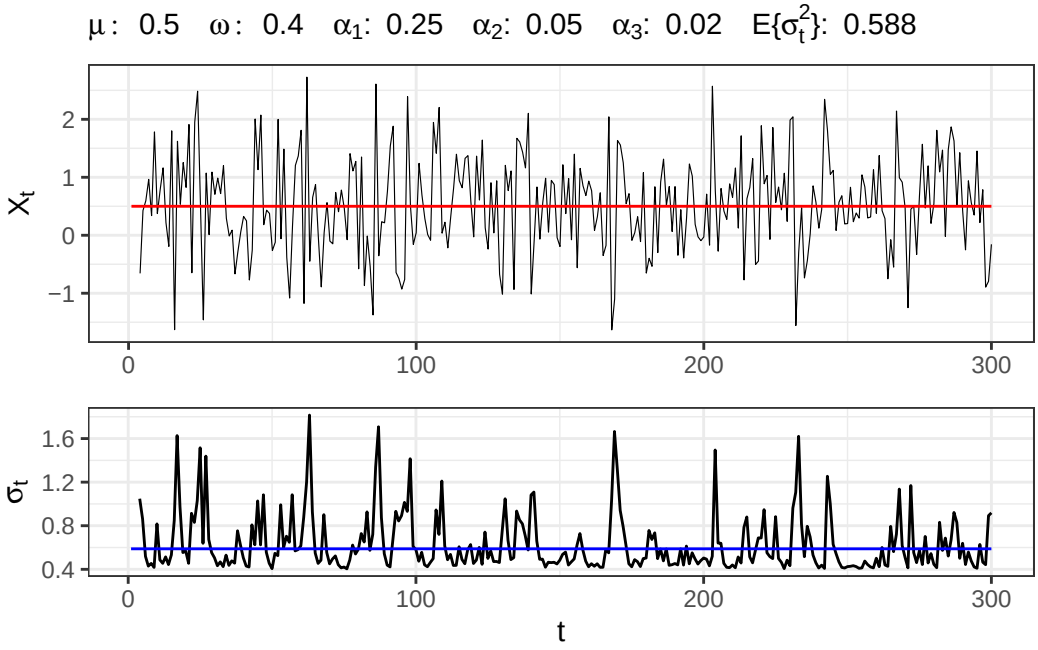


Figure 21.2: On the top an ARCH(3) simulation with its long term mean (red). On the bottom the correspondent stochastic variance with its long term mean (blue).

21.2 GARCH(p,q) process

As done with the ARCH(p), with generalized auto regressive conditional heteroskedasticity (GARCH) we model the dependency of the conditional second moment. It represents a more parsimonious way to express the conditional variance. A GARCH(p,q) processx is defined as:

$$y_t = \mu + \sigma_t u_t$$
$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i (y_{t-i} - \mu)^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2$$

where $u_t \sim \text{MDS}(0, 1)$.

> **Stationarity GARCH(p,q)**
>
> Regarding the variance, the long-term expectation of σ_t^2 reads
>
> $$\mathbb{E}\{\sigma_t^2\} = \frac{\omega}{1 - \sum_{i=1}^{p} \alpha_i - \sum_{j=1}^{q} \beta_j}. \tag{21.2}$$
>
> where $\omega \neq 0$. It is clear that to ensure that the process is stationary the following condition on the GARCH parameters must be satisfied:
>
> $$\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1,$$
>
> with $\alpha_i \geq 0$ and $\beta_j \geq 0$ for all $i \in \{1, \dots, p\}$ and $j \in \{1, \dots, q\}$.

21.2.1 Example: GARCH(1,1) process

Let's simulate an GARCH(1,1) process with normal residuals, i.e.

$$y_t = \mu + \sigma_t u_t$$
$$\sigma_t^2 = \omega + \alpha_1 (y_{t-1} - \mu)^2 + \beta_1 \sigma_{t-1}^2$$

where $u_t \sim \mathcal{N}(0, 1)$.

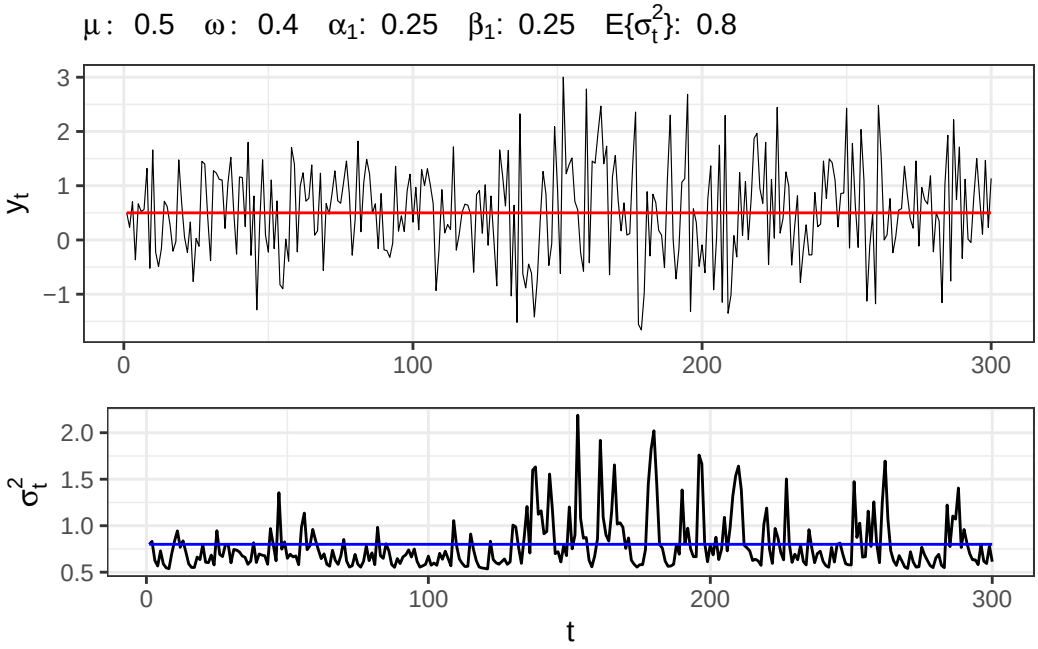


Figure 21.3: On the top a GARCH(1,1) simulation with its long term mean (red). On the bottom the correspondent stochastic variance with its long term mean (blue).

21.2.2 Example: GARCH(2,3) process

Let's simulate a GARCH(2,3) process with normal residuals, i.e.

$$
\begin{aligned}
y_t &= \mu + \sigma_t u_t \\
\sigma_t^2 &= \omega + \sum_{i=1}^{2} \alpha_i (y_{t-i} - \mu)^2 + \sum_{j=1}^{3} \beta_j \sigma_{t-j}^2
\end{aligned}
$$

where $u_t \sim \mathcal{N}(0, 1)$.

21.2.3 Example: GARCH(3,2) process

Let's simulate a GARCH(3,2) process with normal residuals, i.e.

$$
\begin{aligned}
y_t &= \mu + \sigma_t u_t \\
\sigma_t^2 &= \omega + \sum_{i=1}^{3} \alpha_i (y_{t-i} - \mu)^2 + \sum_{j=1}^{2} \beta_j \sigma_{t-j}^2
\end{aligned}
$$

where $u_t \sim \mathcal{N}(0, 1)$.

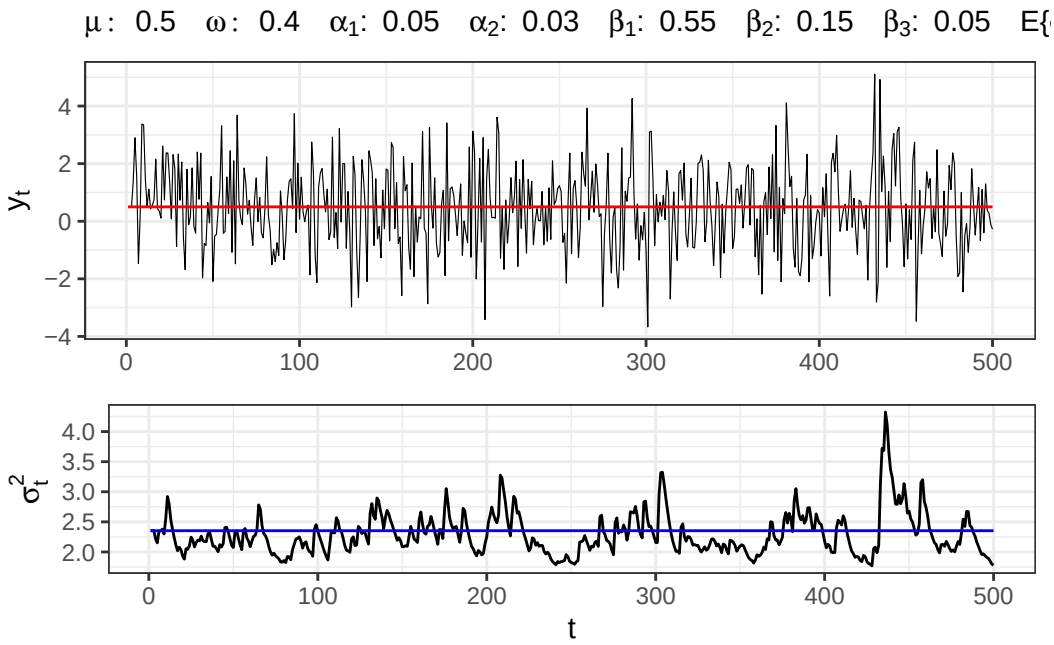


Figure 21.4: On the top a GARCH(2,3) simulation with its long term mean (red). On the bottom the correspondent stochastic variance with its long term mean (blue).

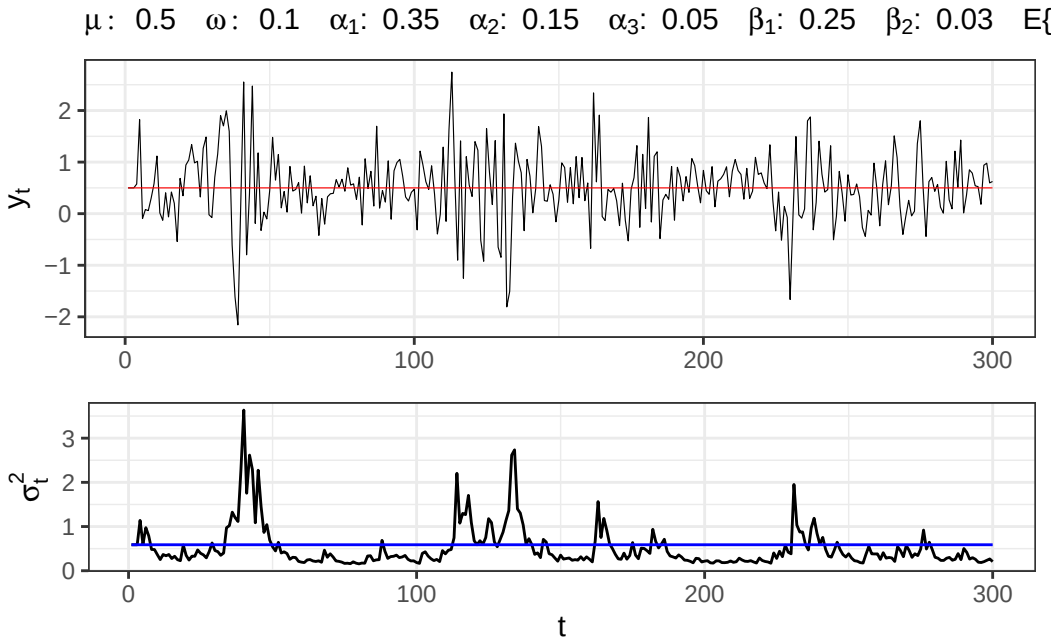


Figure 21.5: On the top a GARCH(3,2) simulation with its long term mean (red). On the bottom the correspondent stochastic variance with its long term mean (blue).

21.3 IGARCH

Many variants of the standard GARCH process were developed in literature. For example, the Integrated Generalized Auto regressive Conditional heteroskedasticity (IGARCH(p,q)) is a restricted version of the GARCH, where the persistent parameters sum up to one, and imply a unit root in the GARCH process with the condition

$$\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j = 1.$$

In GARCH(p, q), the sum of coefficients being less than 1 ensures stationarity (finite unconditional variance) while in IGARCH(p, q), the sum is exactly 1, so the process is nonstationary in variance: the conditional variance has a persistent memory, and the shocks to volatility accumulate over time. The process is strictly stationary under some conditions (see Nelson (1990)), but it has infinite unconditional variance.

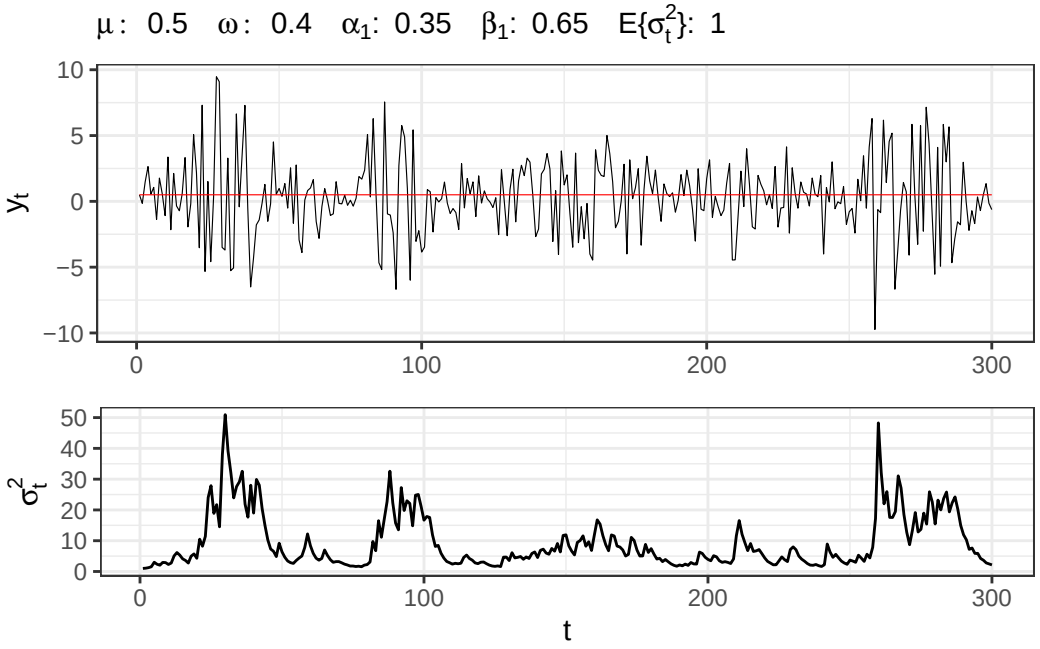


Figure 21.6: On the top an iGARCH(1,1) simulation with its long term mean (red). On the bottom the correspondent stochastic variance with its long term mean (blue).

21.4 GARCH-M

The GARCH in-mean (GARCH-M) model adds a stochastic term into the mean equation. This is motivated especially in financial theories (e.g., risk-return trade-off) suggesting that

expected returns may depend on volatility, i.e.

$$
\begin{aligned}
y_t &= \mu + \sigma_t^2 \lambda + e_t \\
e_t &= \sigma_t u_t \\
\sigma_t^2 &= \omega + \sum_{i=1}^{p} \alpha_i e_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2
\end{aligned}
$$

The effect of the parameter λ is a shift of the mean of the process. If y_t are for example financial returns a $\lambda > 0$ would imply that higher volatility increases expected return, consistent with risk-premium theories. Instead, when $\lambda < 0$, it could reflect behavioral phenomena or model misspecification. The unconditional mean of y_t became

$$\mathbb{E}\{y_t\} = \mu + \mathbb{E}\{\sigma_t^2\}\lambda,$$

while the conditional mean

$$\mathbb{E}\{y_{t+h} \mid \mathcal{F}_t\} = \mu + \lambda \mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_t\}.$$

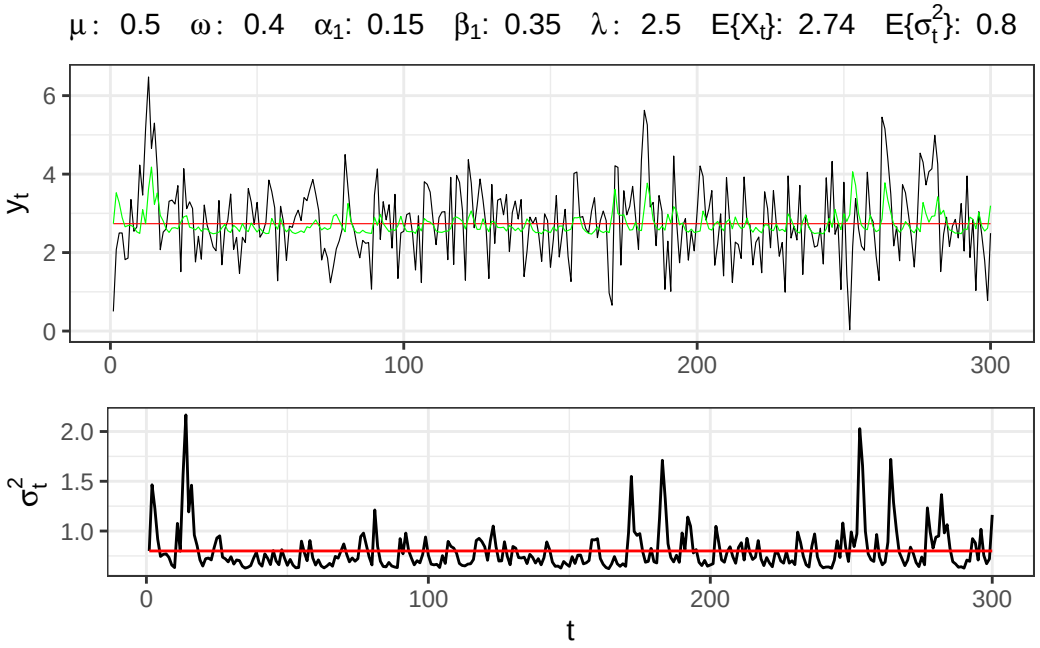


Figure 21.7: GARCH-M(1,1) simulation (top) with the conditional (green) and long term (red) expected values. Simulated GARCh variance (bottom) with the long term expected value (red).

22 GARCH(1,1) moments

Let's consider a very general setup for GARCH(1,1) process where we do not assume that the GARCH residuals

1. Has expected value equal to zero, i.e. $\mathbb{E}\{u_t\} = 0$.
2. Has NOT necessary a second moment that is constant and equal to one, i.e. $\mathbb{E}\{u_t^2\}$ possibly time dependent (deterministic).
3. Has NOT necessary a fourth moment that is constant and equal to 3, i.e. $\mathbb{E}\{u_t^4\}$ possibly time dependent (deterministic).

The process we refer to has the form of a standard GARCH(1,1), i.e.

$$\begin{aligned} y_t &= \sigma_t u_t \\ \sigma_t^2 &= \omega + \alpha_1 y_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \end{aligned}$$

where $u_t \sim \text{WN}(\mu_t, \sigma_t)$.

22.1 First moment σ_t^2

22.1.1 Short-term

Given the information at time $t-1$, the expected value of the GARCH variance after h-steps can be expanded as:

$$\mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\} = \omega \left(1 + \sum_{j=1}^{h-1} \prod_{i=1}^{j} \lambda_{t+h-i}\right) + \sigma_t^2 \prod_{j=1}^{h} \lambda_{t+h-j}, \tag{22.1}$$

with $h \geq 1$ and where in general

$$\lambda_{t+h-i} = \alpha_1 \mathbb{E}\{u_{t+h-i}^2\} + \beta_1. \tag{22.2}$$

The iteration between two consecutive times reads

$$\mathbb{E}\{\sigma_{t+h-s}^2 \mid \mathcal{F}_{t-1}\} = \omega + \lambda_{t+h-s-1} \mathbb{E}\{\sigma_{t+h-s-1}^2 \mid \mathcal{F}_{t-1}\}.$$

Proof: GARCH(1,1) iterative expectation

Proof. Let's start by taking the conditional expectation of the GARCH(1,1) variance at time t given the information up to $t-1$. In this case it is fully known at time t given the information in $t-1$, i.e.

$$\mathbb{E}\{\sigma_t^2 \mid \mathcal{F}_{t-1}\} = \omega + \alpha_1 y_{t-1}^2 + \beta_1 \sigma_{t-1}^2.$$

Then, let's iterate the expectation at time $t+1$, i.e.

$$\mathbb{E}\{\sigma_{t+1}^2 \mid \mathcal{F}_{t-1}\} = \omega + \alpha_1 \mathbb{E}\{y_t^2\} + \beta_1 \mathbb{E}\{\sigma_t^2 \mid \mathcal{F}_{t-1}\},$$

and let's substitute the expression for the squared residuals $y_t^2 = \sigma_t^2 u_t^2$. Since σ_t^2 at time $t-1$ is known we have:

$$\mathbb{E}\{\sigma_{t+1}^2 \mid \mathcal{F}_{t-1}\} = \omega + (\alpha_1 \mathbb{E}\{u_t^2\} + \beta_1)\mathbb{E}\{\sigma_t^2 \mid \mathcal{F}_{t-1}\} = \omega + \lambda_t \sigma_t^2,$$

where

$$\lambda_t = \alpha_1 \mathbb{E}\{u_t^2\} + \beta_1.$$

Iterating the expectation at time $t+2$

$$\begin{aligned}
\mathbb{E}\{\sigma_{t+2}^2 \mid \mathcal{F}_{t-1}\} &= \omega + \lambda_{t+1}\mathbb{E}\{\sigma_{t+1}^2 \mid \mathcal{F}_{t-1}\} \\
&= \omega + \lambda_{t+1}\left(\omega + \lambda_t \sigma_t^2\right) \\
&= \omega(1 + \lambda_{t+1}) + \lambda_{t+1}\lambda_t \sigma_t^2
\end{aligned}$$

Then, at time $t+3$

$$\begin{aligned}
\mathbb{E}\{\sigma_{t+3}^2 \mid \mathcal{F}_{t-1}\} &= \omega + \lambda_{t+2}\,\mathbb{E}\{\sigma_{t+2}^2 \mid \mathcal{F}_{t-1}\} \\
&= \omega + \lambda_{t+2}\left(\omega(1 + \lambda_{t+1}) + \lambda_{t+1}\lambda_t \sigma_t^2\right) \\
&= \omega\left(1 + \lambda_{t+2} + \lambda_{t+2}\lambda_{t+1}\right) + \lambda_{t+2}\lambda_{t+1}\lambda_t \sigma_t^2
\end{aligned}$$

In general, after h-steps one can write the expansion

$$\mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\} = \omega\left(1 + \sum_{j=1}^{h-1}\prod_{i=1}^{j}\lambda_{t+h-i}\right) + \sigma_t^2 \prod_{j=1}^{h}\lambda_{t+h-j}$$

with the convention that an empty product is 1. □

Example: GARCH(1,1) iterative first moment

Example 22.1. Expanding the expression in Equation 22.1 with $h = 1$ one obtain

$$\begin{aligned}\mathbb{E}\{\sigma^2_{t+1} \mid \mathcal{F}_{t-1}\} &= \omega\left(1 + \sum_{j=1}^{0}\prod_{i=1}^{j}\lambda_{t+1-i}\right) + \sigma^2_t\prod_{j=1}^{1}\lambda_{t+1-j} = \\ &= \omega + \sigma^2_t\lambda_t\end{aligned}$$

with $h = 2$

$$\begin{aligned}\mathbb{E}\{\sigma^2_{t+2} \mid \mathcal{F}_{t-1}\} &= \omega\left(1 + \sum_{j=1}^{1}\prod_{i=1}^{j}\lambda_{t+2-i}\right) + \sigma^2_t\prod_{j=1}^{2}\lambda_{t+2-j} = \\ &= \omega(1 + \lambda_{t+1}) + \sigma^2_t(\lambda_{t+1}\lambda_t)\end{aligned}$$

with $h = 3$

$$\begin{aligned}\mathbb{E}\{\sigma^2_{t+3} \mid \mathcal{F}_{t-1}\} &= \omega\left(1 + \sum_{j=1}^{2}\prod_{i=1}^{j}\lambda_{t+3-i}\right) + \sigma^2_t\prod_{j=1}^{3}\lambda_{t+3-j} = \\ &= \omega(1 + \lambda_{t+2} + \lambda_{t+1}\lambda_{t+2}) + \sigma^2_t(\lambda_{t+2}\lambda_{t+1}\lambda_t)\end{aligned}$$

and so on.

Moment	Step	Formula	Iteration	Difference. (%)
$\mathbb{E}\{\sigma^2_{t+0} \mid \mathcal{F}_{t-1}\}$	0	1.200	1.200	0%
$\mathbb{E}\{\sigma^2_{t+1} \mid \mathcal{F}_{t-1}\}$	1	1.235	1.235	0%
$\mathbb{E}\{\sigma^2_{t+2} \mid \mathcal{F}_{t-1}\}$	2	1.250	1.250	0%
$\mathbb{E}\{\sigma^2_{t+3} \mid \mathcal{F}_{t-1}\}$	3	1.258	1.258	0%
$\mathbb{E}\{\sigma^2_{t+4} \mid \mathcal{F}_{t-1}\}$	4	1.261	1.261	0%
$\mathbb{E}\{\sigma^2_{t+5} \mid \mathcal{F}_{t-1}\}$	5	1.263	1.263	0%
$\mathbb{E}\{\sigma^2_{t+6} \mid \mathcal{F}_{t-1}\}$	6	1.263	1.263	0%
$\mathbb{E}\{\sigma^2_{t+7} \mid \mathcal{F}_{t-1}\}$	7	1.264	1.264	0%
$\mathbb{E}\{\sigma^2_{t+8} \mid \mathcal{F}_{t-1}\}$	8	1.264	1.264	0%
$\mathbb{E}\{\sigma^2_{t+9} \mid \mathcal{F}_{t-1}\}$	9	1.264	1.264	0%
$\mathbb{E}\{\sigma^2_{t+10} \mid \mathcal{F}_{t-1}\}$	10	1.264	1.264	0%

Table 22.1: Forecasted expectation of GARCH(1,1) variance with iteration and with formula.

22.1.2 Long-term

If $\mathbb{E}\{u_t^2\}$ is constant for all t then, $\lambda = \alpha_1 \mathbb{E}\{u_t^2\} + \beta_1$, became a constant and the formula simplifies to

$$\mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\} = \sigma_\infty^2 + \lambda^h(\sigma_t^2 - \sigma_\infty^2), \tag{22.3}$$

where σ_∞^2 denotes the long-term expected GARCH variance as $h \to \infty$, i.e.

$$\lim_{h\to\infty} \mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\} = \frac{\omega}{1-\lambda} = \sigma_\infty^2. \tag{22.4}$$

It follows that, under the standard assumption that $u_t \sim \text{WN}(0, 1)$, then one obtain the classic expression

$$\mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\} = \sigma_\infty^2 + (\alpha_1 + \beta_1)^h(\sigma_t^2 - \sigma_\infty^2) \tag{22.5}$$

where σ_∞^2 denotes the unconditional expectation as Equation 22.4 with $\mathbb{E}\{u_t^2\} = 1$.

GARCH(1,1) long-term expectation

Proof. Let's verify the formula in Equation 22.5 for constant λ_t (Equation 22.2) for all t, i.e.

$$\lambda = \alpha_1 \mathbb{E}\{u_t^2\} + \beta_1$$

In this case the iterative formula (Equation 22.1) simplifies, i.e.

$$\begin{aligned}
\mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\} &= \omega\left(1 + \sum_{j=1}^{h-1}\prod_{i=1}^{j}\lambda\right) + \sigma_t^2\prod_{j=1}^{h}\lambda = \\
&= \omega\sum_{j=0}^{h-1}\lambda^j + \sigma_t^2\lambda^h = \\
&= \omega\left(\frac{1-\lambda^h}{1-\lambda}\right) + \sigma_t^2\lambda^h \\
&= \frac{\omega}{1-\lambda} + \lambda^h\left(\sigma_t^2 - \frac{\omega}{1-\lambda}\right)
\end{aligned}$$

Taking the limit as $h \to \infty$ gives the long term stationary variance, i.e.

$$\lim_{h\to\infty} \mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\} = \frac{\omega}{1-\lambda} = \sigma_\infty^2 \iff \lambda < 1$$

Hence, the general expression became

$$\mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\} = \sigma_\infty^2 + \lambda^h(\sigma_t^2 - \sigma_\infty^2)$$

□

Example: GARCH(1,1) long-term first moment

Example 22.2.

Moment	Formula	MonteCarlo	Difference (%)
σ^2_∞	1.264	1.264	-0.0056%

Table 22.2: Forecasted long-term expectation of GARCH(1,1) variance with formula and by Monte Carlo simulations.

22.2 Second moment σ_t^2

22.2.1 Short term

The second moment admits an iterative formula, i.e.

$$\mathbb{E}\{\sigma^4_{t+h} \mid \mathcal{F}_{t-1}\} = \sum_{i=1}^{h} b_{t+h-i} \prod_{j=1}^{i-1} \gamma_{t+h-j} + \sigma_t^4 \prod_{j=1}^{h} \gamma_{t+h-j} \tag{22.6}$$

where

$$\gamma_{t+h-j} = \alpha_1^2 \mathbb{E}\{u^4_{t+h-j}\} + \beta_1(2\alpha_1 \mathbb{E}\{u^2_{t+h-j}\} + \beta_1) \tag{22.7}$$

while

$$b_{t+h-i} = \omega(\omega + 2\lambda_{t+h-i}\mathbb{E}\{\sigma^2_{t+h-i} \mid \mathcal{F}_{t-1}\}) \tag{22.8}$$

with λ_{t+h-i} as in Equation 22.2.

Proof: Iterative formula for the second moment of GARCH(1,1) variance

Proof. Starting from

$$\sigma^4_{t+1} = (\omega + \alpha_1 e_t^2 + \beta_1 \sigma_t^2)^2$$

and substitute the definition of $e_t^2 = \sigma_t^2 u_t^2$, i.e.

$$\begin{aligned}\sigma^4_{t+1} &= (\omega + (\alpha_1 u_t^2 + \beta_1)\sigma_t^2)^2 = \\ &= \omega^2 + 2\omega(\alpha_1 u_t^2 + \beta_1)\sigma_t^2 + (\alpha_1 u_t^2 + \beta_1)^2 \sigma_t^4\end{aligned}$$

Then, let's take the conditional expectation on both sides:

$$\begin{aligned}\mathbb{E}\{\sigma^4_{t+1} \mid \mathcal{F}_t\} &= \omega^2 + \mathbb{E}\{(\alpha_1 u_t^2 + \beta_1)^2\}\mathbb{E}\{\sigma_t^4 \mid \mathcal{F}_t\} + 2\omega(\alpha_1 \mathbb{E}\{u_t^2\} + \beta_1)\mathbb{E}\{\sigma_t^2 \mid \mathcal{F}_t\} = \\ &= \omega^2 + \mathbb{E}\{(\alpha_1 u_t^2 + \beta_1)^2\}\mathbb{E}\{\sigma_t^4 \mid \mathcal{F}_t\} + 2\omega\lambda_t \mathbb{E}\{\sigma_t^2 \mid \mathcal{F}_t\}\end{aligned}$$

where $\lambda_t = \alpha_1 \mathbb{E}\{u_t^2\} + \beta_1$. Then note that:

$$\begin{aligned}\mathbb{E}\{(\alpha_1 u_t^2 + \beta_1)^2\} &= \mathbb{E}\{\alpha_1^2 u_t^4 + \beta_1^2 + 2\alpha_1\beta_1 u_t^2\} = \\ &= \alpha_1^2 \mathbb{E}\{u_t^4\} + \beta_1(2\alpha_1 \mathbb{E}\{u_t^2\} + \beta_1) = \gamma_t\end{aligned}$$

Hence, we can write the expectation in terms of the previous expectation.

$$\begin{aligned}\mathbb{E}\{\sigma_{t+1}^4 \mid \mathcal{F}_t\} &= \omega^2 + \gamma_t \mathbb{E}\{\sigma_t^4 \mid \mathcal{F}_t\} + 2\omega\lambda_t \mathbb{E}\{\sigma_t^2 \mid \mathcal{F}_t\} = \\ &= \omega(\omega + 2\lambda_t \mathbb{E}\{\sigma_t^2 \mid \mathcal{F}_t\}) + \gamma_t \mathbb{E}\{\sigma_t^4 \mid \mathcal{F}_t\} = \\ &= b_t + \gamma_t \mathbb{E}\{\sigma_t^4 \mid \mathcal{F}_t\}\end{aligned}$$

where $b_t = \omega(\omega + 2\lambda_t \mathbb{E}\{\sigma_t^2 \mid \mathcal{F}_t\})$. Iterating at time $t+2$:

$$\begin{aligned}\mathbb{E}\{\sigma_{t+2}^4 \mid \mathcal{F}_t\} &= b_{t+1} + \gamma_{t+1} \mathbb{E}\{\sigma_{t+1}^4 \mid \mathcal{F}_t\} = \\ &= b_{t+1} + \gamma_{t+1} b_t + \gamma_{t+1}\gamma_t \mathbb{E}\{\sigma_t^4 \mid \mathcal{F}_t\}\end{aligned}$$

At time $t+3$

$$\begin{aligned}\mathbb{E}\{\sigma_{t+3}^4 \mid \mathcal{F}_t\} &= b_{t+2} + \gamma_{t+2} \mathbb{E}\{\sigma_{t+2}^4 \mid \mathcal{F}_t\} = \\ &= b_{t+2} + \gamma_{t+2} b_{t+1} + \gamma_{t+2}\gamma_{t+1} b_t + \gamma_{t+2}\gamma_{t+1}\gamma_t \mathbb{E}\{\sigma_t^4 \mid \mathcal{F}_t\}\end{aligned}$$

Hence, in general

$$\mathbb{E}\{\sigma_{t+h}^4 \mid \mathcal{F}_t\} = \sum_{i=1}^{h} b_{t+h-i} \prod_{j=1}^{i-1} \gamma_{t+h-j} + \sigma_t^4 \prod_{j=1}^{h} \gamma_{t+h-j}$$

where we denote as

$$\begin{aligned}\gamma_t &= \alpha_1^2 \mathbb{E}\{u_t^4\} + \beta_1(2\alpha_1 \mathbb{E}\{u_t^2\} + \beta_1) \\ \lambda_t &= \alpha_1 \mathbb{E}\{u_t^2\} + \beta_1 \\ b_t &= \omega^2 + 2\lambda_t \omega \mathbb{E}\{\sigma_t^2 \mid \mathcal{F}_t\}\end{aligned}$$

□

Example: GARCH(1,1) iterative second moment

Example 22.3. With $h = 1$

$$\begin{aligned}\mathbb{E}\{\sigma_{t+1}^4 \mid \mathcal{F}_t\} &= \sum_{i=1}^{1} b_{t+1-i} \prod_{j=1}^{i-1} \gamma_{t+1-j} + \sigma_t^4 \prod_{j=1}^{1} \gamma_{t+1-j} = \\ &= b_t + \gamma_t \sigma_t^4\end{aligned}$$

With $h = 2$

$$\begin{aligned}\mathbb{E}\{\sigma^4_{t+2} \mid \mathcal{F}_t\} &= \sum_{i=1}^{2} b_{t+2-i} \prod_{j=1}^{i-1} \gamma_{t+2-j} + \sigma^4_t \prod_{j=1}^{2} \gamma_{t+2-j} = \\ &= b_{t+1} \prod_{j=1}^{0} \gamma_{t+2-j} + b_t \prod_{j=1}^{1} \gamma_{t+2-j} + \sigma^4_t \gamma_{t+1}\gamma_t = \\ &= b_{t+1} + b_t\gamma_{t+1} + \sigma^4_t \gamma_{t+1}\gamma_t\end{aligned}$$

With $h = 3$

$$\begin{aligned}\mathbb{E}\{\sigma^4_{t+3} \mid \mathcal{F}_t\} &= \sum_{i=1}^{3} b_{t+3-i} \prod_{j=1}^{i-1} \gamma_{t+3-j} + \sigma^4_t \prod_{j=1}^{3} \gamma_{t+3-j} = \\ &= b_{t+2} \prod_{j=1}^{0} \gamma_{t+3-j} + b_{t+1} \prod_{j=1}^{1} \gamma_{t+3-j} + b_t \prod_{j=1}^{2} \gamma_{t+3-j} + \sigma^4_t \gamma_{t+2}\gamma_{t+1}\gamma_t = \\ &= b_{t+2} + b_{t+1}\gamma_{t+2} + b_t\gamma_{t+2}\gamma_{t+1} + \sigma^4_t \gamma_{t+2}\gamma_{t+1}\gamma_t\end{aligned}$$

and so on.

Moment	step	Formula	Iteration	Difference
$\mathbb{E}\{\sigma^4_{t+0} \mid \mathcal{F}_{t-1}\}$	0	1.440000	1.440000	0%
$\mathbb{E}\{\sigma^4_{t+1} \mid \mathcal{F}_{t-1}\}$	1	1.559380	1.559380	0%
$\mathbb{E}\{\sigma^4_{t+2} \mid \mathcal{F}_{t-1}\}$	2	1.609123	1.609123	0%
$\mathbb{E}\{\sigma^4_{t+3} \mid \mathcal{F}_{t-1}\}$	3	1.630762	1.630762	0%
$\mathbb{E}\{\sigma^4_{t+4} \mid \mathcal{F}_{t-1}\}$	4	1.640413	1.640413	0%
$\mathbb{E}\{\sigma^4_{t+5} \mid \mathcal{F}_{t-1}\}$	5	1.644775	1.644775	0%
$\mathbb{E}\{\sigma^4_{t+6} \mid \mathcal{F}_{t-1}\}$	6	1.646762	1.646762	0%
$\mathbb{E}\{\sigma^4_{t+7} \mid \mathcal{F}_{t-1}\}$	7	1.647669	1.647669	0%
$\mathbb{E}\{\sigma^4_{t+8} \mid \mathcal{F}_{t-1}\}$	8	1.648085	1.648085	0%
$\mathbb{E}\{\sigma^4_{t+9} \mid \mathcal{F}_{t-1}\}$	9	1.648275	1.648275	0%
$\mathbb{E}\{\sigma^4_{t+10} \mid \mathcal{F}_{t-1}\}$	10	1.648363	1.648363	0%

Table 22.3: Forecasted second moment of GARCH(1,1) variance with iteration and with formula.

22.2.2 Long-term

If $\mathbb{E}\{u_t^2\}$ and $\mathbb{E}\{u_t^3\}$ are constant for all t then the formula simplifies to

$$\mathbb{E}\{\sigma_{t+h}^4 \mid \mathcal{F}_{t-1}\} = (\omega^2 + 2\omega\lambda\sigma_\infty^2)\left(\frac{1-\gamma^h}{1-\gamma}\right) + 2\omega\sigma_\infty^2\lambda^h\left(\frac{\lambda(\lambda^h - (1+\gamma)^h)}{\lambda^h(\lambda-1-\gamma)}\right) + \sigma_t^4\gamma^h \tag{22.9}$$

where σ_∞^2 denotes the long-term expected GARCH variance as $h \to \infty$, i.e.

$$\sigma_\infty^4 = \lim_{h\to\infty} \mathbb{E}\{\sigma_{t+h}^4 \mid \mathcal{F}_{t-1}\} = \frac{\omega^2(1+\alpha_1\mathbb{E}\{u_t^2\}+\beta_1)}{(1-\alpha_1\mathbb{E}\{u_t^2\}-\beta_1)(1-\alpha_1^2\mathbb{E}\{u_t^4\}-2\alpha_1\beta_1\mathbb{E}\{u_t^2\}-\beta_1^2)}. \tag{22.10}$$

It follows that, under normality $u_t \sim \mathcal{N}(0,1)$ we have that $\mathbb{E}\{u_t^2\} = 1$ and $\mathbb{E}\{u_t^4\} = 3$. Substituting, one obtain the same result as in Bollerslev (1986), i.e.

$$\sigma_\infty^4 = \frac{\omega^2(1+\alpha_1+\beta_1)}{(1-\alpha_1-\beta_1)(1-3\alpha_1^2-2\alpha_1\beta_1-\beta_1^2)}.$$

i GARCH(1,1) long-term second moment

Proof. Under the assumption of constant second and fourth moments of u_t one can simplify the expressions, i.e.

$$\begin{aligned}
\gamma &= \alpha_1^2\mathbb{E}\{u_t^4\} + \beta_1(2\alpha_1\mathbb{E}\{u_t^2\} + \beta_1) \\
b_{t+h-i} &= \omega(\omega + 2\lambda\mathbb{E}\{\sigma_{t+h-i}^2 \mid \mathcal{F}_{t-1}\}) \\
\lambda &= \alpha_1\mathbb{E}\{u_t^2\} + \beta_1
\end{aligned}$$

Recalling the expression of the expectation of the GARCH variance with constant moments in Equation 22.3 one can write

$$b_{t+h-i} = \omega^2 + 2\omega\lambda\sigma_\infty^2 + 2\omega\lambda^{h-i-1}\sigma_\infty^2$$

Substituting the above expression into Equation 22.6 one obtain

$$\begin{aligned}
\mathbb{E}\{\sigma_{t+h}^4 \mid \mathcal{F}_{t-1}\} &= \sum_{i=1}^{h}(\omega^2 + 2\omega\lambda\sigma_\infty^2 + 2\omega\lambda^{h-i-1}\sigma_\infty^2)\gamma^{i-1} + \sigma_t^4\gamma^h \\
&= (\omega^2 + 2\omega\lambda\sigma_\infty^2)\sum_{i=1}^{h}\gamma^{i-1} + 2\omega\sigma_\infty^2\lambda^h\sum_{i=1}^{h}\left(\frac{1}{\lambda}\right)^{i-1}\gamma^{i-1} + \sigma_t^4\gamma^h = \\
&= (\omega^2 + 2\omega\lambda\sigma_\infty^2)\sum_{i=0}^{h-1}\gamma^{i} + 2\omega\sigma_\infty^2\lambda^h\sum_{i=0}^{h-1}\left(\frac{1+\gamma}{\lambda}\right)^{i} + \sigma_t^4\gamma^h
\end{aligned}$$

Notably

$$\sum_{i=0}^{h-1} \gamma^i = \frac{1-\gamma^h}{1-\gamma}$$

and

$$\sum_{i=0}^{h-1} \left(\frac{1+\gamma}{\lambda}\right)^i = \frac{1-\frac{(1+\gamma)^h}{\lambda^h}}{1-\frac{1+\gamma}{\lambda}} = \frac{\lambda(\lambda^h-(1+\gamma)^h)}{\lambda^h(\lambda-1-\gamma)}$$

Hence,

$$\mathbb{E}\{\sigma_{t+h}^4 \mid \mathcal{F}_{t-1}\} = (\omega^2 + 2\omega\lambda\sigma_\infty^2)\left(\frac{1-\gamma^h}{1-\gamma}\right) + 2\omega\sigma_\infty^2\lambda^h\left(\frac{\lambda(\lambda^h-(1+\gamma)^h)}{\lambda^h(\lambda-1-\gamma)}\right) + \sigma_t^4\gamma^h$$

Taking the limit as $h \to infty$, the second and third terms converges to zero if $\lambda < 1$, therefore

$$\lim_{h\to\infty} \mathbb{E}\{\sigma_{t+h}^4 \mid \mathcal{F}_{t-1}\} = \frac{\omega^2 + 2\omega\lambda\sigma_\infty^2}{1-\gamma}$$

More explicitly,

$$\begin{aligned}\lim_{h\to\infty} \mathbb{E}\{\sigma_{t+h}^4 \mid \mathcal{F}_{t-1}\} &= \frac{\omega^2(1+2\frac{\lambda}{1-\lambda})}{1-\alpha_1^2\mathbb{E}\{u_t^4\}-2\alpha_1\beta_1\mathbb{E}\{u_t^2\}-\beta_1^2} = \\ &= \frac{\omega^2(1+\lambda)}{(1-\lambda)(1-\alpha_1^2\mathbb{E}\{u_t^4\}-2\alpha_1\beta_1\mathbb{E}\{u_t^2\}-\beta_1^2)} = \\ &= \frac{\omega^2(1+\alpha_1\mathbb{E}\{u_t^2\}+\beta_1)}{(1-\alpha_1\mathbb{E}\{u_t^2\}-\beta_1)(1-\alpha_1^2\mathbb{E}\{u_t^4\}-2\alpha_1\beta_1\mathbb{E}\{u_t^2\}-\beta_1^2)}\end{aligned}$$

Moment	Formula	MonteCarlo	Difference
σ_∞^4	1.648437	1.648712	-0.0167%

Table 22.4: Forecasted long-term second moment of GARCH(1,1) variance with formula and by Monte Carlo simulations.

□

22.3 Variance σ_t^2

22.3.1 Short term

The conditional variance of σ_t^2 with $h \geq 1$ reads

$$\mathbb{V}\{\sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\} = \mathbb{E}\{\sigma_{t+h}^4 \mid \mathcal{F}_{t-1}\} - (\mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\})^2.$$

Example: GARCH(1,1) iterative variance

Example 22.4.

Moment	step	Formula	Iteration	Difference
$\mathbb{V}\{\sigma_{t+1}^2 \mid \mathcal{F}_{t-1}\}$	1	0.0352420	0.0352420	0%
$\mathbb{V}\{\sigma_{t+2}^2 \mid \mathcal{F}_{t-1}\}$	2	0.0455820	0.0455820	0%
$\mathbb{V}\{\sigma_{t+3}^2 \mid \mathcal{F}_{t-1}\}$	3	0.0489759	0.0489759	0%
$\mathbb{V}\{\sigma_{t+4}^2 \mid \mathcal{F}_{t-1}\}$	4	0.0502199	0.0502199	0%
$\mathbb{V}\{\sigma_{t+5}^2 \mid \mathcal{F}_{t-1}\}$	5	0.0507180	0.0507180	0%
$\mathbb{V}\{\sigma_{t+6}^2 \mid \mathcal{F}_{t-1}\}$	6	0.0509296	0.0509296	0%
$\mathbb{V}\{\sigma_{t+7}^2 \mid \mathcal{F}_{t-1}\}$	7	0.0510227	0.0510227	0%
$\mathbb{V}\{\sigma_{t+8}^2 \mid \mathcal{F}_{t-1}\}$	8	0.0510646	0.0510646	0%
$\mathbb{V}\{\sigma_{t+9}^2 \mid \mathcal{F}_{t-1}\}$	9	0.0510835	0.0510835	0%
$\mathbb{V}\{\sigma_{t+10}^2 \mid \mathcal{F}_{t-1}\}$	10	0.0510922	0.0510922	0%

Table 22.5: Forecasted variance of GARCH(1,1) variance with iteration and with formula.

22.3.2 Long term

The long-term variance of σ_t^2 reads explicitely

$$\mathbb{V}\{\sigma_t^2\} = \sigma_\infty^4 - (\sigma_\infty^2)^2.$$

Example: GARCH(1,1) long-term variance

Example 22.5.

Moment	Formula	MonteCarlo	Difference
$\mathbb{V}\{\sigma_\infty^2\}$	0.0510995	0.0511939	-0.1847%

Table 22.6: Forecasted long-term variance of GARCH(1,1) variance with formula and by Monte Carlo simulations.

22.4 First moment σ_t

22.4.1 Short term

The expected value of the GARCH std. deviation can be approximated as $\sigma_{t+h} = (\sigma_{t+h}^2)^{1/2}$ with a Taylor expansion

$$\mathbb{E}\{\sigma_{t+h} \mid \mathcal{F}_{t-1}\} \approx \sqrt{\mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\}} + \frac{1}{8}\frac{\mathbb{V}\{\sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\}}{\mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\}^{\frac{3}{2}}}.$$

Approximated GARCH(1,1) std. deviation with Taylor expansion.

Proof. Let's X be a non-negative random variable, and let's say one want to approximate:

$$\mathbb{E}\{\sqrt{X}\}.$$

Let $f(x) = \sqrt{x}$ and expand $f(x)$ around the point $\mu = \mathbb{E}\{X\}$, i.e.

$$\sqrt{X} \approx \sqrt{\mu} + \frac{1}{2}\mu^{-1/2}(X-\mu) - \frac{1}{8}\mu^{-3/2}(X-\mu)^2,$$

and take the expectation

$$\mathbb{E}\{\sqrt{X}\} \approx \sqrt{\mu} + \frac{1}{2}\mu^{-1/2}\left(\mathbb{E}\{X\} - \mu\right) - \frac{1}{8}\mu^{-3/2}\mathbb{E}\{(X-\mu)^2\},$$

where $\mathbb{E}\{X\} - \mu = 0$ and $\mathbb{E}\{(X-\mu)^2\} = \mathbb{V}\{X\}$. Applying this result to the random variable σ_{t+h}^2 with $1 < h$ and let's approximate around the expected value with a Taylor expansion, i.e.

$$\sigma_{t+h} = \sqrt{\sigma_{t+h}^2} \approx \sqrt{\mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_t\}} + \frac{1}{2}\frac{(\sigma_{t+h}^2 - \mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_t\})}{\sqrt{\mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_t\}}} - \frac{1}{8}\frac{(\sigma_{t+h}^2 - \mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_t\})^2}{\sqrt{\mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_t\}^3}},$$

Taking the expectation gives the result.

$$\mathbb{E}\{\sigma_{t+h} \mid \mathcal{F}_t\} \approx \sqrt{\mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_t\}} - \frac{1}{8}\frac{\mathbb{V}\{\sigma_{t+h}^2 \mid \mathcal{F}_t\}}{\sqrt{\mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_t\}^3}}.$$

□

Example: GARCH(1,1) std.deviation iterative expectation.

Example 22.6.

Moment	step	Formula	Iteration	Difference
$\mathbb{E}\{\sigma_{t+0} \mid \mathcal{F}_{t-1}\}$	0	1.095445	1.095445	0%
$\mathbb{E}\{\sigma_{t+1} \mid \mathcal{F}_{t-1}\}$	1	1.107896	1.107896	0%
$\mathbb{E}\{\sigma_{t+2} \mid \mathcal{F}_{t-1}\}$	2	1.114145	1.114145	0%
$\mathbb{E}\{\sigma_{t+3} \mid \mathcal{F}_{t-1}\}$	3	1.117128	1.117128	0%
$\mathbb{E}\{\sigma_{t+4} \mid \mathcal{F}_{t-1}\}$	4	1.118522	1.118522	0%
$\mathbb{E}\{\sigma_{t+5} \mid \mathcal{F}_{t-1}\}$	5	1.119168	1.119168	0%
$\mathbb{E}\{\sigma_{t+6} \mid \mathcal{F}_{t-1}\}$	6	1.119466	1.119466	0%
$\mathbb{E}\{\sigma_{t+7} \mid \mathcal{F}_{t-1}\}$	7	1.119603	1.119603	0%
$\mathbb{E}\{\sigma_{t+8} \mid \mathcal{F}_{t-1}\}$	8	1.119666	1.119666	0%
$\mathbb{E}\{\sigma_{t+9} \mid \mathcal{F}_{t-1}\}$	9	1.119694	1.119694	0%
$\mathbb{E}\{\sigma_{t+10} \mid \mathcal{F}_{t-1}\}$	10	1.119708	1.119708	0%

Table 22.7: Forecasted expectation of GARCH(1,1) std. deviation with iteration and with formula.

22.4.2 Long term

The unconditional expected GARCH(1,1) std. deviation

$$\sigma_\infty \approx \sqrt{\sigma_\infty^2} + \frac{1}{8}\frac{\mathbb{V}\{\sigma_\infty^2\}}{(\sigma_\infty^2)^{\frac{3}{2}}}.$$

Example: GARCH(1,1) std. deviation long-term expectation.

Example 22.7.

Moment	Formula	MonteCarlo	Difference
σ_∞	1.119719	1.120497	-0.0695%

Table 22.8: Long-term expectation of GARCH(1,1) std. deviation with approximated formula and by Monte Carlo simulations.

22.5 Variance σ_t

22.5.1 Short term

The variance of the GARCH std. deviation can be approximated

$$\mathbb{V}\{\sigma_{t+h} \mid \mathcal{F}_{t-1}\} \approx \mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\} - \mathbb{E}\{\sigma_{t+h} \mid \mathcal{F}_{t-1}\}^2.$$

Example: GARCH(1,1) std. deviation iterative variance

Example 22.8.

Moment	step	Formula	Iteration	Difference
$\mathbb{V}\{\sigma_{t+1} \mid \mathcal{F}_{t-1}\}$	1	0.0071262	0.0071262	0%
$\mathbb{V}\{\sigma_{t+2} \mid \mathcal{F}_{t-1}\}$	2	0.0090968	0.0090968	0%
$\mathbb{V}\{\sigma_{t+3} \mid \mathcal{F}_{t-1}\}$	3	0.0097164	0.0097164	0%
$\mathbb{V}\{\sigma_{t+4} \mid \mathcal{F}_{t-1}\}$	4	0.0099365	0.0099365	0%
$\mathbb{V}\{\sigma_{t+5} \mid \mathcal{F}_{t-1}\}$	5	0.0100227	0.0100227	0%
$\mathbb{V}\{\sigma_{t+6} \mid \mathcal{F}_{t-1}\}$	6	0.0100589	0.0100589	0%
$\mathbb{V}\{\sigma_{t+7} \mid \mathcal{F}_{t-1}\}$	7	0.0100747	0.0100747	0%
$\mathbb{V}\{\sigma_{t+8} \mid \mathcal{F}_{t-1}\}$	8	0.0100817	0.0100817	0%
$\mathbb{V}\{\sigma_{t+9} \mid \mathcal{F}_{t-1}\}$	9	0.0100849	0.0100849	0%
$\mathbb{V}\{\sigma_{t+10} \mid \mathcal{F}_{t-1}\}$	10	0.0100864	0.0100864	0%

Table 22.9: Forecasted variance of GARCH(1,1) std. deviation with iteration and with formula.

22.5.2 Long term

The long-term variance of the GARCH std. deviation can be approximated as

$$\mathbb{V}\{\sigma_\infty\} \approx \sigma_\infty^2 - (\sigma_\infty)^2.$$

Example: GARCH(1,1) std. deviation long-term variance

Example 22.9.

Moment	Formula	MonteCarlo	Difference
$\mathbb{V}\{\sigma_\infty\}$	0.0079976	0.008414	-5.2065%

Table 22.10: Long-term variance of GARCH(1,1) std. deviation with approximated formula and by Monte Carlo simulations.

22.6 Third moment σ_t

22.6.1 Short term

The expected value of σ_{t+h}^3 can be approximated with a Taylor expansion, i.e. $\sigma_{t+h}^3 = (\sigma_{t+h}^2)^{3/2}$. Then, we can approximate

$$\mathbb{E}\{\sigma_t^3 \mid \mathcal{F}_{t-1}\} \approx \mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\}^{\frac{3}{2}} + \frac{3}{8}\frac{\mathbb{V}\{\sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\}}{\sqrt{\mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\}}}.$$

Example: GARCH(1,1) std. deviation iterative third moment.

Example 22.10.

Moment	step	Formula	Iteration	Difference
$\mathbb{E}\{\sigma_{t+0}^{\frac{3}{2}} \mid \mathcal{F}_{t-1}\}$	0	1.314534	1.314534	0%
$\mathbb{E}\{\sigma_{t+1}^{\frac{3}{2}} \mid \mathcal{F}_{t-1}\}$	1	1.383623	1.383623	0%
$\mathbb{E}\{\sigma_{t+2}^{\frac{3}{2}} \mid \mathcal{F}_{t-1}\}$	2	1.413526	1.413526	0%
$\mathbb{E}\{\sigma_{t+3}^{\frac{3}{2}} \mid \mathcal{F}_{t-1}\}$	3	1.426837	1.426837	0%
$\mathbb{E}\{\sigma_{t+4}^{\frac{3}{2}} \mid \mathcal{F}_{t-1}\}$	4	1.432849	1.432849	0%
$\mathbb{E}\{\sigma_{t+5}^{\frac{3}{2}} \mid \mathcal{F}_{t-1}\}$	5	1.435585	1.435585	0%
$\mathbb{E}\{\sigma_{t+6}^{\frac{3}{2}} \mid \mathcal{F}_{t-1}\}$	6	1.436836	1.436836	0%
$\mathbb{E}\{\sigma_{t+7}^{\frac{3}{2}} \mid \mathcal{F}_{t-1}\}$	7	1.437408	1.437408	0%
$\mathbb{E}\{\sigma_{t+8}^{\frac{3}{2}} \mid \mathcal{F}_{t-1}\}$	8	1.437670	1.437670	0%
$\mathbb{E}\{\sigma_{t+9}^{\frac{3}{2}} \mid \mathcal{F}_{t-1}\}$	9	1.437791	1.437791	0%
$\mathbb{E}\{\sigma_{t+10}^{\frac{3}{2}} \mid \mathcal{F}_{t-1}\}$	10	1.437846	1.437846	0%

Table 22.11: Forecasted third moment of GARCH(1,1) std. deviation with iteration and with formula.

22.6.2 Long term

With a Taylor approximation, the long term third moment of the GARCH std. deviation reads

$$\sigma_\infty^3 \approx (\sigma_\infty^2)^{\frac{3}{2}} + \frac{3}{8}\frac{\mathbb{V}\{\sigma_\infty^2\}}{\sqrt{\sigma_\infty^2}}.$$

Example: GARCH(1,1) std. deviation long-term third moment.

Example 22.11.

Moment	Formula	MonteCarlo	Difference
$\sigma_\infty^{\frac{3}{2}}$	1.437893	1.436906	0.0686%

Table 22.12: Long-term third moment of GARCH(1,1) std. deviation with approximated formula and by Monte Carlo simulations.

22.7 Covariance

The covariance between two GARCH variances at time t and $t+h$ reads:

$$\mathbb{C}v\{\sigma_t^2 \cdot \sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\} = \left(\prod_{i=1}^{h} \lambda_{t+h-i}\right) \mathbb{V}\{\sigma_t^2 \mid \mathcal{F}_{t-1}\}$$

For a fixed t and general s and h,

$$\mathbb{C}v\{\sigma_{t+s}^2 \cdot \sigma_{t+h}^2 \mid \mathcal{F}_t\} = \left(\prod_{i=1}^{\max(s,h)} \lambda_{t+\max(s,h)-i}\right) \mathbb{V}\{\sigma_{\min(s,h)}^2 \mid \mathcal{F}_t\}$$

Proof: Iterative formula for the covariance between GARCH(1,1) variances

Proof. Leveraging the tower property and conditioning one can write:

$$\mathbb{E}\{\sigma_t^2 \cdot \sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\} = \mathbb{E}\{\sigma_t^2 \cdot \mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_t\} \mid \mathcal{F}_{t-1}\}$$

From our previous result we have:

$$\mathbb{E}\{\sigma_t^2 \cdot \sigma_{t+h}^2\} = \mathbb{E}\left\{\sigma_t^2 \cdot \left(\omega \sum_{j=0}^{h-1} \prod_{i=1}^{j} \lambda_{t+h-i} + \prod_{i=1}^{h} \lambda_{t+h-i} \cdot \sigma_t^2\right) \mid \mathcal{F}_{t-1}\right\}$$

Hence,

$$\mathbb{E}\{\sigma_t^2 \cdot \sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\} = \left(\omega \sum_{j=0}^{h-1} \prod_{i=1}^{j} \lambda_{t+h-i}\right) \mathbb{E}\{\sigma_t^2 \mid \mathcal{F}_{t-1}\} + \left(\prod_{i=1}^{h} \lambda_{t+h-i}\right) \mathbb{E}\{\sigma_t^4 \mid \mathcal{F}_{t-1}\}$$

By definition the covariance:

$$\begin{aligned}\mathbb{C}v\{\sigma_t^2 \cdot \sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\} &= \mathbb{E}\{\sigma_t^2 \cdot \sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\} - \mathbb{E}\{\sigma_t^2\} \cdot \mathbb{E}\{\sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\} = \\ &= \left(\omega \sum_{j=0}^{h-1} \prod_{i=1}^{j} \lambda_{t+h-i}\right) \mathbb{E}\{\sigma_t^2\} + \left(\prod_{i=1}^{h} \lambda_{t+h-i}\right) \mathbb{E}\{\sigma_t^4\} - \mathbb{E}\{\sigma_t^2\} \left[\omega \sum_{j=0}^{h-1} \prod_{i=1}^{j} \lambda_{t+h-i} + \prod_{i=1}^{h} \lambda_{t+h-}\right.\end{aligned}$$

and simplify the first and last terms cancel and the one remain with

$$\begin{aligned}\mathbb{C}v\{\sigma_t^2 \cdot \sigma_{t+h}^2 \mid \mathcal{F}_{t-1}\} &= \left(\prod_{i=1}^{h} \lambda_{t+h-i}\right) \mathbb{E}\{\sigma_t^4 \mid \mathcal{F}_{t-1}\} - \left(\prod_{i=1}^{h} \lambda_{t+h-i}\right) \mathbb{E}\{\sigma_t^2 \mid \mathcal{F}_{t-1}\}^2 = \\ &= \left(\prod_{i=1}^{h} \lambda_{t+h-i}\right) \mathbb{V}\{\sigma_t^2 \mid \mathcal{F}_{t-1}\}\end{aligned}$$

□

Example: GARCH(1,1) covariance between variances.

Example 22.12.

Moment	step	Formula	MonteCarlo	Difference
$\mathbb{C}v\{\sigma_{t+0}^2, \sigma_{t+0}^2 \mid \mathcal{F}_{t-1}\}$	0	0.0234411	0.0233719	0.2952%
$\mathbb{C}v\{\sigma_{t+1}^2, \sigma_{t+1}^2 \mid \mathcal{F}_{t-1}\}$	1	0.0107530	0.0106399	1.0516%
$\mathbb{C}v\{\sigma_{t+2}^2, \sigma_{t+2}^2 \mid \mathcal{F}_{t-1}\}$	2	0.0049316	0.0048333	1.9931%
$\mathbb{C}v\{\sigma_{t+3}^2, \sigma_{t+3}^2 \mid \mathcal{F}_{t-1}\}$	3	0.0022608	0.0021966	2.8387%
$\mathbb{C}v\{\sigma_{t+4}^2, \sigma_{t+4}^2 \mid \mathcal{F}_{t-1}\}$	4	0.0010353	0.0009730	6.0243%
$\mathbb{C}v\{\sigma_{t+5}^2, \sigma_{t+5}^2 \mid \mathcal{F}_{t-1}\}$	5	0.0004730	0.0003460	26.8662%
$\mathbb{C}v\{\sigma_{t+6}^2, \sigma_{t+6}^2 \mid \mathcal{F}_{t-1}\}$	6	0.0002149	0.0001076	49.9429%
$\mathbb{C}v\{\sigma_{t+7}^2, \sigma_{t+7}^2 \mid \mathcal{F}_{t-1}\}$	7	0.0000962	0.0000169	82.3939%
$\mathbb{C}v\{\sigma_{t+8}^2, \sigma_{t+8}^2 \mid \mathcal{F}_{t-1}\}$	8	0.0000411	0.0000221	46.2582%
$\mathbb{C}v\{\sigma_{t+9}^2, \sigma_{t+9}^2 \mid \mathcal{F}_{t-1}\}$	9	0.0000146	0.0000417	-186.6511%
$\mathbb{C}v\{\sigma_{t+10}^2, \sigma_{t+10}^2 \mid \mathcal{F}_{t-1}\}$	10	0.0000000	0.0000382	-Inf%

Table 22.13: Forecasted covariance between GARCH(1,1) variances with formula and by monte Carlo simulations

For the covariance between the GARCH std. deviations, we apply a Taylor approximation around the mean of σ_{t+s} and σ_{t+h}, using the delta method:

$$\mathbb{C}v\{\sigma_{t+s}, \sigma_{t+h} \mid \mathcal{F}_t\} \approx \frac{\mathbb{C}v\{\sigma^2_{t+s}, \sigma^2_{t+h} \mid \mathcal{F}_t\}}{4\sqrt{\mathbb{E}\{\sigma^2_{t+s} \mid \mathcal{F}_t\}\mathbb{E}\{\sigma^2_{t+h} \mid \mathcal{F}_t\}}}.$$

Approximated GARCH(1,1) covariance with Taylor expansion.

Proof. Considering the product of

$$\sigma_{t+h} = \sqrt{\sigma^2_{t+h}} \approx \sqrt{\mathbb{E}\{\sigma^2_{t+h} \mid \mathcal{F}_t\}} + \frac{1}{2}\frac{(\sigma^2_{t+h} - \mathbb{E}\{\sigma^2_{t+h} \mid \mathcal{F}_t\})}{\sqrt{\mathbb{E}\{\sigma^2_{t+h} \mid \mathcal{F}_t\}}},$$

and applying the covariance between σ_{t+h} and σ_{t+s} with $1 < s < h$ one obtain the result. □

Example: GARCH(1,1) covariance between std. deviations

Example 22.13.

Moment	step	Formula	MonteCarlo	Difference
$\mathbb{C}v\{\sigma_{t+0}, \sigma_{t+0} \mid \mathcal{F}_{t-1}\}$	0	0.0040696	0.0039547	2.8232%
$\mathbb{C}v\{\sigma_{t+1}, \sigma_{t+1} \mid \mathcal{F}_{t-1}\}$	1	0.0018146	0.0018278	-0.7272%
$\mathbb{C}v\{\sigma_{t+2}, \sigma_{t+2} \mid \mathcal{F}_{t-1}\}$	2	0.0008217	0.0008375	-1.9265%
$\mathbb{C}v\{\sigma_{t+3}, \sigma_{t+3} \mid \mathcal{F}_{t-1}\}$	3	0.0003745	0.0003818	-1.9551%
$\mathbb{C}v\{\sigma_{t+4}, \sigma_{t+4} \mid \mathcal{F}_{t-1}\}$	4	0.0001710	0.0001698	0.7476%
$\mathbb{C}v\{\sigma_{t+5}, \sigma_{t+5} \mid \mathcal{F}_{t-1}\}$	5	0.0000781	0.0000610	21.8277%
$\mathbb{C}v\{\sigma_{t+6}, \sigma_{t+6} \mid \mathcal{F}_{t-1}\}$	6	0.0000354	0.0000191	46.248%
$\mathbb{C}v\{\sigma_{t+7}, \sigma_{t+7} \mid \mathcal{F}_{t-1}\}$	7	0.0000159	0.0000023	85.7078%
$\mathbb{C}v\{\sigma_{t+8}, \sigma_{t+8} \mid \mathcal{F}_{t-1}\}$	8	0.0000068	0.0000027	60.4525%
$\mathbb{C}v\{\sigma_{t+9}, \sigma_{t+9} \mid \mathcal{F}_{t-1}\}$	9	0.0000024	0.0000058	-142.9085%
$\mathbb{C}v\{\sigma_{t+10}, \sigma_{t+10} \mid \mathcal{F}_{t-1}\}$	10	0.0000000	0.0000057	-Inf%

Table 22.14: Forecasted covariance between GARCH(1,1) std. deviation with formula and by monte Carlo simulations

Part V

Tests

23 Hypothesis tests

A **statistical hypothesis** is a claim about the value of a parameter or population characteristic. In any hypothesis-testing problem, there are always two competing hypotheses under consideration

1. The *null hypothesis* $\mathcal{H}_0$ representing the status quo.
2. The alternative hypothesis $\mathcal{H}_1$ representing the research.

The objective of hypothesis testing is to decide, based on sample information, if the alternative hypotheses is actually supported by the data. One usually do new research to challenge the existing beliefs.

> ⚠ Is there strong evidence for the alternative?
>
> Let's consider that you want to establish if the *null hypothesis* $\mathcal{H}_0$ is not supported by the data. One usually assume to work under $\mathcal{H}_0$, then if the sample does not strongly contradict H0, we will continue to believe in the plausibility of the null hypothesis. There are only two possible conclusions: **Reject** $\mathcal{H}_0$ or **Fail to reject** $\mathcal{H}_0$.

Definition 23.1. The **test statistic** $T(X_n)$ is a function of a sample X_n and is used to make a decision about whether the null hypothesis should be rejected or not. In theory, there are an infinite number of possible tests that could be devised. The choice of a particular test procedure must be based on the probability the test will produce incorrect results. In general, two kind of errors are related with test statistics, i.e.

1. A **type I error** is when the null hypothesis is rejected, but it is true.
2. A **type II error** is not rejecting the null when it is false.

The **p-value** is in general related to the probability of the type I error. So, the smaller the P-value, the more evidence there is in the sample data against the null hypothesis and for the alternative hypothesis.

In general, before performing a test one establish a significance level α (the desired type I error probability), that defines the rejection region. Then the decision rule is:

$$\begin{aligned} \text{Reject } \mathcal{H}_0 &\iff \text{p-value } \leq \alpha \\ \text{Do not reject } \mathcal{H}_0 &\iff \text{p-value } > \alpha \end{aligned}$$

The p-value can be thought of as the smallest significance level at which $\mathcal{H}_0$ can be rejected and the calculation of the P-value depends on whether the test is upper, lower, or two-tailed.

For example, let's consider a sample X_n of data. Then, a statistical test consists of the following:

1. an assumption about the distribution of the data, often expressed in terms of a statistical model $\mathcal{M}$;
2. a null hypothesis H_0 and an alternative hypothesis H_1 which make specific statements about the data;
3. a test statistic $T(X_n)$ which is a function of the data and whose distribution under the null hypothesis is known;
4. a significance level α which imposes an upper bound on the probability of rejecting H_0, given that H_0 is true.

The general procedure for a statistical hypothesis test can be summarized as follows:

1. **Inputs**: consider a null hypothesis H_0 and the significance level α.
2. **Critical value**: compute the value t_α that determine the partitions the set of possible values of $T(X_n)$ into *rejection* and *non rejection* regions.
3. **Output**: compare the observed test statistic $T(X_n)$ computed on the sample with the critical value t_α. If it is in the rejection region, H_0 is rejected in favor of H_1. Otherwise, the test fails to reject H_0.

Step	Description
Inputs	H_0, α.
Critical value	Critical level $t(\alpha)$
Output	Rejection or not depending on $T(X_n)$

In general, two kind of tests are available:

- A **two-tailed test** is appropriate if the estimated value is greater or less than a certain range of values, for example, whether a test taker may score above or below a specific range of scores.
- A **one-tailed test** is appropriate if the estimated value may depart from the reference value in only one direction, left or right, but not both.

23.1 Left and right tailed tests

For example, let's simulate a sample X_n of $n = 500$ observations from a normal distribution (i.e. $X_n \sim \mathcal{N}(2, 4^2)$) and consider the following sets of hypothesis, i.e.

$$\mathcal{H}_0 : \mu(X) = 2.4 \qquad \mathcal{H}_1 : \mu(X) \neq 2.4$$

The statistic test is defined as

$$T(X_n) = \sqrt{500} \cdot \frac{\mu(X_n) - 2.4}{\sigma(X_n)} \overset{\mathcal{H}_0}{\sim} t(499).$$

Since it is a **two-tailed test** the critical value for a significance level α, denoted as t_α, is such that:

$$\begin{aligned} &\alpha = \mathbb{P}([T(X_n) < -t_{\alpha/2}] \cup [T(X_n) > t_{\alpha/2}]) \\ &\Updownarrow \\ &t_{\alpha/2} = \mathbb{P}^{-1}(\mathbb{P}(T(X_n) > t_{\alpha/2})), \end{aligned}$$

where $\mathbb{P}^{-1}$ and $\mathbb{P}$ are respectively the quantile and distribution functions of a Student-t. If the statistic test $|T(X_n)| > |t_{\alpha/2}|$, then we reject $\mathcal{H}_0$ and so the mean of the sample is significantly different from 2.4. More precisely, with $\alpha = 0.05$, the critical value of a Student-t with 499 degrees of freedom is $t_{\alpha/2} = 1.9604$.

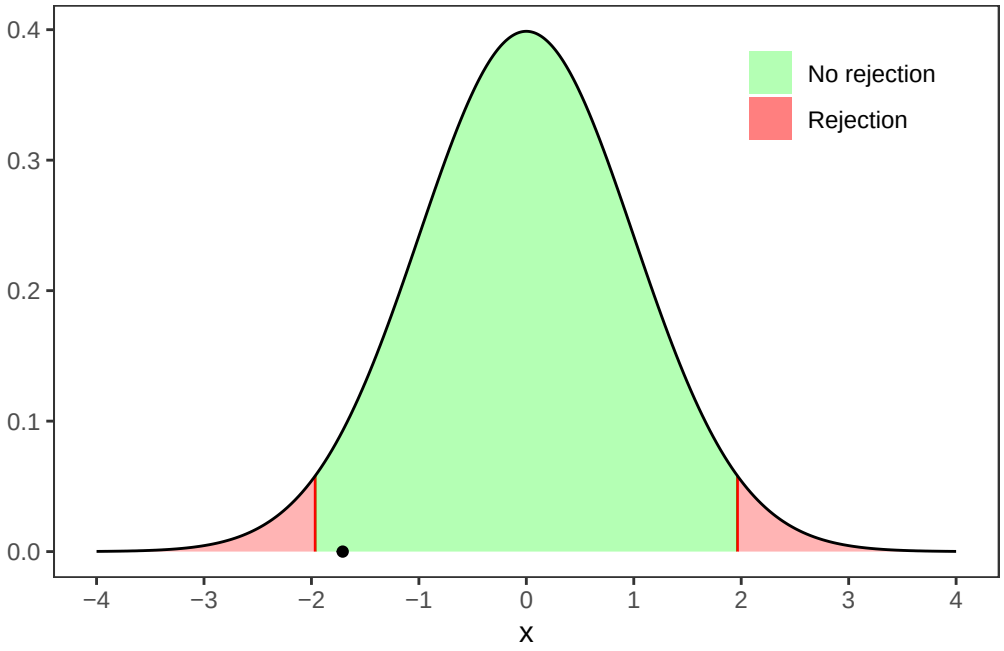


Figure 23.1: Two-tailed test on the mean.

Let's consider another kind of hypothesis,

$$\mathcal{H}_0 : \mu(X) \geq 2.4 \qquad \mathcal{H}_1 : \mu(X) < 2.4$$

The statistic test $T(X_n)$ do not changes, however the null hypothesis implies a **left-tailed test**. Hence, the critical value is t_α is such that $\mathbb{P}(x < t_\alpha) = 0.05$. Applying the quantile function $\mathbb{P}^{-1}$ of a student-t we obtain:

$$\begin{aligned} &\alpha = \mathbb{P}(T(X_n) < t_\alpha) \\ &\Updownarrow \\ &t_\alpha = \mathbb{P}^{-1}(\mathbb{P}(T(X_n) < t_\alpha)), \end{aligned}$$

where $\mathbb{P}^{-1}$ and $\mathbb{P}$ are respectively the quantile and distribution functions of a Student-t. In this case, with $\alpha = 0.05$, the critical value of a Student-t with 499 degrees of freedom is $t_{\alpha/2} = -1.6451$. Therefore, if $T(X_n) < -1.6451$ we do not reject the null hypothesis, i.e. $\mu(X_n)$ is greater than μ_0, otherwise we reject it and $\mu(X_n)$ is lower than μ_0.

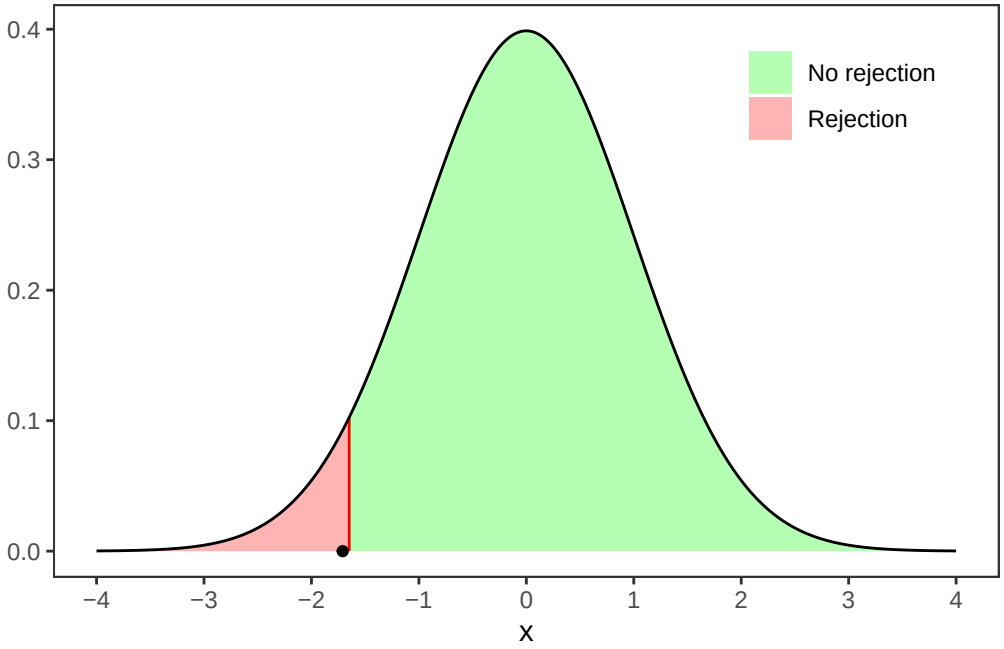


Figure 23.2: Left-tailed test on the mean.

In this case we reject the null hypothesis, hence $\mu(X_n)$ is lower than μ_0.

Lastly, let's consider the **right-tailed** case, i.e.

$$H_0 : \mu(X) \leq 2.4 \qquad H_1 : \mu(X) > 2.4$$

It is always one-side test, but in this case is **right-tailed**. Hence, the critical value t_α is such that

$$\begin{aligned} 1-\alpha &= \mathbb{P}(T(X_n) < t_\alpha) \\ &\Updownarrow \\ t_\alpha &= \mathbb{P}^{-1}(\mathbb{P}(T(X_n) < t_\alpha)), \end{aligned}$$

where $\mathbb{P}^{-1}$ and $\mathbb{P}$ are respectively the quantile and distribution functions of a Student-t. In this case, with $\alpha = 0.05$, the critical value of a Student-t with 499 degrees of freedom is $t_{\alpha/2} = 1.6451$. Therefore, if $T(X_n) < 1.6451$ we do not reject the null hypothesis, i.e. $\mu(X_n)$ is lower than μ_0, otherwise we reject it and $\mu(X_n)$ is greater than μ_0. Coherently with the previous test performed in Figure 23.2, a right railed test is not rejected in Figure 23.3, hence $\mu(X_n)$ is lower than $\mu_0 = 2.4$.

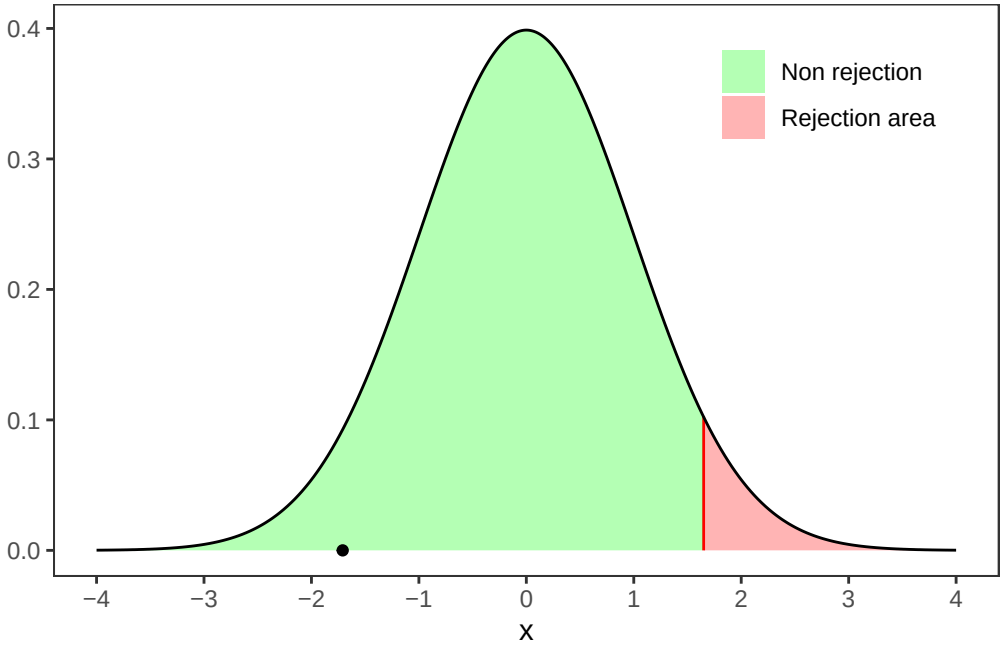


Figure 23.3: Right-tailed test on the mean.

23.2 Tests for the means

Proposition 23.1. *Let's consider the t-test for the mean of a sample of identically and normally distributed random variables $X_n = (x_1, \dots, x_i, \dots, x_n)$. Then the test statistic $T(X_n)$ under $\mathcal{H}_0 : \hat{\mu}(X_n) = \mu_0$ is student-t distributed with $n-1$ degrees of freedom., i.e.*

$$T(X_n) = \frac{\hat{\mu}(X_n) - \mu_0}{\frac{\hat{s}(X_n)}{\sqrt{n}}} \overset{\mathcal{H}_0}{\sim} t_{n-1},$$

where $\hat{\mu}(X_n)$ is the sample mean $\hat{\mu}(X_n)$ and $\hat{\sigma}(X_n)$ the corrected sample variance. Moreover, for $n \to \infty$:

$$T(X_n) \xrightarrow[n\to\infty]{\mathcal{H}_0} \mathcal{N}(0,1).$$

Proof: Proposition 23.1

Proof. In the sample is normally distributed, the sample mean is also normally distributed, i.e.

$$M = \sqrt{n}\frac{\hat{\mu}(X_n) - \mu_0}{\sigma} \sim \mathcal{N}(0,1).$$

Under normality the sample variance, that is a sum of the square of independent and normally distributed random variables, follows a χ^2 distribution with $n-1$ degrees of freedom, i.e.

$$V = \frac{(n-1)\hat{s}^2(X_n)}{\sigma^2} \sim \chi^2_{n-1}.$$

Notably, the ratio of a standard normal and a χ^2 random variables (each one divided by the respective degrees of freedom) is exactly the definition of a Student-t random variable as in Equation 33.2. Hence, the ratio between the statistics M and V divided by their degrees of freedom reads

$$\frac{M}{\sqrt{\frac{V}{n-1}}} = \sqrt{n}\frac{\hat{\mu}(X_n) - \mu_0}{\sigma}\sqrt{\frac{\sigma^2}{\hat{s}^2(X_n)}} = \sqrt{n}\frac{\hat{\mu}(X_n) - \mu_0}{\hat{s}^2(X_n)} \sim t_{n-1}.$$

The statistic test under H_0 follows a Student-t distribution with $n-1$ degrees of freedom. Notably, for large IID samples the statistic converges to a normal random variable independently from the distribution of X. □

23.2.1 Test for two means and equal variances

Let's consider two independent Gaussian populations with equal variance, i.e.

$$X_1 \sim \mathcal{N}(\mu_1, \sigma^2), \quad X_2 \sim \mathcal{N}(\mu_2, \sigma^2)$$

Then, let's consider two samples of unequal size, n_1 and n_2, with unknown means μ_1 and μ_2 and an equal unknown variance σ^2. Then, given the null hypothesis

$$H_0 = \mu_1 - \mu_2 = \mu_\Delta,$$

the test statistic

$$T(X_{n_1}, X_{n_2}) = \frac{\mu(X_{n_1}) - \mu(X_{n_2}) - \mu_\Delta}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2},$$

is Student-t distributed with $n_1 + n_2 - 2$ degrees of freedom and

$$s_p = \sqrt{\frac{(n_1 - 1)\hat{s}^2(X_{n_1}) + (n_2 - 1)\hat{s}^2(X_{n_2})}{n_1 + n_2 - 2}},$$

where $\hat{s}^2(X_{n_1})$ and $\hat{s}^2(X_{n_2})$ are the sample corrected variances (Equation 9.11) of the two samples.

23.2.2 Test for two means and unequal variances

Let's consider two independent Gaussian populations with different variance, i.e.

$$X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), \quad X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2).$$

Then, let's consider two samples of unequal size, n_1 and n_2, with unknown means μ_1 and μ_2 and an unequal unknown variance σ^2. Then, given the null hypothesis

$$H_0 = \mu_1 - \mu_2 = \mu_\Delta,$$

Welch (1938) - Welch (1947) proposes a test statistic

$$T(X_{n_1}, X_{n_2}) = \frac{\mu(X_{n_1}) - \mu(X_{n_2})}{\sqrt{\frac{\hat{s}^2(X_{n_1})}{n_1} + \frac{\hat{s}^2(X_{n_2})}{n_2}}} \approx t_\nu,$$

that follows approximately a Student t-distribution under the null hypothesis, but with fractional degrees of freedom computed using the Welch–Satterthwaite approximation. This is a weighted average of the degrees of freedom from each group, reflecting the uncertainty due to unequal variances, i.e.

$$\nu = \frac{\left(\frac{\hat{s}^2(X_{n_1})}{n_1} + \frac{\hat{s}^2(X_{n_1})}{n_2}\right)^2}{\frac{(\hat{s}^2(X_{n_1}))^2}{n_1^2(n_1-1)} + \frac{(\hat{s}^2(X_{n_1}))^2}{n_2^2(n_2-1)}}.$$

where ν is not necessary an integer.

23.3 Tests for the variances

23.3.1 F-test for two variances

Consider two independent normal samples, i.e.

$$X_{n_1} \sim \mathcal{N}(\mu_1, \sigma_1^2), \quad X_{n_2} \sim \mathcal{N}(\mu_2, \sigma_2^2),$$

where n_1 and n_2 are the number of observations in each sample.

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2.$$

Knowing that the sample variance is chi2 distributed (Equation 9.15) let's define the variables:

$$T_1 = (n_1 - 1)\frac{\hat{s}_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}, \quad T_2 = (n_2 - 1)\frac{\hat{s}_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}.$$

Then, since the ration of two independent χ^2 divided by their respective degrees of freedom is F-distributed (Equation 33.3) the statistic F is defined as:

$$T(X_{n_1}, X_{n_2}) = \frac{\frac{T_1}{n_1-1}}{\frac{T_2}{n_1-1}} = \frac{\frac{\hat{s}_1^2}{\sigma_1^2}}{\frac{\hat{s}_2^2}{\sigma_2^2}} = \frac{\hat{s}_1^2\sigma_2^2}{\hat{s}_2^2\sigma_1^2} \sim \mathrm{F}_{n_1-1,n_2,-1}$$

Under H_0 the two variances are assumed to be equal, i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$, thus the statistic simplifies in:

$$T(X_{n_1}, X_{n_2}) \overset{H_0}{=} \frac{\hat{s}_1^2}{\hat{s}_2^2} \sim \mathrm{F}_{n_1-1,n_2,-1}$$

This means that the null hypothesis of equal variances can be rejected when F is as extreme or more extreme than the critical value obtained from the F-distribution with degrees of freedom $n_1 - 1$ and $n_2 - 1$ using a significance level α.

24 Autocorrelation tests

24.1 Durbin-Watson test

The aim of the Durbin-Watson test is to verify if a time series presents autocorrelation or not. Specifically, let's consider a time series $X_t = (x_1, \dots, x_i, \dots, x_t)$, then evaluating an AR(1) model, i.e.

$$x_t = \phi_1 x_{t-1} + u_t \tag{24.1}$$

we would like to verify if ϕ_1 is significantly different from zero. The test statistic, denoted as DW, is computed as:

$$\text{DW} = \frac{\sum_{i=2}^{t}(x_i - x_{i-1})^2}{\sum_{i=2}^{t} x_{i-1}^2} \approx 2(1 - \phi_1)$$

The null hypothesis H_0 is the *absence of autocorrelation*, i.e.

$$H_0 : \phi_1 = 0 \quad H_1 : \phi_1 \neq 0$$

Under H_0 the Durbin-Watson statistic is approximated as $DW \approx 2(1-0) = 2$. The test always generates a statistic between 0 and 4. However, there is not a known distribution for critical values. Hence to establish if we can reject or not H_0 when we have values very different from 2, we should look at the tables.

24.2 Breush-Godfrey

The Breush-Godfrey test is similar to Durbin-Watson, but it allows for multiple lags in the regression. In order to perform the test let's fit an AR(p) model on the a time series $X_t = (x_1, \dots, x_i, \dots, x_t)$, i.e.

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + u_t \tag{24.2}$$

The null hypothesis H_0 is the *absence of autocorrelation*, i.e.

$$H_0 : \phi_1 = \cdots = \phi_p = 0$$
$$H_1 : \phi_1 \neq 0, \dots, \phi_p \neq 0$$

The null hypothesis H_0 is tested looking at the F statistic that is distributed as a Fisher–Snedecor distribution, i.e F $\sim F_{p,n-p-1}$. Alternatively is is possible to use the LM statistic, i.e. LM $= nR^2 \sim \chi(p)$ where R^2 is the R squared of the regression in Equation 24.2.

24.3 Box–Pierce test

Let's consider a sequence of n IID observations, i.e. $u_t \sim \text{IID}(0, \sigma^2)$. Then, the autocorrelation for the k-lag can be estimated as:

$$\hat{\rho}_k = \mathbb{C}r\{u_t, u_{t-k}\} = \frac{\sum_{t=k}^{n} u_t u_{t-k}}{\sum_{t=k}^{n} u_t^2}.$$

Moreover, since $\hat{\rho}_k \sim N\left(0, \frac{1}{n}\right)$, standardizing $\hat{\rho}_k$ one obtain

$$\sqrt{n}\hat{\rho}_k \sim N(0,1) \implies n\hat{\rho}_k^2 \sim \chi_1^2.$$

It is possible to generalize the result considering m-auto correlations. In specific, let's define a vector containing the first m standardized auto-correlations. Due to the previous result it converges in distribution to a multivariate standard normal, i.e.

$$\sqrt{n}\begin{bmatrix} \hat{\rho}_1 \\ \vdots \\ \hat{\rho}_k \\ \vdots \\ \hat{\rho}_m \end{bmatrix} \xrightarrow[n\to\infty]{d} \mathcal{N}(0_{m\times 0}, \mathbb{I}_{m\times m}).$$

Remembering that the sum of the squares of m-normal random variable is distributed as a $\chi^2(m)$, one obtain the Box–Pierce test as

$$BP_m = n\sum_{k=1}^{m} \hat{\rho}_k^2 \xrightarrow[H_0]{d} \chi_m^2,$$

where the null hypothesis and the alternative are

$$H_0 : \rho_1 = \cdots = \rho_m = 0$$
$$H_1 : \rho_1 \neq 0, \dots, \rho_p \neq 0$$

Note that such test, also known as Portmanteau test, provide an asymptotic result valid only for large samples.

> **IID assumption**
>
> Note that the assumption of the test is that the observations are IID. Therefore, the test do no apply in the case of heteroskedasticity.

24.3.1 Ljung-Box test

Since the Box–Pierce test provide a consistent framework only for large samples, when dealing with a small samples it is preferable to use an alternative version, known as Ljung-box test, defined with a correction factor, i.e.

$$\mathrm{LB}_m = n(n+2)\sum_{k=1}^{m}\frac{\hat{\rho}_k^2}{n-k} \xrightarrow[H_0]{d} \chi^2(m)$$

Independently from the statistic test used, i.e. $Q_m = BP_m$ or $Q_m = LB_m$, in general both are rejected when

$$\begin{cases} \mathrm{Q}_m > \chi^2_{1-\alpha,m} & \mathrm{H}_0 \text{ rejected} \\ \mathrm{Q}_m < \chi^2_{1-\alpha,m} & \mathrm{H}_0 \text{ not rejected} \end{cases}$$

where $\chi^2_{1-\alpha,m}$ is the quantile with probability $1-\alpha$ of the χ^2_m distribution with m degrees of freedom. If we reject H_0, the time series presents autocorrelation, otherwise if H_0 is non rejected we have no autocorrelation.

25 Normality tests

Tests of normality are statistical inference procedures designed to test that the underlying distribution of a random variable is normally distributed. There is a long history of these tests, and there are a plethora of them available for use, i.e. Jarque and Bera (1980), D'Agostino and Pearson (1973), Ralph B. D'agostino and Jr. (1990). Such kind of tests are based on the comparison of the sample skewness and kurtosis with the skewness and kurtosis of a normal distribution, hence their estimation is crucial.

25.1 Jarque-Brera test

If X is an independent and identically distributed process, the asymptotic distribution of the skewness and kurtosis in Equation 9.19 holds for $X \sim \mathcal{N}$. Hence, we construct an *omnibus test* for normality. Standardizing the skewness and kurtosis, under

$$H_0 : \beta_1(X_n) = 0,\ \beta_2(X_n) = 3,$$

we have

$$Z_1(X_n) = \sqrt{n}\frac{b_1(X_n)}{\sqrt{6}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1),$$
$$Z_2(X_n) = \sqrt{n}\frac{b_2(X_n) - 3}{\sqrt{24}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1),$$

where $b_1(X_n)$ and $b_2(X_n)$ are defined respectively in Equation 9.17 and in Equation 9.20. It is possible to prove that $Z_1(X_n)$ and $Z_2(X_n)$ are independent. Since the sum of two independent standard normal squared random variables follows a χ^2_v distribution with $v = 2$ degrees of freedom, let's rewrite the Jarque-Brera statistic as:

$$\text{JB}(X_n) = Z_1(X_n)^2 + Z_2(X_n)^2 \xrightarrow[n\to\infty]{d} \chi^2_2.$$

However, if n is small the $\text{JB}(X_n)$ over-reject the null hypothesis H_0, i.e. type I error.

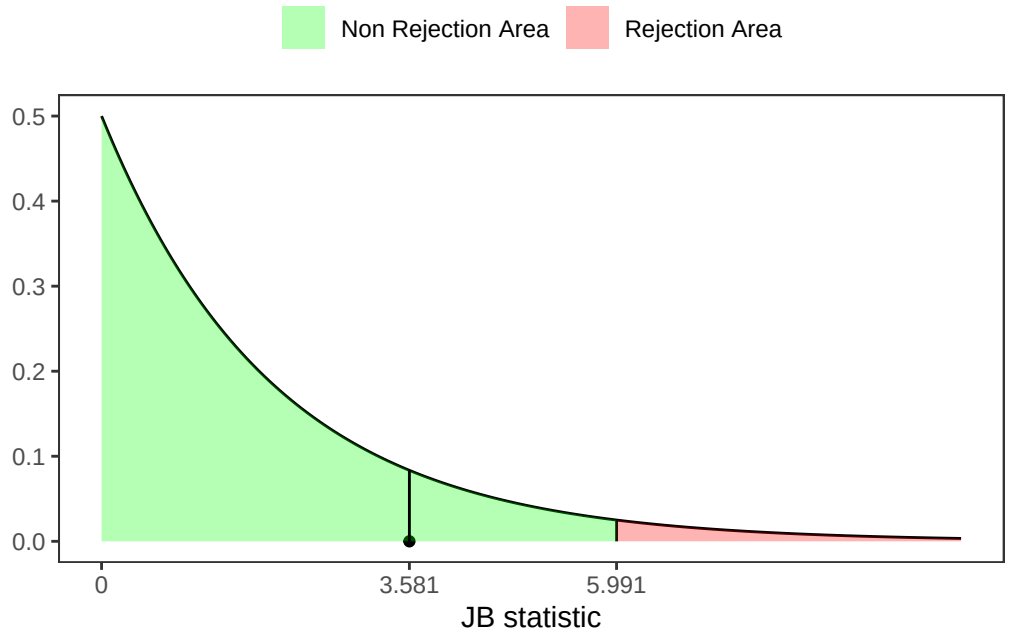


Figure 25.1: Jarque-Brera test for normality with a simulated Normal sample of 50 observations with $\alpha = 0.05$.

25.2 Urzua-Jarque-Brera test

Let's substitute the asymptotic moments with the exact sample moments of skewness and kurtosis. Following Urzúa (1996), let's write the new omnibus test statistic as:

$$\tilde{Z}_1(X_n) = \frac{b_1(X_n)}{\sqrt{\mathbb{V}\{b_1(X_n)\}}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1),$$
$$\tilde{Z}_2(X_n) = \frac{b_2(X_n) - \mathbb{E}\{b_2(X_n)\}}{\sqrt{\mathbb{V}\{b_2(X_n)\}}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1),$$

where the exact moments $\mathbb{V}\{b_1(X_n)\}$, $\mathbb{E}\{b_2(X_n)\}$ and $\mathbb{V}\{b_2(X_n)\}$ are defined respectively in Equation 9.18, Equation 9.21, Equation 9.22. Hence, the Urzua-Jarque-Brera test adjusted for small samples is computed as:

$$\text{UJB}(X_n) = \tilde{Z}_1(X_n)^2 + \tilde{Z}_2(X_n)^2 \xrightarrow[n\to\infty]{d} \chi^2_2.$$

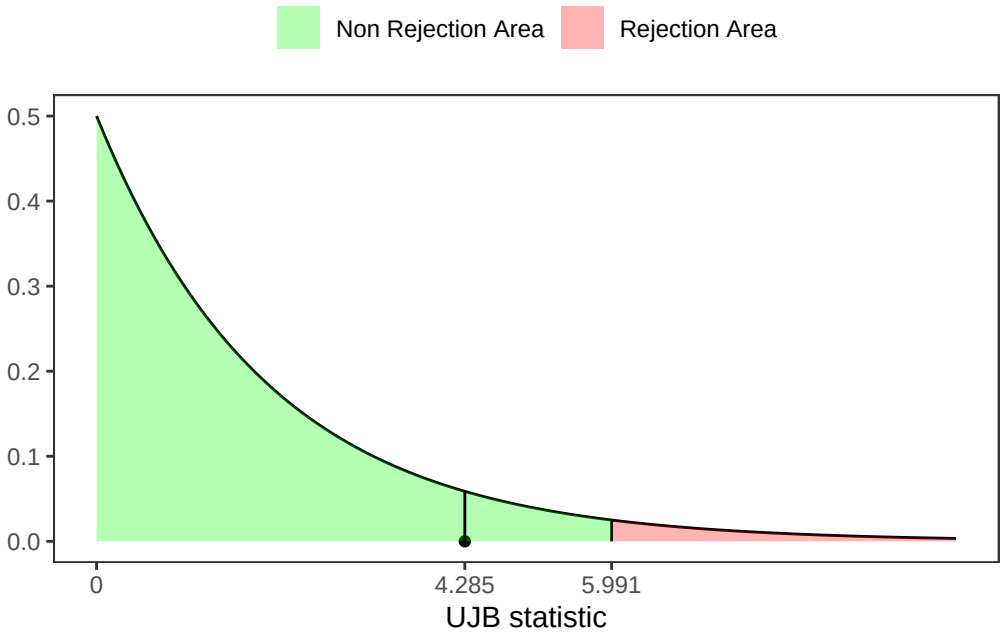


Figure 25.2: Urzua-Jarque-Brera test for normality with a simulated Normal sample of 50 observations with $\alpha = 0.05$.

25.3 D'Agostino skewness test

D'Agostino and Pearson (1973) proposed an alternative way to test that the skewness is different from zero. Starting from the statistic $\tilde{Z}_1(X_n)$, compute:

$$\beta_2(b_1) = \frac{3(n^2 + 27n - 70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)},$$

and

$$W^2 = \sqrt{2\beta_2(b_1) - 1} - 1,$$
$$\delta = \frac{1}{\sqrt{\ln(W)}},$$
$$\alpha = \sqrt{\frac{2}{W^2 - 1}}.$$

The statistic test for skewness is defined as:

$$Z_1^*(X_n) = \delta \log \left\{ \frac{\tilde{Z}_1(X_n)}{\alpha} + \sqrt{\left(\frac{\tilde{Z}_1(X_n)}{\alpha} \right)^2 + 1} \right\} =$$
$$= \delta \sinh^{-1} \left(\frac{\tilde{Z}_1(X_n)}{\alpha} \right),$$

where $Z_1^*(X_n) \sim \mathcal{N}(0,1)$.

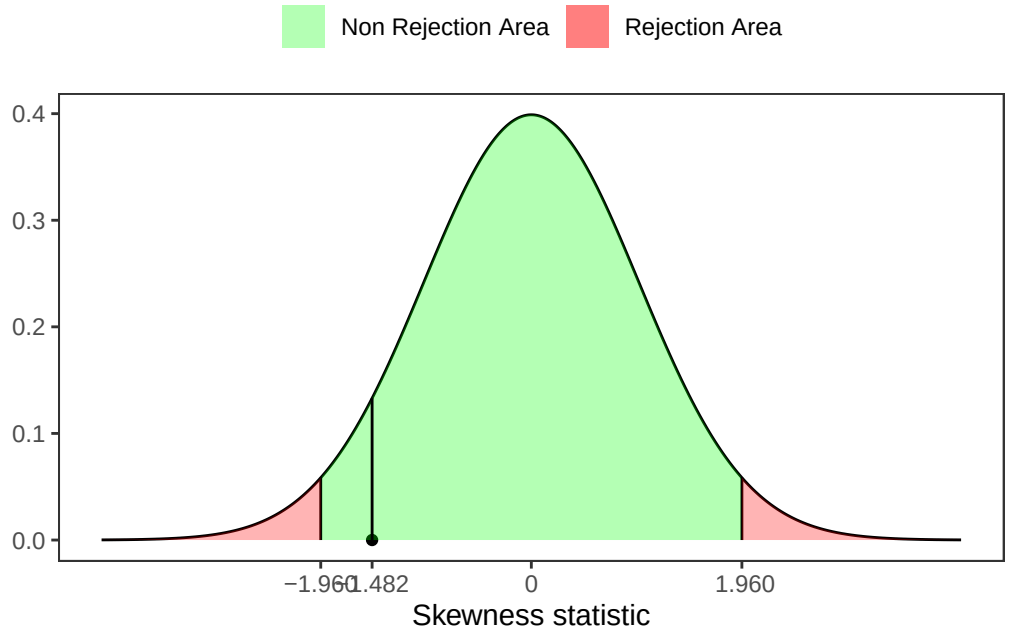


Figure 25.3: Skewness test with a simulated Normal sample of 50 observations with $\alpha = 0.05$.

25.4 Anscombe Kurtosis test

Anscombe and Glynn (1983) proposed a test for skewness. Starting from the statistic $\tilde{Z}_2(X_n)$, let's compute the third standardized moment of b_2:

$$\beta_1(b_2) = \frac{6(n^2 - 5n + 2)}{(n+7)(n+9)} \sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}},$$

and

$$A = 6 + \frac{8}{\beta_1(b_2)} \left[\frac{2}{\beta_1(b_2)} + \sqrt{1 + \frac{4}{\beta_1(b_2)}} \right].$$

The statistic test for kurtosis is defined as:

$$Z_2^*(X_n) = \sqrt{\frac{9A}{2}} \left(1 - \frac{2}{9A} - \left[\frac{1 - 2/A}{1 + \tilde{Z}_2(X_n)\sqrt{2/(A-4)}} \right]^{1/3} \right),$$

where $Z_2^*(X_n) \sim \mathcal{N}(0,1)$.

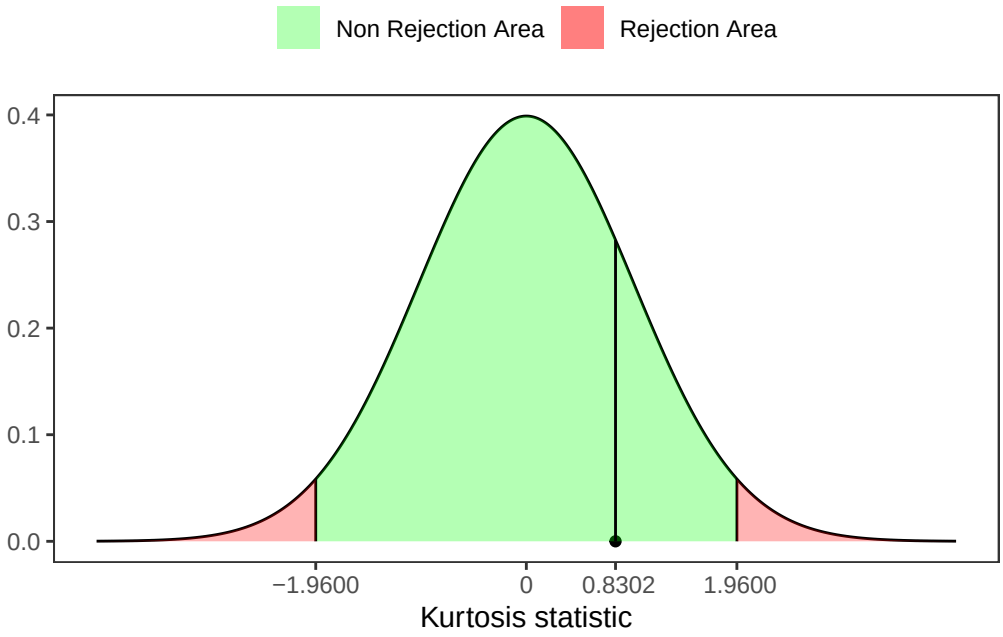


Figure 25.4: Kurtosis test with a simulated Normal sample of 50 observations with $\alpha = 0.05$.

25.5 D'Agostino-Pearson K^2 test

Finally in Ralph B. D'agostino and Jr. (1990), there is also another omnibus test based on the statistics $Z_1^*(X_n)$, $Z_2^*(X_n)$, i.e.

$$K^2 = (Z_1^*(X_n))^2 + (Z_2^*(X_n))^2 \sim \chi^2(2).$$

25.6 Kolmogorov-Smirnov Test

The Kolmogorov–Smirnov test can be used to verify whether a samples is drawn from a reference distribution. In the case of q sample with dimension n, the KS statistic is defined as:

$$KS_n = \sup_{\forall x} |F_n(x) - F(x)|. \tag{25.1}$$

In this settings, the test statistic follows the Kolmogorov distribution, i.e.

$$F_{KS}(\sqrt{n} \cdot KS_n < x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2}.$$

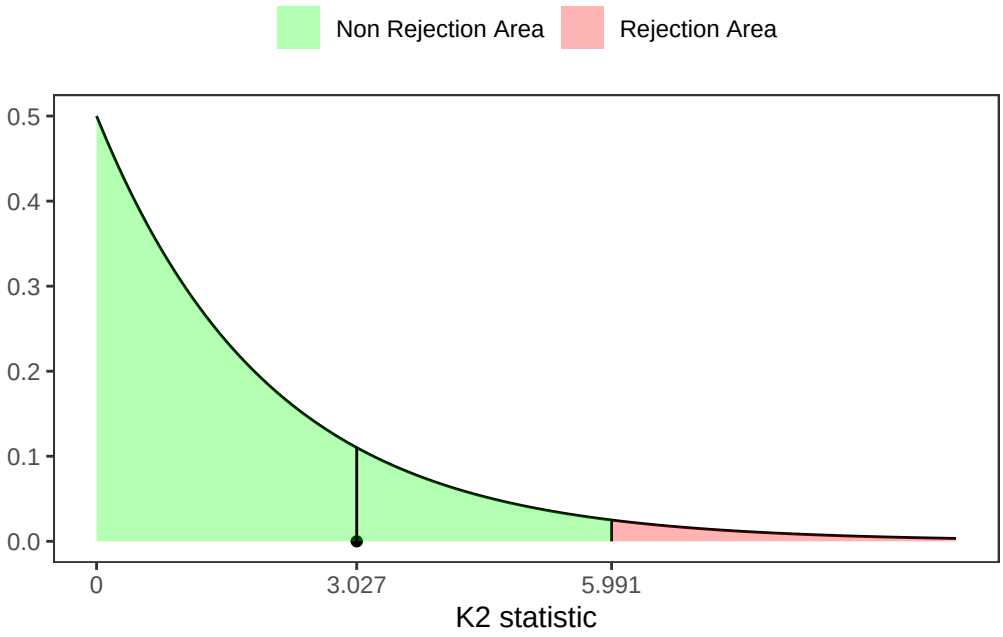


Figure 25.5: Agostino normality test with a simulated Normal sample of 50 observations with $\alpha = 0.05$.

The null distribution of this statistic is calculated under the null hypothesis that the sample is drawn from the reference distribution F, i.e.

$$H_0 : F_n(X) = F(X) \quad H_1 : F_n(X) \neq F(X)$$

For large samples, H_0 is rejected at a given confidence level α if:

$$\sqrt{n} \cdot KS_n > F_{KS}^{-1}(F_{KS}(\sqrt{n} \cdot KS_n < K_\alpha)),$$

where K_α represents the critical value and F_{KS}^{-1} the quantile function of the Kolmogorov distribution.

25.6.1 Example 1: KS test for normality

Let's simulate 500 observations of a stationary normal random variable, i.e. $X_n \sim \mathcal{N}(0.2, 1)$.

Table 25.1: KS-test for normality on a Normal's random sample with $\alpha = 0.05$.

n	α	KS_n	**Critical Level**	H_0
5000	5%	0.8651	1.358	Non-Rejected

1. Simulated stationary sample
2. Set the interval values for upper and lower band. This is done to avoid including outliers, however it is possible to use also the minimum and maximum.
3. Empirical cdf
4. Compute the KS-statistic as in Equation 25.1.
5. Compute the rejection level

25.6.2 Example 2: KS test for normality

Let's simulate 500 observations of a stationary t-student random variable with increasing degrees of freedom ν, i.e. $X_n \sim t(\nu)$. Then, we compare the obtained result with a standard normal random variable, i.e. $\mathcal{N}(0,1)$. It is known that increasing the degrees of freedom of a student-t imply the convergence to a standard normal. Hence, we expect that from a certain ν awards the null hypothesis will be no more rejected.

Table 25.2: KS-test for normality on t-student's random samples with $\alpha = 0.05$.

ν	n	α	KS_n	**Critical Level**	H_0
1	5000	5%	9.356	1.358	Rejected
5	5000	5%	2.974	1.358	Rejected
10	5000	5%	1.751	1.358	Rejected
15	5000	5%	1.462	1.358	Rejected
20	5000	5%	1.285	1.358	Non-Rejected
30	5000	5%	1.190	1.358	Non-Rejected

26 Stationarity tests

26.1 Dickey–Fuller test

The Dickey–Fuller test tests the null hypothesis that a unit root is present in an auto regressive (AR) model. The alternative hypothesis is different depending on which version of the test is used, usually is "stationary" or "trend-stationary". Let's consider an AR(1) model, i.e.

$$x_t = \mu + \delta t + \phi_1 x_{t-1} + u_t, \tag{26.1}$$

or equivalently adding and subtracting x_{t-1}

$$\Delta x_t = \mu + \delta t + (1 - \phi_1) x_{t-1} + u_t. \tag{26.2}$$

The hypothesis of the Dickey–Fuller test are:

$$\begin{aligned} &\mathrm{H}_0 : \phi_1 = 1 \text{ (non stationarity)} \\ &\mathrm{H}_1 : \phi_1 < 1 \text{ (stationarity)} \end{aligned}$$

The Dickey–Fuller statistic (DF) is computed as:

$$\mathrm{DF} = \frac{1 - \phi_1}{\mathbb{S}d\{1 - \phi_1\}}$$

However, since the test is done over the residual term rather than raw data, it is not possible to use the t-distribution to provide critical values. Therefore, the statistic DF has a specific distribution.

26.2 Augmented Dickey–Fuller test

The augmented Dickey–Fuller is a more general version of the Dickey–Fuller test for a general AR(p) model, i.e.

$$\Delta x_t = \mu + \delta t + \gamma x_{t-1} + \sum_{i=1}^{p} \phi_i \Delta x_{t-i}$$

The hypothesis of the augmented Dickey–Fuller test are:

$$\begin{aligned} &\mathrm{H}_0 : \gamma = 0 \ \text{ (non stationarity)} \\ &\mathrm{H}_1 : \gamma < 0 \ \text{ (stationarity)} \end{aligned}$$

The augmented Dickey–Fuller statistic (ADF) is computed as:

$$\text{ADF} = \frac{\gamma}{\mathbb{S}d\{\gamma\}}$$

As in the simpler case, the critical values are computed using a specific table for the Dickey–Fuller test.

26.3 Kolmogorov-Smirnov test

The Kolmogorov–Smirnov two-sample test (KS) can be used to test whether two samples came from the same distribution. Let's define the empirical distribution function F_n of n-independent and identically distributed ordered observations $X_{(i)}$ as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,x]}(X_{(i)}).$$

The KS statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution functions of two random samples. The distribution of the KS statistic under the null hypothesis assumes that the samples are drawn from the same distribution, i.e.

$$\begin{aligned} \mathrm{H}_0 &: X \text{ is stationary} \\ \mathrm{H}_1 &: X \text{ is not stationary} \end{aligned}$$

The statistic test for two samples with dimension n_1 and n_2 is defined as:

$$\text{KS}_{n_1,n_2} = \sup_{\forall x} |F_{n_1}(x) - F_{n_2}(x)|,$$

and for large samples H_0 is rejected with confidence level $1-\alpha$ if:

$$\text{KS}_{n_1,n_2} > \sqrt{-\frac{1}{2n_2}\ln\left(\frac{\alpha}{2}\right)\left(1+\frac{n_2}{n_1}\right)}.$$

Hence, since the statistic is always greater of equal to zero, with a given statistic KS_{n_1,n_2} the p-value with confidence level $\alpha = 2\mathbb{P}(X > \text{KS}_{n_1,n_2})$ reads:

$$\mathbb{P}(X > \text{KS}_{n_1,n_2}) = \exp\left(-\frac{2n_2}{1+\frac{n1}{n2}}\text{KS}^2_{n_1,n_2}\right)$$

KS test for time series

To apply the test in a time series settings, use a random index to split the original series in two sub-series. Then the KS can be applied as usual.

26.3.1 Examples

Check for stationarity

Example 26.1. Let's consider 500 simulated observations of the random variable X drown from a population distributed as $X \sim N(0.4, 1)$. Then, considering it as a time series, let's sample a random index to split the series in a point. Finally, as shown in Table 26.1 the null hypothesis, i.e. the two samples come from the same distribution, is not reject with the confidence level $\alpha = 5\%$.

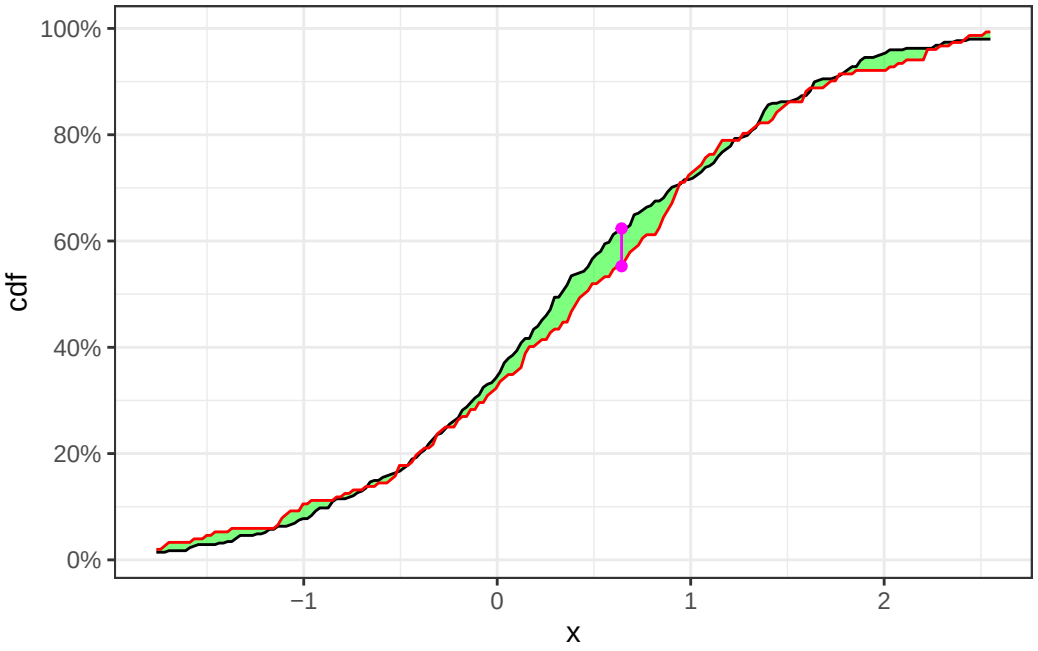


Figure 26.1: Two samples cdfs and KS-statistic (magenta) for a stationary time series.

Table 26.1: KS test for a stationary time series.

Index split	α	n_1	n_2	KS_{n_1,n_2}	p.value	**Critical level**	H_0
348	0.05	348	152	0.07093	0.6282	0.132	Non-Rejected

Check for non-stationarity

Example 26.2. Let's consider 250 simulated observations of the random variable X drown from a population distributed as $Y_{1,t} \sim N(0,1)$ and the following 250 from $Y_{2,t} \sim N(0.3,1)$. Then the non-stationary series will have a structural break at the point 250 and the time series is given by:

$$X_t = \begin{cases} Y_{1,t} & t \leq 250 \\ Y_{2,t} & t > 250 \end{cases}$$

As in Example 26.2 let's split the time series and apply the KS-test. In this case, as shown in Table 26.2 the null hypothesis, i.e. the two samples come from the same distribution, is reject with confidence level $\alpha = 5\%$.

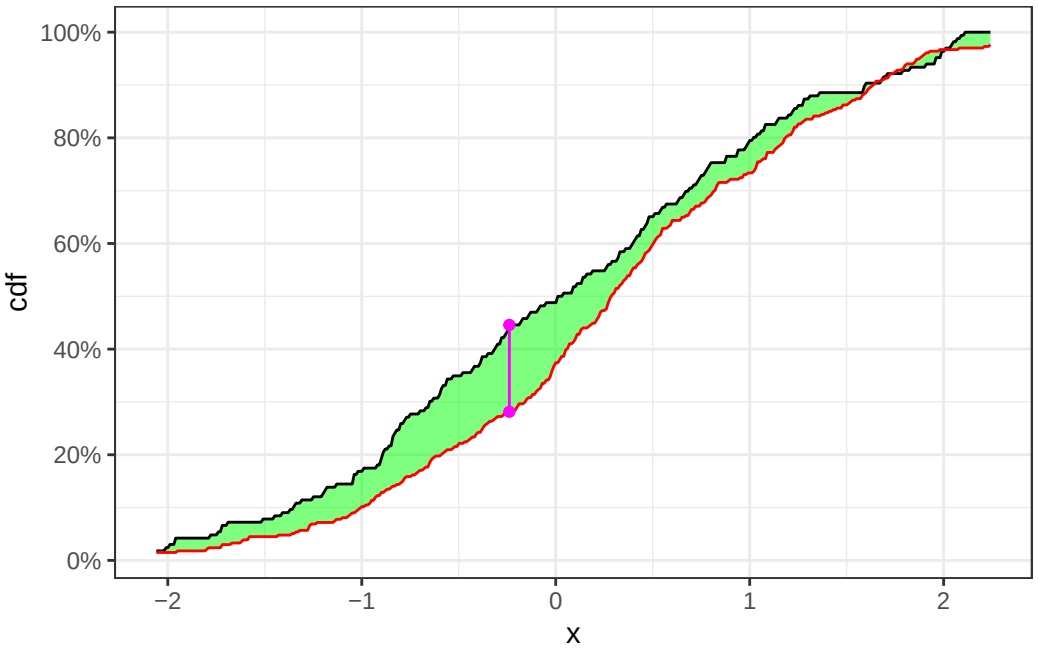


Figure 26.2: Two samples cdfs and KS-statistic (magenta) for a non-stationary time series.

Table 26.2: KS test for a non-stationary time series.

Index split	α	n_1	n_2	KS_{n_1,n_2}	p.value	**Critical level**	H_0
166	0.05	166	334	0.1643	0.00000583	0.1129	Rejected

27 Value at Risk test

Let's consider a ARMA(2,2)-GARCH(1,1) model defined as:

$$\begin{aligned} x_t &= \mu + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \theta_1 e_{t-1} + \theta_2 e_{t-2} + e_t \\ e_t &= \sigma_t u_t \\ \sigma_t^2 &= \omega + \alpha_1 e_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \end{aligned}$$

The Value at Risk (VaR) with confidence level α depends on the distribution of x_t that is implicitly defined from the distribution of u_t. Therefore, the VaR at time t, conditional to the information up to time $t-1$, is implicitly defined as:

$$\mathbb{P}(X_t \leq \text{VaR}_{t|t-1}^{\alpha}) = \alpha.$$

27.1 Normal distribution

Let's consider independent and normally distributed residuals u_t. Then, also the conditional distribution of x_t given the information up to time $t-1$ will be normal with conditional mean and variance given by:

$$\begin{cases} \mathbb{E}\{X_t \mid I_{t-1}\} = \mu + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \theta_1 e_{t-1} + \theta_2 e_{t-2} \\ \mathbb{V}\{X_t \mid I_{t-1}\} = \omega + \alpha_1 e_{t-1}^2 + \beta_1 \sigma_{t-1}^2 = \sigma_t^2 \end{cases}$$

Hence, given the quantile q^{α} of a standard normal with level α, the VaR is computed as:

$$\text{VaR}_{t|t-1}^{\alpha} = \mathbb{E}\{X_t \mid I_{t-1}\} + q^{\alpha}\sqrt{\mathbb{V}\{X_t \mid I_{t-1}\}}.$$

27.2 Gaussian Mixture distribution

Let's consider independent and Gaussian mixture distributed residuals u_t with 2 components. Then, also the conditional distribution of x_t given the information up to time $t-1$ will be a Gaussian mixture with conditional mean and variance given by:

$$\begin{cases} \mathbb{E}\{X_t \mid I_{t-1}, B=1\} = \mu + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \mu_1 \sigma_t \\ \mathbb{E}\{X_t \mid I_{t-1}, B=0\} = \mu + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \mu_0 \sigma_t \\ \mathbb{V}\{X_t \mid I_{t-1}, B=1\} = (\sigma_t \sigma_1)^2 \\ \mathbb{V}\{X_t \mid I_{t-1}, B=0\} = (\sigma_t \sigma_0)^2 \end{cases}$$

Hence, given the quantile q^α of a Gaussian mixture with 2 components with level α, that in general needs to be computed numerically, the VaR is defined as:

$$\text{VaR}_{t|t-1}^{\alpha} = q^\alpha.$$

27.3 Test on the number of violations

Let's define a violation of the $\text{VaR}_{t|t-1}^{\alpha}$ as

$$v_t = \mathbb{1}_{[x_t \leq \text{VaR}_{t|t-1}^{\alpha}]} \sim \text{Bernoulli}(\alpha),$$

and let's define the number of violations of the conditional VaR as follows, i.e.

$$N_t = \sum_{i=1}^{t} v_i.$$

27.3.1 Asymptotic variance

Applying the central limit theorem (CLT) it is possible to prove that the statistic test converges in distribution to a standard normal, i.e.

$$\text{T}_1^{\alpha} = \frac{1}{\sqrt{t}} \sum_{i=1}^{t} \left(\frac{v_i - \alpha}{\sqrt{\alpha(1-\alpha)}} \right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1).$$

Hence, given the null hypothesis $H_0 : \mathbb{P}\{X_t \leq \text{VaR}_{t|t-1}^{\alpha} | I_{t-1}) = \alpha$, that is equivalent to $\mathbb{E}\{e_t\} = \alpha$ we define the critical values at a confidence level α^* as

$$\begin{aligned} \alpha &= \mathbb{P}\{|\text{T}_1^{\alpha}| > t_{\alpha^*/2}\} \\ &\Updownarrow \\ t_{\alpha^*/2} &= \mathbb{P}^{-1}\{\mathbb{P}\{|\text{T}_1^{\alpha}| > t_{\alpha^*/2}\}\} \end{aligned}$$

where $\mathbb{P}$ and $\mathbb{P}^{-1}$ are respectively the distribution and the quantile of a standard normal. Therefore, the null hypothesis is rejected at a confidence level α^* if:

$$\begin{cases} [\text{T}_1^{\alpha} < -t_{\alpha^*/2}] \cup [\text{T}_1^{\alpha} > t_{\alpha^*/2}] & \implies H_0 \text{ rejected} \\ [-t_{\alpha^*/2} < \text{T}_1^{\alpha} < t_{\alpha^*/2}] & \implies H_0 \text{ not rejected} \end{cases}$$

27.3.2 Empirical variance

Instead of using the theoretical variance of e_t, namely $\alpha(1-\alpha)$, let's substitute it with the empirical one, i.e.

$$\alpha(1-\alpha) \longrightarrow \frac{N_t}{t}\left(1-\frac{N_t}{t}\right).$$

Hence, the new statistic test NV_2 converges to NV_1 in probability, therefore also in distribution, i.e.

$$\mathrm{T}_2^\alpha = \frac{1}{\sqrt{t}}\sum_{i=1}^{t}\left(\frac{v_i-\alpha}{\sqrt{\frac{N_t}{t}\left(1-\frac{N_t}{t}\right)}}\right) \xrightarrow[n\to\infty]{p} \mathrm{T}_1^\alpha \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1).$$

For small samples, the following relation between the two statistics should be used:

$$\mathrm{T}_2^\alpha = \frac{\sqrt{\alpha(1-\alpha)}}{\sqrt{\frac{N_t}{t}\left(1-\frac{N_t}{t}\right)}}\mathrm{T}_1^\alpha.$$

27.4 Example: H_0 is not rejected

Instead of simulating exactly $u_t \sim \mathcal{N}(0,1)$, let's simulate residuals that are close to the normal distribution, i.e. $u_t \sim t(25)$.

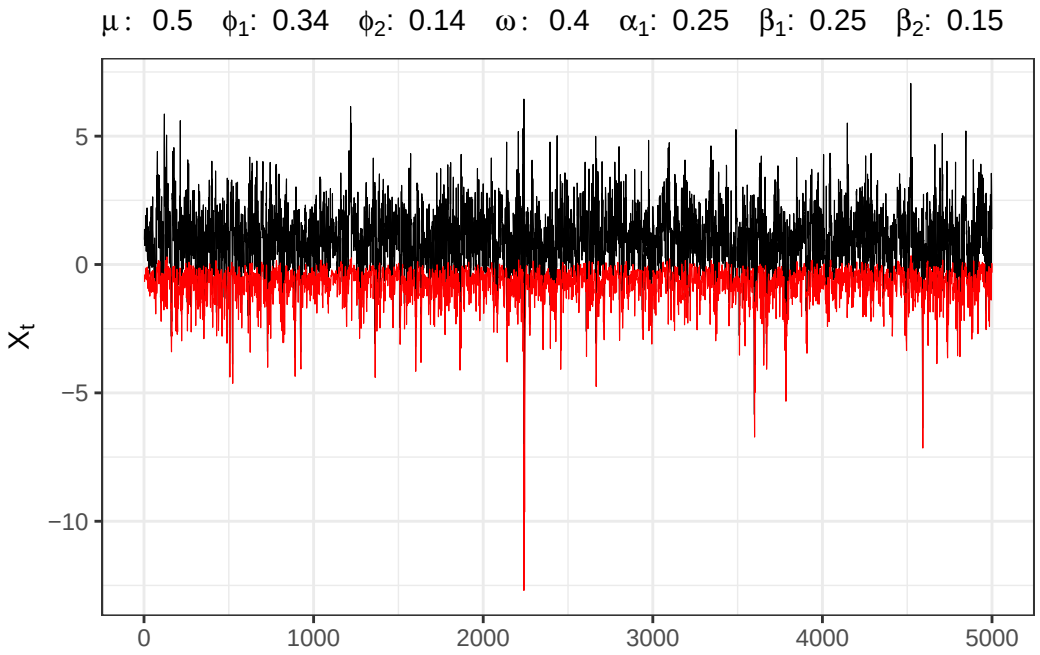


Figure 27.1: AR(2)-GARCH(1,2) simulation with theoric (red) VaR at $\alpha = 0.05$.

Table 27.1: Test for a Student-t with 25 degrees of freedom at $\alpha^* = 0.05$ on the number of violations of the theoric VaR at $\alpha = 0.05$.

n	α	$\frac{N_n}{n}$	$t_{\alpha/2}$	T_1^α	T_2^α	$t_{\alpha/2}$	$H_0(\mathrm{T}_1)$	$H_0(\mathrm{T}_2)$
5000	5%	5.6%	-1.96	1.947	1.845	1.96	Non-Rejected	Non-Rejected

27.5 Example: H_0 is rejected

Instead of simulating exactly $u_t \sim \mathcal{N}(0,1)$, let's simulate residuals that are not close to the normal distribution, i.e. $u_t \sim t(5)$.

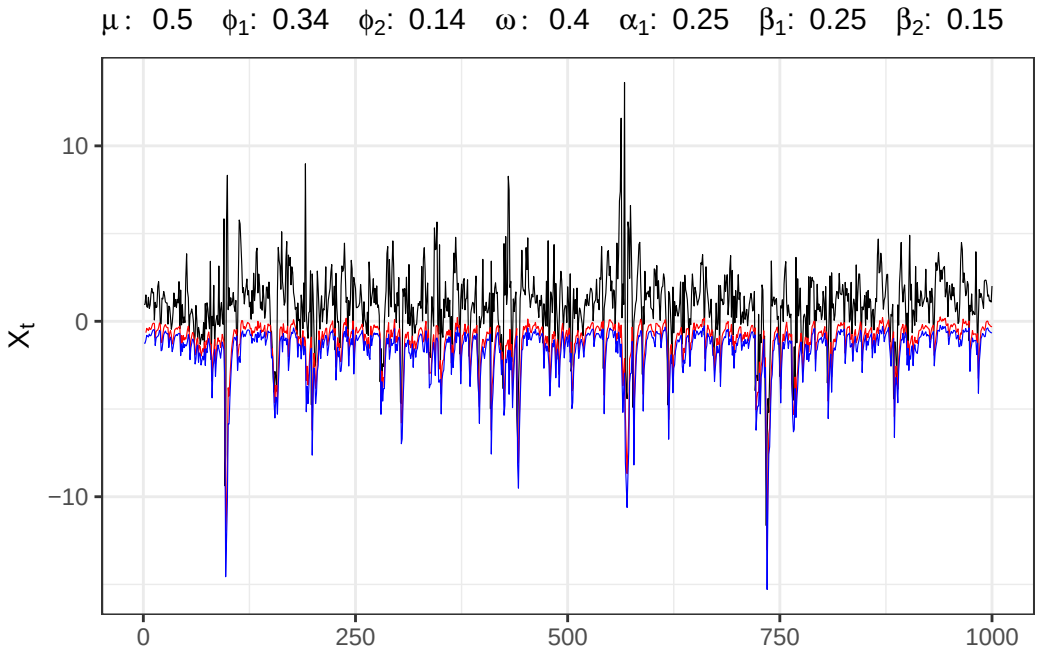


Figure 27.2: AR(2)-GARCH(1,2) simulation with theoric (red) and empirical (blue) VaR at $\alpha = 0.05$.

Computing the test on the normal quantile gives a rejection of the null hypothesis H_0, i.e. the deviation from the VaR is not stochastic and it is not an adequate measure of risk.

Table 27.2: Test for a Student-t with 5 degrees of freedom at $\alpha^* = 0.05$ on the number of violations of the theoric VaR at $\alpha = 0.05$.

n	α	$\frac{N_n}{n}$	$-t_{\alpha/2}$	T_1^α	T_2^α	$t_{\alpha/2}$	$H_0(\mathrm{T}_1)$	$H_0(\mathrm{T}_2)$
1000	5%	8.5%	-1.96	5.078	3.969	1.96	Rejected	Rejected

Setting $\alpha = 0.05$ we obtain an empiric $\hat{q}_\alpha$ equal to -2.07 different from the theoric one of -1.6449.

Part VI

Robustness

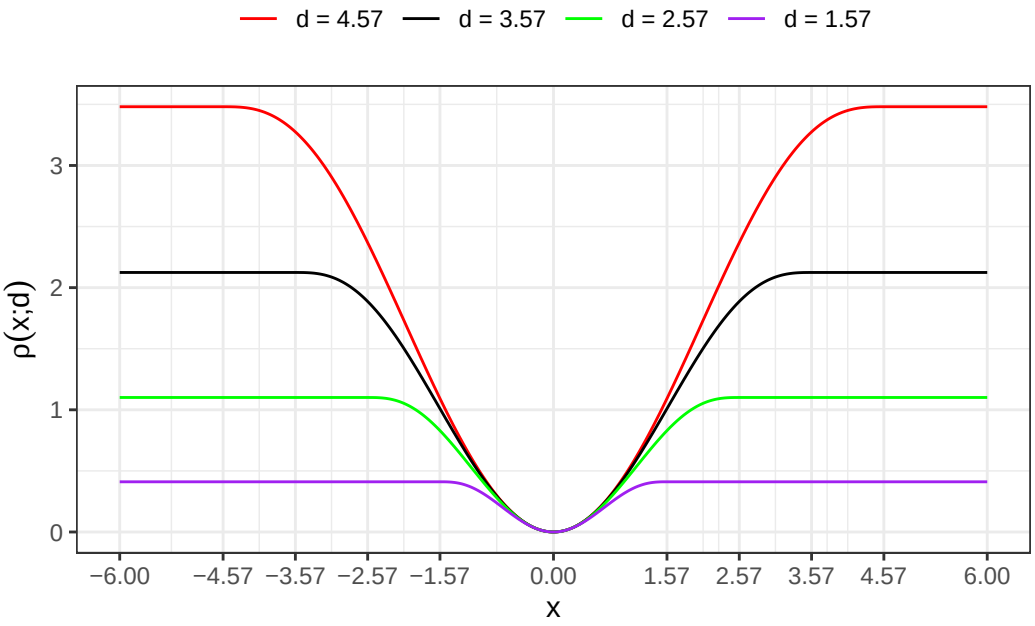
28 Tukey functions

28.1 Tukey's Bisquare

$$\rho_d(x) = \begin{cases} \frac{d^2}{6}\left\{1 - \left[1 - \frac{x^2}{d^2}\right]^3\right\} & |x| \leq d \\ \frac{d^2}{6} & |x| > d \end{cases}$$

28.2 R

```
tukey_bisquare <- function(d){
  function(x){
    x[abs(x) > d] <- NA
    f_x <- (d^2)/6*(1 - (1 - (x/d)^2)^3)
    f_x[is.na(f_x)] <- (d^2)/6
    return(f_x)
  }
}
```

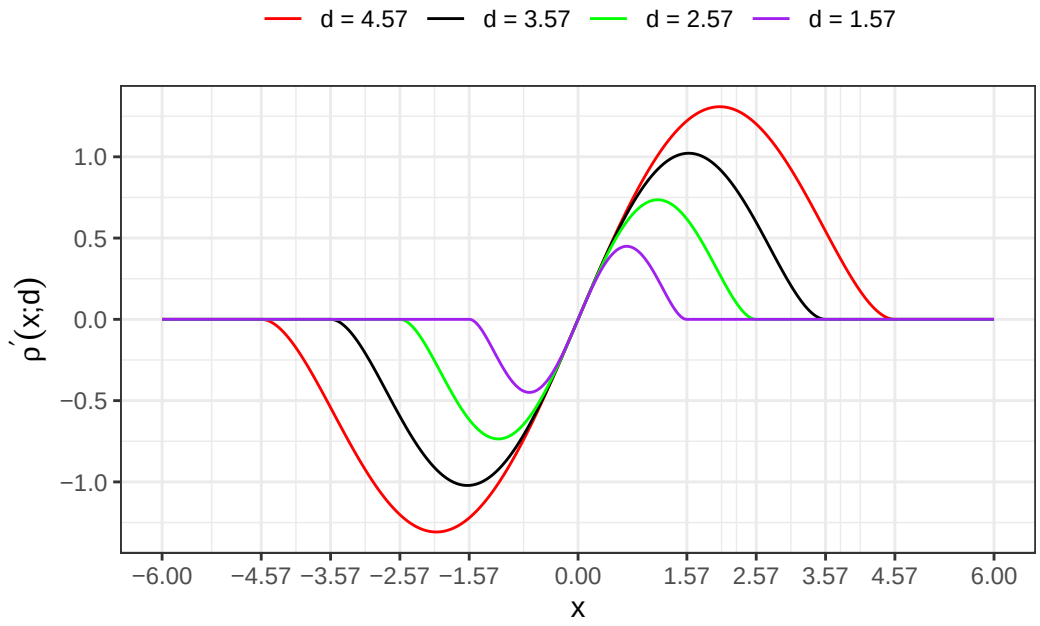


28.2.1 First derivative

$$\rho'(x;d) = \begin{cases} x\left[1-\frac{x^2}{d^2}\right]^2 & |x| \le d \\ 0 & |x| > d \end{cases}$$

28.3 R

```r
# Tukey's Bisquare First Derivative
tukey_bisquare_prime <- function(d){
  function(x){
    x[abs(x) > d] <- NA
    f_x <- x*(1 - (x/d)^2)^2
    f_x[is.na(f_x)] <- 0
    return(f_x)
  }
}
```

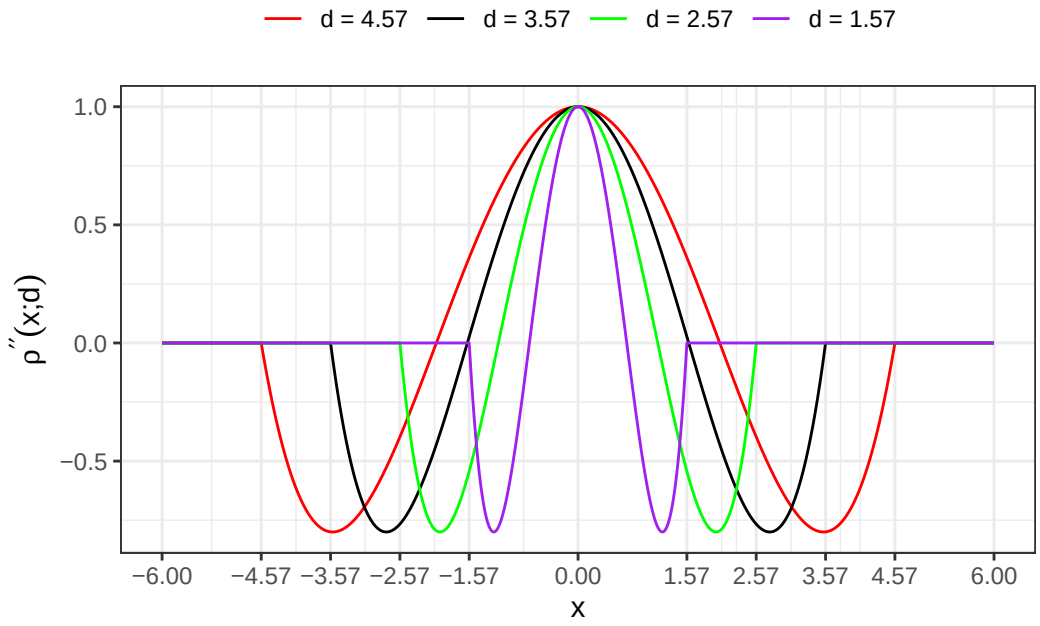


28.3.1 Second derivative

$$\rho''_d(x) = \begin{cases} \left(1 - \frac{x^2}{d^2}\right)\left(1 - \frac{x^2}{d^2} - \frac{4x^2}{d^2}\right) & |x| \leq d \\ 0 & |x| > d \end{cases}$$

28.4 R

```r
# Tukey's Bisquare Second Derivative
tukey_bisquare_second <- function(d){
  function(x){
    x[abs(x) > d] <- NA
    f_x <- (1 - (x/d)^2)*((1 - (x/d)^2) - 4*(x^2)/(d^2))
    f_x[is.na(f_x)] <- 0
    return(f_x)
  }
}
```

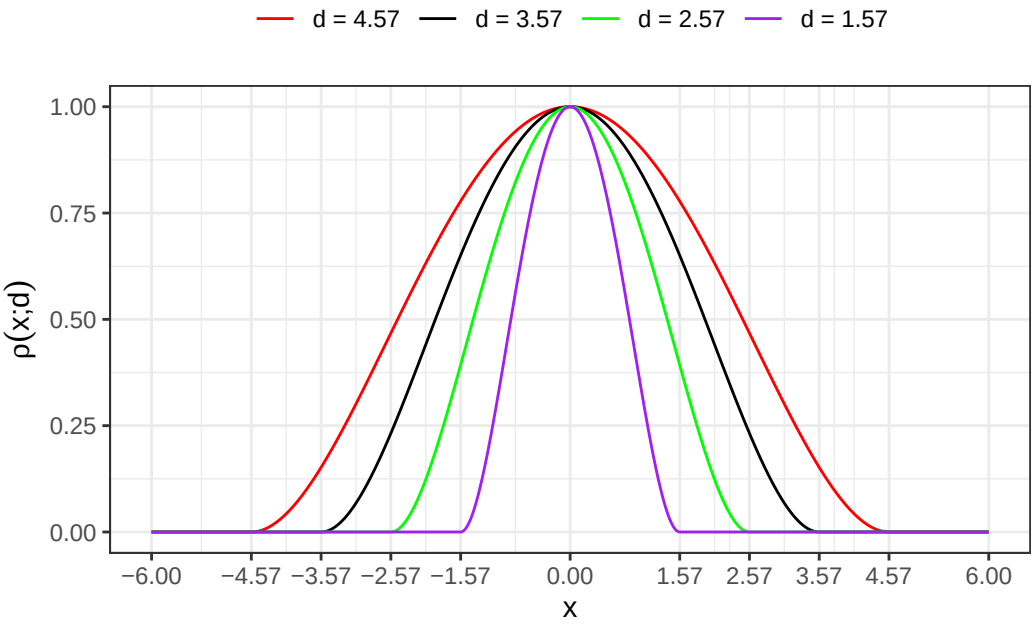


28.5 Tukey Biweight

$$\rho_d(x) = \begin{cases} \left(1 - \frac{x^2}{d^2}\right)^2 & |x| \leq d \\ 0 & |x| > d \end{cases}$$

28.6 R

```
# Tukey's Biweight Function
tukey_biweight <- function(d){
  function(x){
    x[abs(x) > d] <- NA
    f_x <- (1 - (x/d)^2)^2
    f_x[is.na(f_x)] <- 0
    return(f_x)
  }
}
```

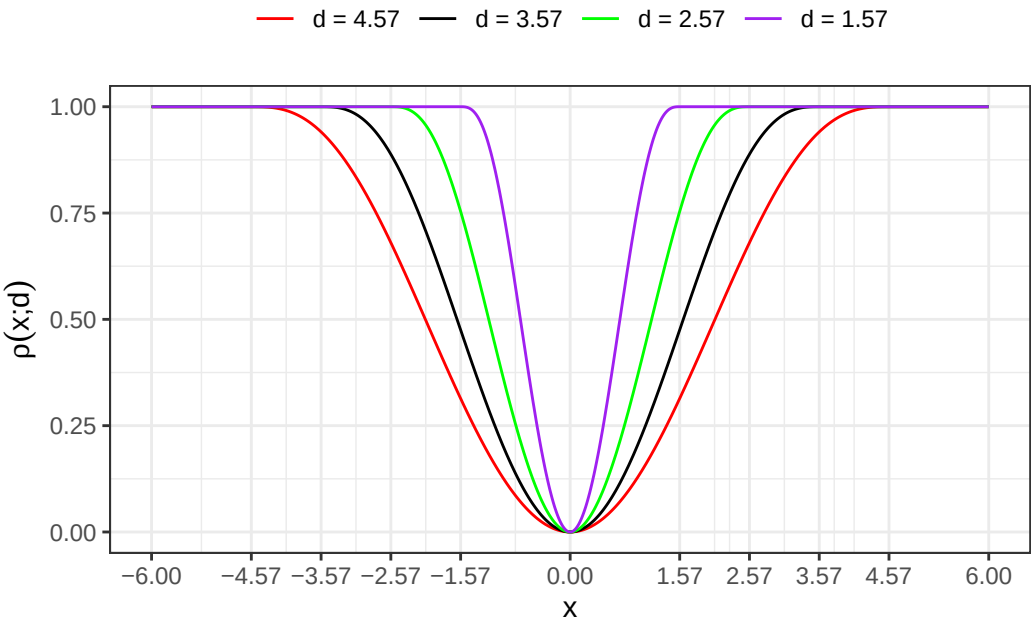


28.7 Tukey-Beaton Bisquare

$$\rho_d(x) = \begin{cases} \frac{3x^2}{d^2} - \frac{3x^4}{d^4} + \frac{x^6}{d^6} & |x| \leq d \\ 1 & |x| > d \end{cases}$$

28.8 R

```r
# Tukey-Beaton Bisquare Function
tukey_beaton_bisquare <- function(d){
  function(x){
    x[abs(x) > d] <- NA
    f_x <- 3*(x/d)^2 - 3*(x/d)^4 + (x/d)^6
    f_x[is.na(f_x)] <- 1
    return(f_x)
  }
}
```

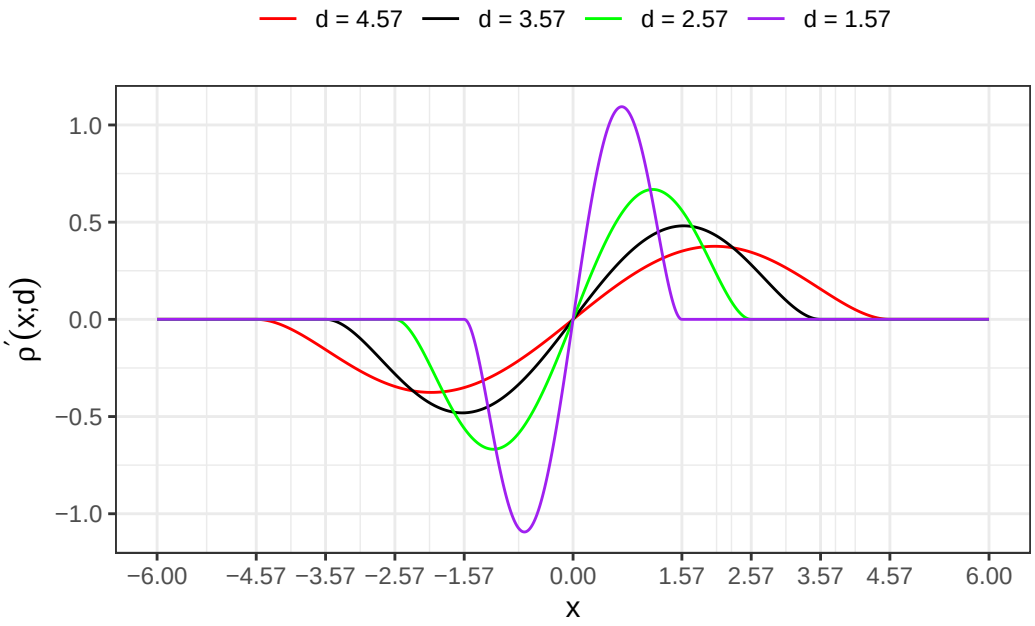


28.8.1 First derivative

$$\rho'_d(x) = \begin{cases} \frac{6x}{d^2} - \frac{12x^3}{d^4} + \frac{6x^5}{d^4} & |x| \leq d \\ 0 & |x| > d \end{cases}$$

28.9 R

```
# Tukey-Beaton Bisquare First Derivative
tukey_beaton_prime <- function(d){
  function(x){
    x[abs(x) > d] <- NA
    f_x <- 6*(1/d^2)*x - 12*(1/d^4)*(x)^3 + 6*(x)^5*(1/d^6)
    f_x[is.na(f_x)] <- 0
    return(f_x)
  }
}
```

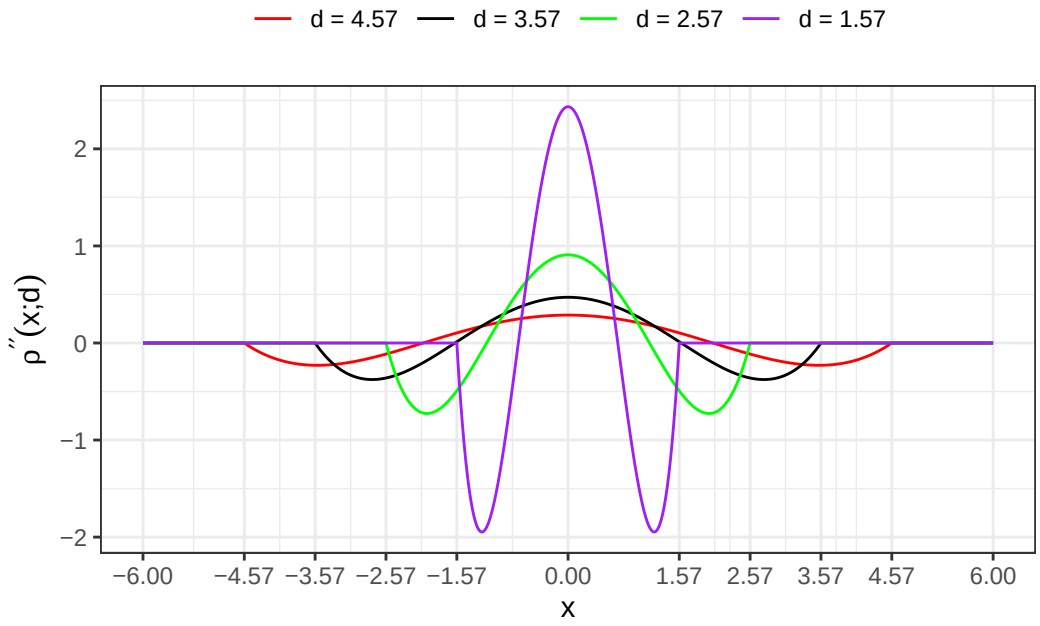


28.9.1 Second derivative

$$\rho_d''(x) = \begin{cases} \frac{6}{d^2} - \frac{36x^2}{d^4} + \frac{30x^4}{d^4} & |x| \leq d \\ 0 & |x| > d \end{cases}$$

28.10 R

```
# Tukey-Beaton Bisquare Second Derivative
tukey_beaton_second <- function(d){
  function(x){
    x[abs(x) > d] <- NA
    f_x <- 6*(1/d^2) - 36*(1/d^4)*(x)^2 + 30*(x)^4*(1/d^6)
    f_x[is.na(f_x)] <- 0
    return(f_x)
  }
}
```


d = 4.57
d = 3.57
d = 2.57
d = 1.57
ρ″(x;d)
2
1
0
−1
−2
−6.00
−4.57
−3.57
−2.57
−1.57
0.00
1.57
2.57
3.57
4.57
6.00
x

Part VII

Distributions

29 Gaussian mixture

Let's consider a linear combination of a Bernoulli and two normal random variables, all assumed to be independent, i.e.

$$X_t \sim B_t \cdot X_{1,t} + (1 - B_t) \cdot X_{0,t}, \tag{29.1}$$

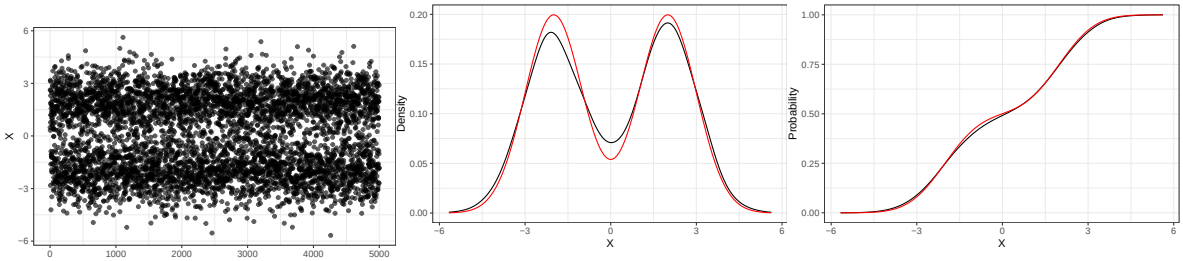
where B is a Bernoulli random variable

$$B_t \sim \text{Bernoulli}(p),$$

and for the i-th component

$$X_{i,t} = \mu_1 + \sigma_1 Z_{i,t}$$

where $Z_{i,t}$ is standard Normal random variable. In compact form a Gaussian Mixture with two components is denoted as $X_t \sim GM(\mu_1, \mu_0, \sigma_1^2, \sigma_0^2, p)$.



(a) Simulated mixture sample. (b) Simulated (black) and true (red) density. (c) Simulated (black) and true (red) distribution.

Figure 29.1: Gaussian Mixture simulation and density function with true parameters $\mu_1 = -2$, $\mu_2 = 2$, $\sigma_1 = 1$, $\sigma_2 = 1$ and $p = 0.5$.

29.1 Distribution and density

Proposition 29.1. *The distribution function of a Gaussian mixture random variable is a weighted sum between the distributions of the components, i.e.:*

$$F_X(x) = p \cdot F_{X_1}(x) + (1 - p) F_{X_0}(x), \tag{29.2}$$

Taking the derivative, it can be easily shown that the density function reads:

$$f_X(x) = p \cdot f_{X_1}(x) + (1-p) f_{X_0}(x). \tag{29.3}$$

In general

$$f_X(x) = \frac{p}{\sigma_1} \cdot \phi\left(\frac{x-\mu_1}{\sigma_1}\right) + \frac{1-p}{\sigma_0} \cdot \phi\left(\frac{x-\mu_0}{\sigma_0}\right), \tag{29.4}$$

where Φ is the cumulative distribution function and ϕ is the density function of a standard normal random variable. An implementation of the density and distribution of a Gaussian Mixture is contained in the R package extraDistr, i.e. `dmixnorm` *for the density and* `pmixnorm` *for the distribution.*

i Proof: Proposition 29.1

Proof. From the formal definition of distribution function of a random variable X

$$F_X(y) = \mathbb{P}(X \leq y) = \mathbb{E}\{\mathbb{1}_{X \leq x}\},$$

where if X is a Gaussian Mixture with two components we can express it as conditional expectation with respect to B, i.e.

$$\begin{aligned} F_X(y) &= \mathbb{E}\{\mathbb{1}_{X \leq x} | B\} = \\ &= \mathbb{E}\{\mathbb{1}_{X \leq x} | B = 0\}\mathbb{P}(B = 0) + \mathbb{E}\{\mathbb{1}_{X \leq x} \mid B = 1\}\mathbb{P}(B = 1) = \\ &= p \cdot \mathbb{P}(X_1 \leq x) + (1-p) \cdot \mathbb{P}(X_0 \leq x) \end{aligned}$$

Hence, standardizing the Normal random variable one obtain

$$F_X(x) = p \cdot \Phi\left(\frac{x-\mu_1}{\sigma_1}\right) + (1-p) \cdot \Phi\left(\frac{x-\mu_0}{\sigma_0}\right),$$

where Φ denotes the distribution function of a standard normal. Knowing that $f_X(x) = \frac{dF_X(x)}{dx}$ and that $\phi_X(x) = \frac{d\Phi(x)}{dx}$, where ϕ is the density function of a standard normal we obtain the result, i.e.

$$f_X(x) = \frac{p}{\sigma_1} \cdot \phi\left(\frac{x-\mu_1}{\sigma_1}\right) + \frac{1-p}{\sigma_0} \cdot \phi\left(\frac{x-\mu_0}{\sigma_0}\right).$$

□

29.2 Moment generating function

Proposition 29.2. *The moment generating function of a Gaussian mixture random variable (Equation 29.1) in t reads:*

$$M_X(u) = p \cdot M_{X_1}(u) + (1-p) \cdot M_{X_0}(u).$$

where for a general $i \in \{0, 1\}$, $M_{X_i}(u)$ is the moment generating function of a Gaussian random variable with moments μ_i, σ_i^2, i.e.

$$M_{X_i}(u) = \exp\left\{\mu_i u + \frac{u^2\sigma_i^2}{2}\right\}$$

i Proof: Proposition 29.2

Proof. Applying the definition of moment generating function and the property of linearity of the expectation on a Gaussian mixture (Equation 29.1), one obtain:

$$\begin{aligned} M_X(u) &= p \cdot \mathbb{E}\{e^{uZ_1}\} + (1-p) \cdot \mathbb{E}\{e^{uZ_0}\} = \\ &= p \cdot M_{X_1}(u) + (1-p) \cdot M_{X_0}(u) \end{aligned}$$

Hence, the moment generating function of X is a linear combination of the moment generating functions of the two components. □

29.3 Esscher transform

Proposition 29.3. *The Esscher transform of a Gaussian mixture random variable reads:*

$$\mathcal{E}_\theta\{f_X\}(x) = p_1(\theta) f_{X_1}(x;\theta) + p_0(\theta) f_{X_0}(x;\theta),$$

where for $i \in \{0, 1\}$:

$$f_X(x;\theta) = \frac{1}{\sigma_i}\phi\left(\frac{x - \mu_i - \theta\sigma_i^2}{\sigma_i}\right),$$

and the distorted probabilities are defined as:

$$p_1(\theta) = p \cdot \frac{M_{X_1}(\theta)}{M_X(\theta)}, \quad p_0(\theta) = (1 - p_1(\theta)).$$

Proof: Proposition 29.3

Proof. In general, the Esscher transform of a density function f_X is defined as:

$$\mathcal{E}_\theta\{f_X\}(x) = \frac{e^{\theta x} f_X(x)}{M_X(\theta)} = \frac{e^{\theta x} f_X(x)}{\int_{-\infty}^{\infty} e^{\theta y} f_X(y) dy}.$$

Let's focus only on the numerator. Substituting the density function of a Gaussian mixture one obtain:

$$\mathcal{E}_\theta\{f_X\}(x) = \frac{e^{\theta X}(p f_{Z_1}(x) + (1-p) f_{Z_2}(x))}{M_X(\theta)}.$$

Let's consider the i-component for $i \in \{0, 1\}$ and let's explicit the density function, i.e.

$$e^{\theta x} f_{Z_i}(x) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{(x-\mu_i)^2}{2\sigma_i^2} + \theta x\right\}.$$

Let's expand the exponent of the exponential term for the i-th component, i.e.

$$\begin{aligned}\theta x - \frac{(x-\mu)^2}{2\sigma^2} &= \theta x - \frac{x^2 - 2\mu x + \mu^2}{2\sigma^2} = \\ &= -\frac{x^2}{2\sigma^2} + \left(\theta + \frac{\mu}{\sigma^2}\right) x - \frac{\mu^2}{2\sigma^2}\end{aligned}$$

Let's denote with $A = \theta + \frac{\mu}{\sigma^2}$, then complete the square

$$-\frac{x^2}{2\sigma^2} + Ax = -\frac{1}{2\sigma^2}\left[x^2 - 2\sigma^2 Ax + A^2\sigma^4\right] + \frac{A^2\sigma^2}{2} = \frac{(x - A\sigma^2)}{2\sigma^2} + \frac{A^2\sigma^2}{2},$$

Therefore

$$\begin{aligned}\theta x - \frac{(x-\mu_i)^2}{2\sigma_i^2} &= \frac{(x - \mu_i - \theta\sigma^2)}{2\sigma_i^2} - \frac{\mu_i^2}{2\sigma_i^2} + \frac{(\theta\sigma_i^2 + \mu)^2}{2\sigma_i^2} = \\ &= \frac{(x - \mu_i - \theta\sigma_i^2)}{2\sigma_i^2} - \frac{\mu_i^2}{2\sigma_i^2} + \frac{\theta^2}{2} + \frac{\mu_i^2}{2\sigma_i^2} + \theta\mu_i = \\ &= \frac{(x - \mu_i - \theta\sigma_i^2)}{2\sigma_i^2} + \frac{\theta^2}{2} + \theta\mu_i\end{aligned}$$

Hence, adding and subtracting inside the exponential one obtain

$$\begin{aligned}e^{\theta x} f_{Z_i}(x) &= \frac{1}{\sqrt{2\pi\sigma_i}} \exp\left\{-\frac{(x-\mu_i)^2}{2\sigma_i^2} + \theta x - \mu_i\theta - \frac{\theta^2\sigma_i^2}{2}\right\} \exp\left\{\mu_i\theta + \frac{\theta^2\sigma_i^2}{2}\right\} = \\ &= \frac{1}{\sqrt{2\pi\sigma_i}} \exp\left\{\mu_i\theta + \frac{\theta\sigma_i^2}{2}\right\} \exp\left\{-\frac{(x - \mu_i - \theta\sigma_i^2)^2}{2\sigma_i^2}\right\} = \\ &= M_{Z_i}(\theta) f_{Z_i}(x; \theta)\end{aligned}$$

Hence, we obtain the result

$$f_{Z_i}(x;\theta) = \frac{1}{\sigma_i}\phi\left(\frac{x-\mu_i-\theta\sigma_i^2}{\sigma_i}\right).$$

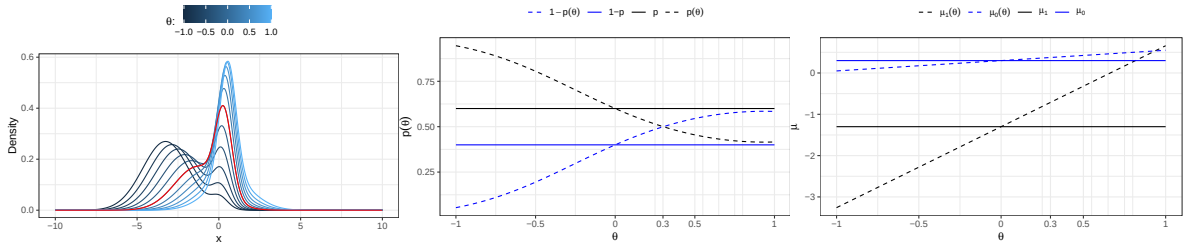
Hence, the Esscher density

$$\mathcal{E}_\theta\{f_X\}(x) = \frac{pM_{Z_1}(\theta)f_{Z_1}(x;\theta) + (1-p)M_{Z_0}(\theta)f_{Z_i}(x;\theta)}{M_X(\theta)}.$$

Hence let's collect the first part and the denominator (mgf of the Gaussian mixture in θ) and define the new probabilities

$$p_1(\theta) = \frac{pM_{X_1}(\theta)}{M_X(\theta)}, \quad p_0(\theta) = \frac{(1-p)M_{X_0}(\theta)}{M_X(\theta)}.$$

□



(a) Esscher densities for different values of θ in $[-1,1]$ and with $\theta = 0$ (red).

(b) Distorted mixture probability for different values of θ in $[-1,1]$ and with $\theta = 0$ (red).

(c) Distorted mixture means for different values of θ in $[-1,1]$ and with $\theta = 0$ (red).

Figure 29.2: Esscher transform of a Gaussian mixture.

29.4 Moments

Proposition 29.4. *The expectation of a Gaussian Mixture random variable (Equation 29.1) reads:*

$$\mathbb{E}\{X\} = p\mu_1 + (1-p)\mu_2, \tag{29.5}$$

and the second moment

$$\mathbb{E}\{X^2\} = p(\mu_1^2+\sigma_1^2) + (1-p)(\mu_2^2+\sigma_2^2). \tag{29.6}$$

Hence, the variance

$$\mathbb{V}\{X\} = p(1-p)(\mu_1-\mu_2)^2 + \sigma_1^2 p + \sigma_2^2(1-p). \tag{29.7}$$

Proof: Proposition 29.4

Proof. Given that X_1, X_0 and B are independent, the expectation is computed as:

$$\begin{aligned}\mathbb{E}\{X\} &= \mathbb{E}\{\mathbb{E}\{X \mid B\}\} = \\ &= \mathbb{E}\{X \mid B=1\}\mathbb{P}(B=1) + \mathbb{E}\{X \mid B=0\}\mathbb{P}(B=0) = \\ &= p\mathbb{E}\{X_1\} + (1-p)\mathbb{E}\{X_0\} = \\ &= p\mu_1 + (1-p)\mu_2\end{aligned}$$

The second moment is computed similarly to the first one, i.e.

$$\begin{aligned}\mathbb{E}\{X^2\} &= \mathbb{E}\{\mathbb{E}\{X^2 \mid B\}\} = \\ &= \mathbb{E}\{X^2 \mid B=1\}\mathbb{P}(B=1) + \mathbb{E}\{X^2 \mid B=0\}\mathbb{P}(B=0) = \\ &= \mathbb{E}\{B\}\mathbb{E}\{X_1^2\} + \mathbb{E}\{1-B\}\mathbb{E}\{X_0^2\} = \\ &= p\mathbb{E}\{X_1^2\} + (1-p)\mathbb{E}\{X_0^2\} = \\ &= p(\mu_1^2+\sigma_1^2) + (1-p)(\mu_2^2+\sigma_2^2)\end{aligned}$$

The variance, by definition, is given by:

$$\mathbb{V}\{X\} = \mathbb{E}\{X^2\} - \mathbb{E}\{X\}^2,$$

where the first moment squared is

$$\begin{aligned}\mathbb{E}\{X\}^2 &= \left(p\mu_1 + (1-p)\mu_2\right)^2 = \\ &= p^2\mu_1^2 + (1-p)^2\mu_2^2 + 2p(1-p)\mu_1\mu_2\end{aligned} \tag{29.8}$$

Hence the variance,

$$\begin{aligned}\mathbb{V}\{X\} &= p(\mu_1^2+\sigma_1^2) + (1-p)(\mu_2^2+\sigma_2^2) - p^2\mu_1^2 - (1-p)^2\mu_2^2 - 2p(1-p)\mu_1\mu_2 = \\ &= p\mu_1^2 + p\sigma_1^2 + \mu_2^2 + \sigma_2^2 - p\mu_2^2 - p\sigma_2^2 - p^2\mu_1^2 - (1-p)^2\mu_2^2 - 2p(1-p)\mu_1\mu_2 = \\ &= \mu_1^2 p(1-p) + p\sigma_1^2 + (1-p)\sigma_2^2 + p(1-p)\mu_2^2 - 2p(1-p)\mu_1\mu_2 = \\ &= p(1-p)(\mu_1^2 - \mu_2^2 - 2\mu_1\mu_2) + p\sigma_1^2 + (1-p)\sigma_2^2 = \\ &= p(1-p)(\mu_1-\mu_2)^2 + p\sigma_1^2 + (1-p)\sigma_2^2\end{aligned}$$

Equivalently, with the law of total variance:

$$\mathbb{V}\{X\} = \mathbb{V}\{\mathbb{E}\{X \mid B\}\} + \mathbb{E}\{\mathbb{V}\{X \mid B\}\}$$

where

$$\begin{aligned}\mathbb{E}\{\mathbb{V}\{X \mid B\}\} &= \mathbb{E}\{\sigma_1^2 B + \sigma_0^2(1-B)\} = \\ &= \sigma_1^2 p + \sigma_0^2(1-p)\end{aligned}$$

Then

$$\begin{aligned}\mathbb{E}\{X \mid B\} &= \mu_1 B + \mu_0(1-B) = \\ &= \mu_0 + (\mu_1 - \mu_0)B\end{aligned}$$

and therefore

$$\begin{aligned}\mathbb{V}\{\mathbb{E}\{X \mid B\}\} &= \mathbb{V}\{\mu_0 + (\mu_1 - \mu_0)B\} = \\ &= (\mu_1 - \mu_0)^2 \mathbb{V}\{B\} = \\ &= (\mu_1 - \mu_0)^2 p(1-p)\end{aligned}$$

The total variance:

$$\begin{aligned}\mathbb{V}\{X\} &= \mathbb{V}\{\mathbb{E}\{X \mid B\}\} + \mathbb{E}\{\mathbb{V}\{X \mid B\}\} = \\ &= (\mu_1 - \mu_0)^2 p(1-p) + \sigma_1^2 p + \sigma_0^2(1-p)\end{aligned}$$

□

29.4.1 Special Cases

Proposition 29.5. *If the random variable X (Equation 29.1) is centered in zero, i.e.*

$$\mathbb{E}\{X\} = p\mu_1 + (1-p)\mu_0 = 0,$$

then, the following expression holds

$$(\mu_1 - \mu_0)^2 p(1-p) = p\mu_1^2 + (1-p)\mu_0^2$$

ℹ Proof: Proposition 29.5

Proof. Let's show that the following expressions

$$\text{LHS} = p(1-p)(\mu_1 - \mu_0)^2 \equiv p\mu_1^2 + (1-p)\mu_0^2 = \text{RHS}$$

are equivalent under the constraint

$$\mathbb{E}\{X\} = p\mu_1 + (1-p)\mu_0 = 0$$

Firstly let's note that if the mixture is centered the ration between μ_1 and μ_0 is constant, i.e.

$$\mu_1 p + \mu_0(1-p) = 0 \implies \mu_1 = -\mu_0 r$$

where we define the ratio r as:

$$r = \frac{(1-p)}{p}$$

Let's now expand the LHS and substitute the relation between μ_1 and μ_0:

$$\begin{aligned}\text{LHS} &= (\mu_1 - \mu_0)^2 p(1-p) = \\ &= p(1-p)\mu_1^2 - 2p(1-p)\mu_1\mu_0 + p(1-p)\mu_0^2 = \\ &= p(1-p)\left(r^2\mu_0^2 + 2r\mu_0^2 + \mu_0^2\right) = \\ &= p(1-p)\mu_0^2(r^2 + 2r + 1) = \\ &= p(1-p)\mu_0^2(r+1)^2\end{aligned}$$

Then, we note that

$$(r+1) = \left(\frac{1-p}{p} + 1\right) = \frac{1-p+p}{p} = \frac{1}{p} \implies \text{LHS} = \mu_0^2 \frac{1-p}{p}$$

Now, let's consider the RHS

$$\begin{aligned}\text{RHS} &= p\mu_1^2 + (1-p)\mu_0^2 = \\ &= pr^2\mu_0^2 + (1-p)\mu_0^2 = \\ &= \mu_0^2(pr^2 + 1 - p) = \\ &= \mu_0^2 \frac{1-p}{p}\end{aligned}$$

since

$$\begin{aligned}(pr^2 + 1 - p) &= \left(p\frac{(1-p)^2}{p^2} + 1 - p\right) = \\ &= \left(\frac{(1-p)^2 + p(1-p)}{p}\right) = \\ &= (1-p)\left(\frac{1-p+p}{p}\right) = \\ &= \frac{1-p}{p}\end{aligned}$$

Hence, the RHS and LHS are equal. □

29.4.2 Central moments

Proposition 29.6. *The second central moment of a Gaussian mixture reads:*

$$\kappa_2\{X\} = (\delta_1^2 + \sigma_1^2)p + (\delta_0^2 + \sigma_0^2)(1-p) = \mathbb{V}\{X\}$$

where for $i \in \{0, 1\}$, $\delta_i = \mu_i - \mathbb{E}\{X\}$.

Proof: Proposition 29.6

Proof. Developing the squares:

$$\delta_1^2 = (\mu_1^2 + \mathbb{E}\{X\}^2 - 2\mu_1\mathbb{E}\{X\})$$
$$\delta_0^2 = (\mu_0^2 + \mathbb{E}\{X\}^2 - 2\mu_0\mathbb{E}\{X\})$$

and summing

$$\delta_1^2 p + \delta_0^2(1-p) = (\mu_1^2 p + \mu_0^2(1-p)) + \mathbb{E}\{X\}^2 - 2(\mu_1 p + \mu_0(1-p))\mathbb{E}\{X\} =$$
$$= (\mu_1^2 p + \mu_0^2(1-p)) - \mathbb{E}\{X\}^2$$

Thus, substituting the result in the initial expression one obtain the result. □

29.5 Estimation

29.5.1 Maximum likelihood

Minimizing the negative log-likelihood gives an estimate of the parameters, i.e.

$$\underset{\mu_1,\mu_2,\sigma_1,\sigma_2,p}{\operatorname{argmin}} \left\{ \sum_{i=1}^{t} \log(f_X(x_i)) \right\},$$

or equivalently maximizing the negative log-likelihood, i.e.

$$\underset{\mu_1,\mu_2,\sigma_1,\sigma_2,p}{\operatorname{argmax}} \left\{ -\sum_{i=1}^{t} \log(f_X(x_i)) \right\}.$$

Example: ML-estimate

Table 29.1: Maximum likelihood estimates for a Gaussian Mixture.

Parameter	True	Estimate	Log-lik	Bias
μ_1	-2.0	-1.9905803	-10332.56	0.0094197
μ_2	2.0	2.0006381	-10332.56	0.0006381
σ_1	1.0	1.0457653	-10332.56	0.0457653
σ_2	1.0	1.0033373	-10332.56	0.0033373
p	0.5	0.4937459	-10332.56	-0.0062541

29.5.2 Moments matching

Let's fix the parameter of the first component, namely μ_1 and σ_1^2 and a certain probability p. Then let's compute the sample estimate of the expectation of X_t, i.e.

$$\mathbb{E}\{X\} = \frac{1}{t}\sum_{i=1}^{t} x_i = \hat{\mu}$$

and the sample variance:

$$\mathbb{V}\{X\} = \frac{1}{t}\sum_{i=1}^{t}(x_i - \hat{\mu})^2 = \hat{\sigma}^2$$

In order to obtain an estimate of the second distribution such that the Gaussian Mixture moments match exactly the sample estimates we solve the system for μ_2 and σ_2^2:

$$\begin{cases} \hat{\mu} = p\mu_1 + (1-p)\mu_2 \\ \hat{\sigma}^2 = p(1-p)(\mu_1 - \mu_2)^2 + \sigma_1^2 p + \sigma_2^2(1-p) \end{cases}$$

which lead to a unique solution, i.e.

$$\begin{aligned} \mu_2 &= \frac{\hat{\mu} - p\mu_1}{1-p} \\ \sigma_2^2 &= \frac{\hat{\sigma}^2 - p\sigma_1^2}{1-p} - p(\mu_1 - \mu_2)^2 \end{aligned}$$

29.5.3 EM

To classify an existing empirical series into two groups (Bernoulli = 0 and Bernoulli = 1) such that the empirical properties (mean and variance) of the groups match the theoretical properties of the original normal distributions, we can use an Expectation-Maximization (EM) algorithm. Here we summarizes the steps and formulas used in the EM algorithm routine to classify an empirical series into two groups such that the empirical properties match the theoretical properties of two normal distributions.

Table 29.2: EM algorithm routine

Step	**Description**
Initialization	Initialize responsibilities and other parameters.
1. **E-step**	Calculate the responsibilities for each data point as: $\gamma_{i1} = \frac{p \cdot f(x_i\|\mu_1,\sigma_1)}{p \cdot f(x_i\|\mu_1,\sigma_1)+(1-p)\cdot f(x_i\|\mu_2,\sigma_2)}$, $\gamma_{i2} = \frac{(1-p)\cdot f(x_i\|\mu_2,\sigma_2)}{p \cdot f(x_i\|\mu_1,\sigma_1)+(1-p)\cdot f(x_i\|\mu_2,\sigma_2)}$. Compute $n_1 = \sum_{i=1}^{n}\gamma_{i1}$ and $n_2 = \sum_{i=1}^{n}\gamma_{i2}$.
2. **M-step**	Update the parameters using the calculated responsibilities:

Step	Description
	Means: $\mu_1 = \frac{1}{n_1}\sum_{i=1}^{n}\gamma_{i1}x_i, \quad \mu_2 = \frac{1}{n_2}\sum_{i=1}^{n}\gamma_{i2}x_i.$
	Variances:
	$\sigma_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}\gamma_{i1}(x_i-\mu_1)^2, \quad \sigma_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}\gamma_{i2}(x_i-\mu_2)^2$
	Bernoulli probability: $p = \frac{n_1}{n}$.
3. **Log-likelihood**	Calculate the log-likelihood for convergence check.
4. **Check convergence**	Check if the change in log-likelihood is below a threshold, otherwise come back to 1.
Output	Series of Bernoulli B_t and the optimal parameters $\{\mu_1, \mu_2, \sigma_1, \sigma_2, p\}$.

Table 29.3: Estimated Gaussian Mixture moments with EM

statistic	emp	opt	hat
Mean	2.516087	2.516087	2.516087
Std. Dev	2.958034	2.957904	2.957879

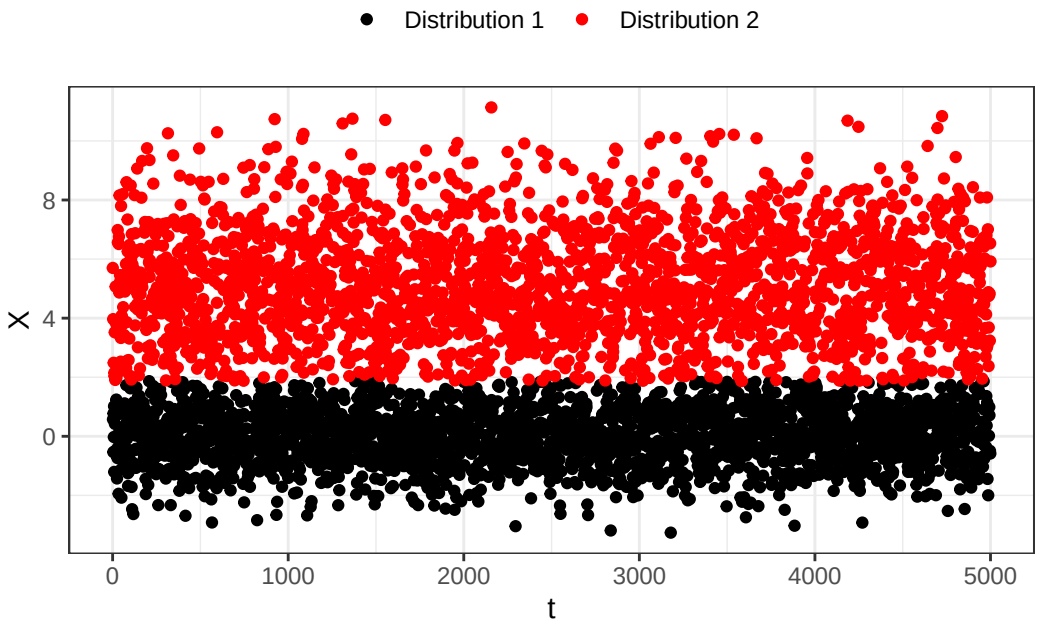


Figure 29.3: Classified simulated series with EM

29.5.4 Matrix moments matching

Proposition 29.7. *Any finite K-component Gaussian mixture with finite moments admits a more parsimonious moment-matching approximation with only two-component Gaussian mixture. We start with the parameters of the mixture at time $t+h$ and we adjust them to match the first three moments of the multinomial mixture. The procedure ensures that the resulting distribution will have mean:*

$$\begin{aligned}\mathbb{E}\{U_{t+h} \mid \mathcal{F}_t\} &= M(t,0,h)\\ \mathbb{V}\{U_{t+h} \mid \mathcal{F}_t\} &= S(t,0,h)\\ \mathbb{E}\{U_{t+h} \mid \mathcal{F}_t\} &= \Omega(t,0,h)\end{aligned}$$

where

$$\Omega(t,0,h) = \mathbb{E}\{U^3_{t+h} \mid \mathcal{F}_t\} = \sum_{j=0}^{h-1} \mathbb{E}\{\psi^3_{t+h-j} \mid \mathcal{F}_t\}\mathbb{E}\{u^3_{t+h-j}\}$$

In general, the variance and the skewness do not converge to a constant number for all t, but will depends on the skewness of the starting point at $t+1$ till the ending point at $t+h$ and needs to be recomputed each time. The parameters of the resulting mixture will be:

$$\mu^\star_{1,t+h} = \sqrt{\Sigma_{t+h}}\ \mu_{1,t+h} \quad \mu^\star_{0,t+h} = \sqrt{\Sigma_{t+h}}\ \mu_{0,t+h}$$

and variances:

$$\sigma^{\star 2}_{1,t+h} = \frac{\left(\Sigma_{t+h} - p_{t+h}\mu^{\star 2}_{1,t+h} - (1-p_{t+h})\mu^{\star 2}_{0,t+h}\right)\cdot 3\mu^\star_{0,t+h}(1-p_{t+h}) - \left(\Omega_{t+h} - p_{t+h}\mu^{\star 3}_{1,t+h} - (1-p_{t+h})\mu^{\star 3}_{0,t+h}\right)(1}{3p_{t+h}(1-p_{t+h})(\mu^\star_{0,t+h} - \mu^\star_{1,t+h})} \tag{29.9}$$

and

$$\sigma^{\star 2}_{0,t+h} = \frac{\left(\Omega_{t+h} - p_{t+h}\mu^{3*}_{1,t+h} - (1-p_{t+h})\mu^{3*}_{0,t+h}\right)p_{t+h} - 3\mu^\star_{1,t+h}\left(\Sigma_{t+h} - p_{t+h}\mu^{\star 2}_{1,t+h} - (1-p_{t+h})\mu^{2*}_{0,t+h}\right)p_{t+h}}{3p_{t+h}(1-p_{t+h})(\mu^*_{0,t+h} - \mu^*_{1,t+h})} \tag{29.10}$$

Proof. Let's consider the following approach. We start by fixing the mixture probabilities at time $t+h$, i.e. p_{t+h}. Then, we consider the means and variances parameters of the mixture at time $t+h$ free to vary and we adjust them to match the first three central moments of the true multinomial mixture with a two component Gaussian mixture. More precisely, let's define:

$$\mu^*_{i,t+h} = \mu_{Y_i}(t,h), \quad \sigma^{2*}_{i,t+h} = \sigma^2_{Y_i}(t,h) \tag{29.11}$$

Recalling the central moments as in **?@eq-proof-ut-moments** we have that, with the parameters defined as in Equation 29.11, the resulting expected value and variance already matches the exact expectation and variance of the multinomial mixture. To improve the match between the two distributions, one can explicit also the third central moment, i.e.

$$\omega_{t+h} = (3\sigma^{2*}_{1,t+h}\mu^*_{1,t+h} + \mu^{3*}_{1,t+h}p_{t+h}) + (3\sigma^{2*}_{0,t+h}\mu^*_{0,t+h} + \mu^{3*}_{0,t+h})(1-p_{t+h})$$

In this way, one can set a system to adjust the variances $\sigma^{*2}_{i,t+h}$ such that the second and the third central moments of the Gaussian mixture with two components matches the ones of the multinomial mixture. More precisely, the third central moment of y_{t+h} reads:

$$\kappa_3\{y_{t+h} \mid \mathcal{F}_t\} = \sum_{j=s}^{h-1} \mathbb{E}\{\psi^3_{t+h-j} \mid \mathcal{F}_t\}\mathbb{E}\{u^3_{t+h-j}\}.$$

Let's denote the target moments as:

$$\Sigma_{t+h} = \kappa_2\{y_{t+h} \mid \mathcal{F}_t\} \quad \Omega_{t+h} = \kappa_3\{y_{t+h} \mid \mathcal{F}_t\}$$

and let's represent the system in matrix form

$$\underbrace{\begin{pmatrix} p_{t+h} & 1-p_{t+h} \\ 3p_{t+h}\mu^*_{1,t+h} & 3(1-p_{t+h}\mu^*_{0,t+h}) \end{pmatrix}}_{\mathbf{D}} \underbrace{\begin{pmatrix} \sigma^{2\star}_{1,t+h} \\ \sigma^{2\star}_{0,t+h} \end{pmatrix}}_{\mathbf{\Sigma^\star_{t+h}}} = \underbrace{\begin{pmatrix} \Sigma_{t+h} - p_{t+h}\mu^{2*}_{1,t+h} - (1-p_{t+h})\mu^{2*}_{0,t+h} \\ \Omega_{t+h} - p_{t+h}\mu^{3*}_{1,t+h} - (1-p_{t+h})\mu^{2*}_{0,t+h} \end{pmatrix}}_{\mathbf{G}}$$

The solution of the system has the form $\mathbf{D}\Sigma^\star = \mathbf{G} \iff \Sigma^\star = \mathbf{D}^{-1}\mathbf{G}$. The determinant of the matrix $\mathbf{D}$ is different from zero only if $\mu^*_{1,t+h} \neq \mu^*_{0,t+h}$, i.e.

$$\det(\mathbf{D}) = 3p_{t+h}(1-p_{t+h})(\mu^*_{0,t+h} - \mu^*_{1,t+h}).$$

By applying Cramer's rule the system can be solved explicitly for $i \in \{0, 1\}$, i.e.

$$\sigma^{2\star}_{i,t+h} = \frac{\det(\mathbf{D}_i)}{\det(\mathbf{D})} \tag{29.12}$$

Where $\mathbf{D}_1$ is obtained by replacing the first column of $\mathbf{D}$ with the first column of $\mathbf{G}$, i.e.

$$\mathbf{D}_1 = \begin{pmatrix} \Sigma_{t+h} - p_{t+h}\mu^{2*}_{1,t+h} - (1-p_{t+h})\mu^{2*}_{0,t+h} & 1-p_{t+h} \\ \Omega_{t+h} - p_{t+h}\mu^{3\star}_{1,t+h} - (1-p_{t+h})\mu^{3*}_{0,t+h} & 3(1-p_{t+h})\mu^*_{0,t+h} \end{pmatrix}$$

Then:

$$\begin{aligned}\det(\mathbf{D}_1) = &\left(\Sigma_{t+h} - p_{t+h}\mu^{\star 2}_{1,t+h} - (1-p_{t+h})\mu^{\star 2}_{0,t+h}\right) \cdot 3(1-p_{t+h})\mu^\star_{0,t+h} + \\ &- (1-p_{t+h})\left(\Omega_{t+h} - p_{t+h}\mu^{\star 3}_{1,t+h} - (1-p_{t+h})\mu^{\star 3}_{0,t+h}\right)\end{aligned}$$

Similarly for the second component $\mathbf{D}_0$ is obtained by replacing the second column of $\mathbf{D}$ with the second column of $\mathbf{G}$, i.e.

$$\mathbf{D}_0 = \begin{pmatrix} p_{t+h} & \Sigma_{t+h} - p_{t+h}\mu^{\star 2}_{1,t+h} - (1-p_{t+h})\mu^{\star 2}_{0,t+h} \\ 3p_{t+h}\mu^\star_{1,t+h} & \Omega_{t+h} - p_{t+h}\mu^{\star 3}_{1,t+h} - (1-p_{t+h})\mu^{\star 3}_{0,t+h} \end{pmatrix}$$

Then:

$$\begin{aligned}\det(\mathbf{D}_0) = &\, p_{t+h} \cdot \left(\Omega_{t+h} - p_{t+h}\mu^{\star 3}_{1,t+h} - (1-p_{t+h})\mu^{\star 3}_{0,t+h}\right) + \\ &- 3p_{t+h}\mu^\star_{1,t+h} \cdot \left(\Sigma_{t+h} - p_{t+h}\mu^{\star 2}_{1,t+h} - (1-p_{t+h})\mu^{\star 2}_{0,t+h}\right)\end{aligned}$$

Substituting and developing Equation 29.12, one obtain the explicit solutions of the system. □

Part VIII

Appendix

30 Calculus

30.1 Fundamental limits

Table 30.1: Fundamental limits

$\lim_{x\to 0} \frac{\sin x}{x} = 1$	$\lim_{x\to 0} \frac{\ln(1+x)}{x} = 1$	$\lim_{x\to 0} \frac{e^x-1}{x} = 1$
$\lim_{x\to 0} \left(1+\frac{1}{x}\right)^x = e$	$\lim_{x\to 0} \frac{\log_a(1+x)}{x} = \log_a e$	$\lim_{x\to 0} \frac{a^x-1}{x} = \ln a$

30.2 Derivatives

Table 30.2: Fundamental derivatives

Function $f(x)$	**Derivative** $f'(x)$
$y = a,\ a \in \mathbb{R}$	$y' = 0$
$y = x^n,\ n \in \mathbb{N}$	$y' = nx^{n-1}$
$y = x^\alpha,\ \alpha \in \mathbb{R}$	$y' = \alpha x^{\alpha-1}$
$y = x^{\frac{1}{n}},\ n > 0$	$y' = \frac{1}{n} x^{\frac{1}{n}-1}$
$y = \sin x$	$y' = \cos x$
$y = \cos x$	$y' = -\sin x$
$y = \tan x$	$y' = \frac{1}{\cos^2 x} = 1 + \tan^2 x$
$y = \cot x$	$y' = -\frac{1}{\sin^2 x} = -(1 + \cot^2 x)$
$y = \arcsin x$	$y' = \frac{1}{\sqrt{1-x^2}}$
$y = \arccos x$	$y' = -\frac{1}{\sqrt{1-x^2}}$
$y = \arctan x$	$y' = \frac{1}{1+x^2}$
$y = \operatorname{arccot}\ x$	$y' = -\frac{1}{1+x^2}$
$y = a^x$	$y' = a^x \ln a$
$y = e^x$	$y' = e^x$
$y = \log_a x$	$y' = \frac{1}{x \ln a}$
$y = \ln x$	$y' = \frac{1}{x}$
$f(x) = c \cdot g(x)$	$f'(x) = c \cdot g'(x)$
$f(x) = g(x) + s(x)$	$f'(x) = g'(x) + s'(x)$
$f(x) = \frac{1}{g(x)}$	$f'(x) = -\frac{g'(x)}{g(x)^2}$

Function $f(x)$	Derivative $f'(x)$

- **Derivative of the product**:

$$[f(x) \cdot g(x)]'(x) = f'(x)g(x) + f(x)g'(x)$$

- **Derivative of the ratio**:

$$\left[\frac{f(x)}{g(x)}\right]'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{[g(x)]^2}$$

- **Derivative of the composition**:

$$[f(g(x))]'(x) = g'(x) \cdot f'(g(x))$$

Example derivative of the composition

For example $f(g(x)) = \ln(1+2x)$, then $f(x) = \ln(x)$ and $g(x) = 1+2x$, hence

$$[\ln(1+2x)]'(x) = [1+2x]'(x) \cdot [\ln(x)]'(1+x)$$

30.2.1 Taylor series

$$f(x) = \sum_{n=1}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + ... \tag{30.1}$$

30.3 Integrals

Table 30.3: Fundamental integrals

Immediate	General
$\int x^n dx = \frac{x^{n+1}}{n+1} + c$	$\int f(x)^n dx = \frac{f(x)^{n+1}}{n+1} + c$
$\int \frac{1}{x} dx = \log(x) + c$	$\int \frac{f'(x)}{f(x)} dx = \log(f(x)) + c$
$\int a^x dx = \log_a(e) + c$	$\int a^{f(x)} f'(x) dx = a^{f(x)} \log_a(e) + c$
$\int e^x dx = e^x + c$	$\int e^x f'(x) dx = e^{f(x)} + c$
$\int \sin(x) dx = -\cos(x) + c$	$\int \sin(f(x)) f'(x) dx = -\cos(f(x)) + c$
$\int \cos(x) dx = \sin(x) + c$	$\int \cos(f(x)) f'(x) dx = \sin(f(x)) + c$

Immediate	General
$\int \frac{1}{\cos(x)^2} dx = \tan(x) + c$	$\int \frac{f'(x)}{\cos(f(x))^2} dx = \tan(f(x)) + c$
$\int \frac{1}{\sin(x)^2} dx = \cot(x) + c$	$\int \frac{f'(x)}{\sin(f(x))^2} dx = \cot(f(x)) + c$
$\int \frac{1}{\sqrt{1-x^2}} dx = \arcsin(x) + c$	$\int \frac{f'(x)}{\sqrt{1-f(x)^2}} dx = \arcsin(f(x)) + c$
$\int \frac{1}{\sqrt{a^2-x^2}} dx = \arcsin(\frac{x}{a}) + c$	$\int \frac{f'(x)}{\sqrt{a^2-f(x)^2}} dx = \arcsin(\frac{f(x)}{a}) + c$
$\int \frac{1}{\sqrt{1+x^2}} dx = \arctan(x) + c$	$\int \frac{f'(x)}{\sqrt{1+f(x)^2}} dx = \arctan(f(x)) + c$

30.3.1 Fundamental theorem

$$f(b) - f(a) = \int_a^b f'(x)dx \iff \int f'(x)dx = f(x) + C$$

30.3.2 Integration by parts

$$\int_a^b f(x)g'(x)dx = [f(x)g(x)]_{x=a}^{x=b} - \int_a^b f'(x)g(x)dx$$

or in compact form:

$$\int_a^b f(x)dg(x) = [f(x)g(x)]_{x=a}^{x=b} - \int_a^b g(x)df(x)dx$$

31 Probability

Definition 31.1. (Absolutely continuous)
Consider a measure space $(\Omega, \mathcal{B})$, then a measure μ is said to be absolutely continuous with respect to ν, namely $\mu << \nu$, iff:

$$\mu << \nu \iff \mu(B) = 0 \implies \nu(B) = 0 \ \ \forall B \in \mathcal{B}$$

Definition 31.2. (Concentration)
Consider a measure space $(\Omega, \mathcal{B})$, then a measure μ **concentrates** on $B \in \mathcal{B}$, if $\mu(B^c) = 0$.

Definition 31.3. (Mutually singular)
Consider a measure space $(\Omega, \mathcal{B})$, then two measures μ and ν are said to be **mutually singular** if for any disjoint set $A \cap B = \emptyset$, $A, B \in \mathcal{B}$ we have that μ concentrates on A and ν concentrates on B.

Definition 31.4. (σ-finite)
Consider a measure space $(\Omega, \mathcal{B})$, then a measure μ is said to be σ-finite if exists a countable partition $B_1, B_2, \ldots, \subset \Omega$ such that $\forall i$ we have that $B_i \in \mathcal{B}$ and $\mu(B_i) < \infty$. In other words, a measure is σ-finite, when we are able to divide the sample space in a countable partition of sets such that each one is in $\mathcal{B}$ and has finite measure. Note that this do not imply that the measure of Ω is finite. In fact, consider for example the Lebesgue measure m on $\mathbb{R}$, then we obtain $m(\mathbb{R}) = \infty$. However, since we can partition $\mathbb{R}$ in a countable series of intervals, each one with finite length, then the Lebesgue measure is σ-finite.

32 Linear Algebra

32.1 Vector multiplication

Consider a vector of this form,

$$\underset{n\times 1}{\mathbf{y}} = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix},$$

then

$$\underset{n\times 1}{\mathbf{y}}\ \underset{n\times 1}{\mathbf{y}}^\top = \underset{n\times 1}{\mathbf{y}}\ \underset{1\times n}{\mathbf{y}} = \underset{n\times n}{\mathbf{y}} = \begin{pmatrix} y_1^2 & \cdots & y_1 y_i & \cdots & y_1 y_n \\ \vdots & \ddots & \vdots & & \vdots \\ y_i y_1 & \cdots & y_i^2 & \cdots & y_i y_n \\ \vdots & & \vdots & \ddots & \vdots \\ y_n y_1 & \cdots & y_n y_i & \cdots & y_n^2 \end{pmatrix},$$

and

$$\underset{n\times 1}{\mathbf{y}}^\top\ \underset{n\times 1}{\mathbf{x}} = \underset{1\times n}{\mathbf{y}}\ \underset{n\times 1}{\mathbf{y}} = \sum_{i=1}^{n} y_i^2 = \underset{1\times 1}{y}.$$

32.2 Matrix multiplication

Consider a matrix of this form,

$$\underset{n\times k}{\mathbf{X}} = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{11} & \cdots & x_{ij} & \cdots & x_{ik} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_k \end{pmatrix},$$

then

$$\underset{k\times k}{\mathbf{M}} = \underset{n\times k}{\mathbf{X}}^{\top} \underset{n\times k}{\mathbf{X}} = \begin{pmatrix} \sum_{i=1}^{n} x_{i1}^2 & \cdots & \sum_{i=1}^{n} x_{i1}x_{ij} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \ddots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ij}x_{i1} & \cdots & \sum_{i=1}^{n} x_{ij}^2 & \cdots & \sum_{i=1}^{n} x_{ij}x_{ik} \\ \vdots & & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{ik}x_{i1} & \cdots & \sum_{i=1}^{n} x_{ik}x_{ij} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{pmatrix},$$

and

$$\underset{k\times 1}{\mathbf{m}} = \underset{n\times k}{\mathbf{X}}^{\top} \underset{n\times 1}{\mathbf{y}} = \begin{pmatrix} \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ij}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{pmatrix}.$$

32.3 Special matrices

32.3.1 Basis vector

$$\mathbf{e}_p^{\top} = \begin{pmatrix} 1 & 0 & \ldots & 0 \end{pmatrix}. \tag{32.1}$$

The standard basis vector of length p with 1 in the first position and 0 elsewhere.

32.3.2 Matrix of ones

In mathematics, a matrix of ones (also called an all-ones matrix) is a matrix where every entry equals 1, i.e.

$$\begin{aligned} \mathbf{J}_2 &= \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} & \mathbf{J}_3 &= \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \\ \mathbf{J}_{3,2} &= \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} & \mathbf{J}_{2,3} &= \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \end{aligned} \tag{32.2}$$

In general, When two indices are provided, e.g., $\mathbf{J}_{3,2}$, the first indicates the number of rows and the second the number of columns. When only one index is given, e.g., $\mathbf{J}_3$, it denotes a square matrix of size 3×3. Some basic matrix operations include:

- $\mathbf{J}_{n,1} \cdot \mathbf{J}_{1,n} = \mathbf{J}_n$
- $\mathbf{J}_{1,n} \cdot \mathbf{J}_{n,1} = \mathbf{J}_1 = 1$

32.3.3 Identity matrix

In linear algebra, the identity matrix of size n is the $n \times n$ square matrix with ones on the main diagonal and zeros elsewhere, i.e.

$$\mathbf{I}_1 = (1)\,,\ \mathbf{I}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ \mathbf{I}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},\ \mathbf{I}_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{32.3}$$

32.4 Determinant

The determinant is a scalar value that can be computed from a square matrix. It provides important information about the matrix, such as whether it is invertible, its volume-scaling factor in linear transformations, and the linear dependence of its rows or columns. For a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, the determinant is denoted $\det(\mathbf{A})$. Consider two matrices $\mathbf{A}_{n \times n}$ and $\mathbf{B}_{n \times n}$, then the determinant satisfies some properties.

1. **Scalar**: $a \det(\mathbf{A}) = a^n \det(\mathbf{A})$ for $a \in \mathbb{R}$.
2. **Transpose**: $\det(\mathbf{A}^\top) = \det(\mathbf{A})$.
3. **Multiplication**: $\det(\mathbf{AB}) = \det(\mathbf{A}) \det(\mathbf{B})$.
4. **Inverse**: $\det(\mathbf{A}^{-1}) = \frac{1}{\det(\mathbf{A})}$.
5. **Rank**: if
   - $\det(\mathbf{A}) \neq 0$ then $rank(\mathbf{A}) = \max = n$.
   - $\det(\mathbf{A}) = 0$ then $rank(\mathbf{A}) < \max = n$.

The determinant of an $n \times n$ matrix can be computed by expanding along any row or column. Expanding along the i-th row gives:

$$\det(\mathbf{A}) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(\mathbf{M}_{ij}). \tag{32.4}$$

where $\mathbf{M}_{ij}$ is the sub-matrix without the i-th row and the j-th column. For example considering a 2×2 matrix, the determinant simplifies to a well-known formula:

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \implies \det(\mathbf{A}) = ad - bc. \tag{32.5}$$

Determinant of a 3×3 matrix with Laplace recursion

Let's consider a generic 3×3 matrix, i.e.

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix},$$

then, let's fix the row $i = 1$ and develop the Laplace expansion (Equation 32.4), i.e.

$$\begin{aligned} \det(\mathbf{A}) &= (-1)^{1+1} a_{11} \mathbf{M}_{11} + (-1)^{1+2} a_{12} \mathbf{M}_{12} + (-1)^{1+3} a_{13} \mathbf{M}_{13} = \\ &= a_{11} \mathbf{M}_{11} - a_{12} \mathbf{M}_{12} + a_{13} \mathbf{M}_{13} \end{aligned} \tag{32.6}$$

where the sub-matrices $\mathbf{M}_{11}$, $\mathbf{M}_{12}$ and $\mathbf{M}_{13}$ read explicitly as:

$$\mathbf{M}_{11} = \begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix}, \quad \mathbf{M}_{12} = \begin{pmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{pmatrix}, \quad \mathbf{M}_{13} = \begin{pmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}.$$

Then, the determinant of 2×2 matrices is easily computable as:

$$\begin{aligned} \det(\mathbf{M}_{11}) &= \det \begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix} = a_{22} a_{33} - a_{23} a_{32} \\ \det(\mathbf{M}_{12}) &= \det \begin{pmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{pmatrix} = a_{21} a_{33} - a_{23} a_{31} \\ \det(\mathbf{M}_{13}) &= \det \begin{pmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} = a_{21} a_{32} - a_{22} a_{31} \end{aligned} \tag{32.7}$$

Finally, coming back to Equation 32.6 and substituting the result in Equation 32.7 one obtain:

$$\det(\mathbf{A}) = a_{11}(a_{22} a_{33} - a_{23} a_{32}) - a_{12}(a_{21} a_{33} - a_{23} a_{31}) + a_{13}(a_{21} a_{32} - a_{22} a_{31}).$$

32.5 Trace

The trace of a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is the sum of its diagonal elements. It is also equal to the sum of the eigenvalues λ_i of $\mathbf{A}$, counted with algebraic multiplicity, i.e.

$$\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii} = \sum_{i=1}^{n} \lambda_i. \tag{32.8}$$

Some properties of the trace operator includes:

1. $\text{tr}(\mathbf{A}^\top) = \text{tr}(\mathbf{A})$.
2. $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B})$.
3. $\text{tr}(a\mathbf{A}) = a \cdot \text{tr}(\mathbf{A})$ for $a \in \mathbb{R}$.
4. $\text{tr}(\mathbf{A}^n) = \sum_{i=1}^{n} \lambda_i^n$ where λ_i is the i-th eigenvalue of $\mathbf{A}$.
5. $\text{tr}(\mathbf{A}^{-1}) = \sum_{i=1}^{n} \frac{1}{\lambda_i}$.

33 Notable relations between distributions

33.1 Chi squared

The chi2 distribution with ν degrees of freedom, namely $\chi^2(\nu)$, is defined as the sum of ν-independent and identically distributed standard normal random variables squared for $i = 1, \dots, \nu$, i.e.

$$\begin{cases} Z_i \sim \mathcal{N}(0,1) & \forall i \\ Z_i \perp Z_j & \forall i \neq j \end{cases} \implies X = Z_1^2 + \cdots + Z_\nu^2 \sim \chi^2(\nu) \tag{33.1}$$

The chi-squared distribution $\chi^2(\nu)$ is a special case of the gamma distribution, i.e.

$$X \sim \chi^2(\nu) \iff X \sim \text{Gamma}\left(\alpha = \frac{\nu}{2}, \theta = 2\right)$$

> **Sum of χ^2 random variables**
>
> The sum of two χ^2 is again χ^2 if and only if $\chi^2_{\nu_1}$ and $\chi^2_{\nu_2}$ are independent, formally:
>
> $$\chi^2(\nu_1) \perp \chi^2(\nu_2) \implies \chi^2(\nu_1) + \chi^2(\nu_2) \sim \chi^2(\nu_1 + \nu_2)$$
>
> If they are not independent their sum is not χ^2 distributed.

33.1.1 Moments

Table 33.1: Moments of a χ^2 random variable

Expectation	Variance	Skewness	Excess Kurtosis
ν	2ν	$\sqrt{\frac{8}{\nu}}$	$\frac{12}{\nu}$

33.1.2 Relations with others distributions

1. $\frac{1}{\nu}\chi^2(\nu) \xrightarrow[\nu\to\infty]{p} 1$.
2. $\frac{\chi^2(\nu)-\nu}{\sqrt{2\nu}} \xrightarrow[\nu\to\infty]{d} \mathcal{N}(0,1)$.
3. If $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ then $\mathbf{x}^T\Sigma^{-1}\mathbf{x} \sim \chi^2(k)$.
4. A generalization of property 3. to non-central distributions: if $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$ then $\mathbf{x}^T\Sigma^{-1}\mathbf{x} \sim \chi^2(k, \delta)$ where $\delta = \mu^T\Sigma\mu$.

33.2 Student-t

The Student-t distribution with ν degrees of freedom, namely $t(\nu)$, is defined as the ratio of two independent random variables. In specific, a standard normal random variable Z and the square root of a $\chi^2(\nu)$ divided by its degrees of freedom ν, i.e.

$$\begin{cases} Z \sim \mathcal{N}(0,1) \\ V \sim \chi^2(\nu) \\ Z \perp V \end{cases} \implies X = \frac{\sqrt{\nu}\, Z}{\sqrt{V}} \sim t(\nu) \tag{33.2}$$

Given a location parameter μ and a scale parameter σ^2 the Student-t random variable admits the following stochastic representation:

$$X = \mu + \sigma\frac{\sqrt{\nu}\, Y}{\sqrt{V}} \sim t(\mu, \sigma^2, \nu)$$

33.2.1 Moments

Table 33.2: Moments of a Student-t random variable $X \sim t(\mu, \sigma, \nu)$.

Expectation	Variance	Skewness	Excess Kurtosis
μ	$\frac{\nu}{\nu-2}\sigma^2$	0	$\frac{6}{\nu-4}$, $\nu > 4$

33.2.2 Relations with others distributions

1. $t(\nu) \xrightarrow[\nu\to\infty]{d} \mathcal{N}(0,1)$.
2. $t(\nu)^2 \equiv F(1,\nu)$.

33.3 Fisher–Snedecor

The Fisher–Snedecor distribution with ν_1 and ν_2 degrees of freedom, often denoted as F, is defined as the ratio of two independent chi2 random variables, each one divided by its degrees of freedom, i.e.

$$\begin{cases} V_1 \sim \chi^2(\nu_1) \\ V_2 \sim \chi^2(\nu_2) \end{cases} \implies X = \frac{\frac{V_1}{\nu_1}}{\frac{V_2}{\nu_2}} \sim \mathrm{F}(\nu_1, \nu_2) \tag{33.3}$$

33.3.1 Relations with others distributions

1. $\nu_2 \, \mathrm{F}(\nu_1, \nu_2) \xrightarrow[\nu_2 \to \infty]{d} \chi^2(\nu_1)$.

References

Anscombe, F. J., and William J. Glynn. 1983. "Distribution of the Kurtosis Statistic B2 for Normal Samples." *Biometrika* 70 (1): 227–34. https://doi.org/10.2307/2335960.

Bollerslev, Tim. 1986. "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics* 31 (3): 307–27. https://doi.org/https://doi.org/10.1016/0304-4076(86)90063-1.

D'Agostino, Ralph, and E. S. Pearson. 1973. "Tests for Departure from Normality. Empirical Results for the Distributions of B2 and √B1." *Biometrika* 60 (3): 613–22. https://doi.org/10.1093/biomet/60.3.613.

Jarque, Carlos M., and Anil K. Bera. 1980. "Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals." *Economics Letters* 6 (3): 255–59. https://doi.org/https://doi.org/10.1016/0165-1765(80)90024-5.

Nelson, Daniel B. 1990. "Stationarity and Persistence in the GARCH(1,1) Model." *Econometric Theory* 6 (3): 318–34. http://www.jstor.org/stable/3532198.

Pearson, Egon S. 1931. "Note on Tests for Normality." *Biometrika* 22 (3/4): 423–24. https://doi.org/10.2307/2332104.

Ralph B. D'agostino, Albert Belanger, and Ralph B. D'agostino Jr. 1990. "A Suggestion for Using Powerful and Informative Tests of Normality." *The American Statistician* 44 (4): 316–21. https://doi.org/10.1080/00031305.1990.10475751.

Urzúa, Carlos M. 1996. "On the Correct Use of Omnibus Tests for Normality." *Economics Letters* 53 (3): 247–51. https://doi.org/10.1016/S0165-1765(96)00923-8.

Welch, B. L. 1938. "The Significance of the Difference Between Two Means When the Population Variances Are Unequal." *Biometrika* 29 (3/4): 350–62. https://doi.org/10.1093/biomet/29.3-4.350.

———. 1947. "The Generalization of "Student's" Problem When Several Different Population Variances Are Involved." *Biometrika* 34 (1-2): 28–35. https://doi.org/10.1093/biomet/34.1-2.28.

Wold, Herman. 1939. "A Study in the Analysis of Stationary Time Series. By Herman Wold." *Journal of the Institute of Actuaries* 70 (1): 113–15. https://doi.org/10.1017/S0020268100011574.