Statistics and Econometrics

Notes for a Graduate Sequence in Econometrics

Paolo Zacchia

January 28, 2022

Contents

Part I

Probability and Statistics

Lecture 1

Random Variables

This lecture is a self-contained introduction to basic probability theory, including random variables and univariate probability distribution functions. In reviewing concepts that are fundamental towards later subjects, special care and emphasis are placed on establishing notation conventions that are adopted throughout all lectures. Examples are chosen so to facilitate a more extensive treatment of univariate and multivariate probability distributions, which constitute the subject of later lectures.

1.1 Events and Probabilities

Probability theory is a branch of mathematics that concerns the analysis of phenomena of uncertain occurrence. It constitutes a common mathematical framework for the measurement of the odds that some specific phenomena manifests themselves in the real world – their **probability** – at any point in time: past, present and future. As such, probability theory is the foundation of **statistics** and related disciplines such as econometrics.

The elaboration of probability theory requires a common mathematical characterization for all the possible occurrences of the phenomena that are potentially subject to its analysis. For this sake, probability theory borrows from **set theory** and models phenomena as, indeed, sets of alternatives.

Definition 1.1. Sample Space. The set $\mathbb{S}$ collecting all possible outcomes associated with a certain phenomenon is called the **sample space**.

A basic example of a sample space is the one associated with the classical experiment about tossing a coin: $\mathbb{S}_{coin} = \{Head, Tail\}$. A more expanded sample space is that of grades in a university exam: with letter grades for example (but not allowing for plus and minus), $\mathbb{S}_{exam} = \{A, B, C, D, E, F\}$.

The former are both examples of **countable sample spaces** characterized by a finite number of elements. There are also countable sample spaces with an **infinite** number of events: for example, the number of emails that one receives during a day can be expressed as $\mathbb{S}_{emails} = \{0, 1, 2, \dots\} = \mathbb{N}_0$. Other phenomena are modeled through **uncountable sample spaces** with an infinite number of elements. For example, the sample space associated with the income of an individual is represented by the nonnegative portion of the real line: $\mathbb{S}_{income} = \mathbb{R}_+$, whereas net wealth (assets minus liabilities) can also be negative: $\mathbb{S}_{wealth} = \mathbb{R}$. One can definitely construct even more complex, multidimensional sample spaces. For example, the net wealth of a household of two with separate financial positions is the collection of two numbers. It follows that $\mathbb{S}_{household} = \mathbb{S}_{wealth.1} \times \mathbb{S}_{wealth.2} = \mathbb{R}^2$.

The characterization of phenomena as sets of occurrences allows for a suitable definition of **events**, that is, combinations of occurrences.

Definition 1.2. Event. Any subset of a sample space $\mathbb{S}$, including $\mathbb{S}$ itself, is an **event**.

The definition of events as subsets allows to think about the probability of well-defined *groups* of alternatives. In the coin case, there are four events: $\mathbb{A}_{null} = \varnothing$, $\mathbb{A}_{head} = \{Head\}$, $\mathbb{A}_{tail} = \{Tail\}$, $\mathbb{A}_{full} = \mathbb{S}_{coin} = \{Head, Tail\}$: clearly the null event can never happen while the "full" event (either head or tail) should always happen, but this is a matter of associating probabilities to events, not about the definition of events. In the case of grades, one can think about events such as being above (or below) a passing grade such as C, so that $\mathbb{A}_{passing} = \{A, B, C\}$, $\mathbb{A}_{failing} = \{D, E, F\}$, *et cetera*. Similarly, one can split the the income sample space into segments like tax brackets.

Because events are subsets, standard set operations such as **union** ($\cup$), **intersection** ($\cap$) and **complementation** ($\mathbb{A}^c$) extend to them. Moreover, the following properties apply (the proof is left as an exercise).

Theorem 1.1. Properties of Events. *Let $\mathbb{A}_S$, $\mathbb{B}_S$ and $\mathbb{C}_S$ be three events associated with the sample space $\mathbb{S}$. The following properties hold.*

a.	*Commutativity:*	$\mathbb{A}_S \cup \mathbb{B}_S = \mathbb{B}_S \cup \mathbb{A}_S$ $\mathbb{A}_S \cap \mathbb{B}_S = \mathbb{B}_S \cap \mathbb{A}_S$
b.	*Associativity:*	$\mathbb{A}_S \cup (\mathbb{B}_S \cup \mathbb{C}_S) = (\mathbb{A}_S \cup \mathbb{B}_S) \cup \mathbb{C}_S$ $\mathbb{A}_S \cap (\mathbb{B}_S \cap \mathbb{C}_S) = (\mathbb{A}_S \cap \mathbb{B}_S) \cap \mathbb{C}_S$
c.	*Distributive Laws:*	$\mathbb{A}_S \cap (\mathbb{B}_S \cup \mathbb{C}_S) = (\mathbb{A}_S \cap \mathbb{B}_S) \cup (\mathbb{A}_S \cap \mathbb{C}_S)$ $\mathbb{A}_S \cup (\mathbb{B}_S \cap \mathbb{C}_S) = (\mathbb{A}_S \cup \mathbb{B}_S) \cap (\mathbb{A}_S \cup \mathbb{C}_S)$
d.	*DeMorgan's Laws:*	$(\mathbb{A}_S \cup \mathbb{B}_S)^c = \mathbb{A}_S^c \cap \mathbb{B}_S^c$ $(\mathbb{A}_S \cap \mathbb{B}_S)^c = \mathbb{A}_S^c \cup \mathbb{B}_S^c$

It is useful to characterize events that do not overlap, in the sense that no occurrence – that is, no element of the sample space $\mathbb{S}$ – which is contained in one is also contained in the other.

Definition 1.3. Disjoint Events. Two events $\mathbb{A}_1$ and $\mathbb{A}_2$ are **disjoint** or **mutually exclusive** if $\mathbb{A}_1 \cap \mathbb{A}_2 = \varnothing$. The events in a collection $\mathbb{A}_1, \mathbb{A}_2, \ldots$ are **pairwise disjoint** or **mutually exclusive** if $\mathbb{A}_i \cap \mathbb{A}_j = \varnothing$ for all $i \neq j$.

Intuitively, disjoint events cannot simultaneously happen. It is also useful to characterize collections of disjoint events covering the entire sample space.

Definition 1.4. Partition. The events in a collection $\mathbb{A}_1, \mathbb{A}_2, \ldots$ form a **partition** of the sample space $\mathbb{S}$ if they are pairwise disjoint and $\cup_{i=1}^{Z} \mathbb{A}_i = \mathbb{S}$ if the collection is of finite dimension Z; $\cup_{i=1}^{\infty} \mathbb{A}_i = \mathbb{S}$ if the collection has an infinite number of elements.

For example, the collection of events $\mathbb{A}_{passing}$ and $\mathbb{A}_{failing}$ is a partition of the (simplified) letter grades sample space. One can obtain an infinite partition of the income sample space by splitting the positive real line into an infinite number of non-overlapping segments, such as – given some positive integer K – the following sequence.

$$\mathbb{A}_1 = [0, K)\,,\ \mathbb{A}_2 = [K, 2K)\,,\ \ldots,\ \mathbb{A}_n = [(n-1)\,K, nK)\,,\ \ldots$$

These notions are *almost* sufficient to provide a formal characterization of a **probability function**, that is, a function assigning to each event of a sample space a value that measures the chance of any occurrence allowed by that event. The formal mathematical definition of probability functions that is illustrated next follows the **axiomatic foundations** of probability theory as originally developed by Andrej Nikolaevič Kolmogorov. However, it is necessary to first discuss yet another mathematical notion, concerning the properties of the (collection of) events that are the domain of probability functions. This is the concept of **sigma algebra** (σ-**algebra**).

Definition 1.5. Sigma Algebra. Given some set $\mathbb{S}$, a **sigma algebra** or a **Borel field** is a collection of subsets of $\mathbb{S}$, which is denoted as $\mathcal{B}$, that satisfies the following properties:

a. $\varnothing \in \mathcal{B}$;

b. for any subset $\mathbb{A} \in \mathcal{B}$, it is $\mathbb{A}^c \in \mathcal{B}$;

c. for any *countable* sequence of subsets $\mathbb{A}_1, \mathbb{A}_2, \cdots \in \mathcal{B}$, it is $\cup_{i=1}^{\infty} \mathbb{A}_i \in \mathcal{B}$.

It is easy to see that properties **b.** and **c.** together with DeMorgan's Law also imply $\cap_{i=1}^{\infty} \mathbb{A}_i \in \mathcal{B}$ for any appropriate countable sequence of subsets.

The notion of sigma algebra is quite general that it is usually possible to find many sigma algebras for some individual set such as a sample space $\mathbb{S}$. For example, a *trivial* sigma algebra is the one constituted by just the empty set $\varnothing$, and the original set (e.g. $\mathbb{S}$). For finite and (or) countable sets such as the two realizations of the coin experiment and the list of letter grades, *any* collection of subsets is an adequate sigma algebra (it is a good exercise to prove this). For such sets, probability functions are usually formulated upon the largest sigma algebra that contains *all* the subsets.

The case of uncountable sets is somewhat more complicated as certain collections of subsets are not sigma algebras. In most cases, the uncountable set of interest is a subset of $\mathbb{R}^K$ for some given integer $K \geq 1$, and the sigma algebra upon which the probability functions are built is the collection of all *connected* sets, their union and intersections;[1] in the case of $\mathbb{R}$ for example, connected sets take the form

$$[a, b], \quad (a, b], \quad [a, b), \quad (a, b)$$

for any two $a, b \in \mathbb{R}$ with $a \leq b$. It appears then that the notion of sigma algebra is quite general to allow for a wide class of reasonable collections of subsets or events; note though that collections that are not sigma algebras exist and probability functions cannot be applied to them.[2]

The definition of probability function is thus in order.

Definition 1.6. Probability Function. Given a sample space $\mathbb{S}$ and an associated sigma algebra $\mathcal{B}$, a **probability function** $\mathbb{P}$ is a function with domain $\mathcal{B}$ that satisfies the three **axioms of probability**:

a. $\mathbb{P}(\mathbb{A}) \geq 0 \ \forall \mathbb{A} \in \mathcal{B}$;

b. $\mathbb{P}(\mathbb{S}) = 1$;

c. given a *countable* sequence of *pairwise disjoint* subsets $\mathbb{A}_1, \mathbb{A}_2, \cdots \in \mathcal{B}$, then $\mathbb{P}\left(\cup_{i=1}^{\infty} \mathbb{A}_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(\mathbb{A}_i)$.

[1] For readers unfamiliar with topology, a *connected set* is somewhat informally defined as a set that cannot be partitioned into two nonempty subsets such that each subset has no points in common with the set closure of the other subset. For example, the subset of $\mathbb{R}$ defined as $\mathbb{A} = \{x : x \in [a, b) \vee x \in (b, c], a < b < c\}$ is not connected, because it can be partitioned in such a way that defies the above definition.

[2] Here is an example of a collection $\mathcal{B}'$ of subsets of $\mathbb{R}$ which is *not* a sigma algebra. Suppose that $\mathcal{B}'$ contains all the finite disjoint unions of sets of the form

$$(-\infty, a], \quad (a, b], \quad (b, \infty), \quad \varnothing, \quad \mathbb{R}$$

then $\cup_{i=1}^{\infty} \left(0, \frac{i-1}{i}\right] = (0, 1) \notin \mathcal{B}'$ which contradicts the definition of sigma algebra.

Note from the definition that it is especially easy to construct probability functions that satisfy the three axioms for any suitable sigma algebra $\mathcal{B}$ of all finite and/or countable sample spaces $\mathbb{S}$. In such cases, it is sufficient to assign to each element $s \in \mathbb{S}$ a number $p(s) \geq 0$ such that $\sum_{s \in \mathbb{S}} p(s) = 1$.[3] For uncountable sample spaces $\mathbb{S}$, it is best to see a probability function as a particular instance of a **measure** (which is a more general mathematical concept, whose treatment is outside the scope of this chapter).

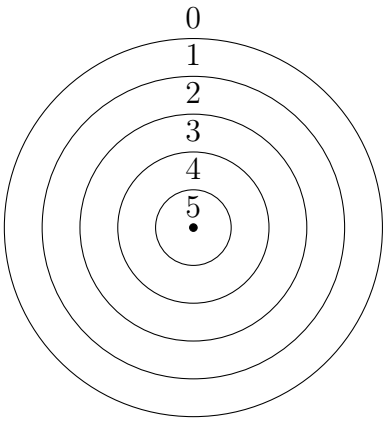


Figure 1.1: Probabilities on a Dartboard

Example 1.1. Probabilities on a Dartboard. One useful example of a probability function for an uncountable sample space is that of the dartboard displayed in Figure 1.1. Consider dart players who score points depending on how closely to the center of the dartboard they land their darts; to each area divided by two contiguous concentric circles corresponds a varying number of points, and zero points are attributed if a player fails to hit the dartboard. One may want to calculate the probability that a player scores a specific number of points by throwing just a single dart. Clearly, the sample space in this particular setting is the set of all the points that the dart can potentially hit (on the dartboard as well as outside it) and it is clearly uncountable, because there are infinitely many such points. Yet it is intuitive to see each region of the dartboard as a separate *event*, that *failure* of hitting the board is another event, that all such events are *pairwise disjoint*, and that together with the empty set they form a *sigma algebra*. Appropriate probability functions, depending say on the skill of each player, can be based on the sample space partition defined by these events.

[3] In the coin example, one has $p(Head) \geq 0$, $p(Tail) \geq 0$ and $p(Head)+p(Tail) = 1$. Similarly, a probability function for the simplified letter grades would assign to each letter a nonnegative number such that their sum equals 1.

It is easier to think about the probability function for a naïve or unskilled player who, if hits the dartboard at all, does so purely at random, so that the probability to score a given number of nonzero points is *proportional* to the area of each dartboard section. Suppose that, for example, the distance of all circles from the center of the dartboard is given by $(I + 1 - i)\, r$, where I is the maximum number of points that are attainable (5 in the Figure), i is the number of points associated with each "ring" of the dartboard, and $r > 0$ is the distance between any two contiguous rings (all equidistant from one another) as well as between the innermost circle and the center. In such a case, the area corresponding to each ring measures as follows.

$$\text{Area associated with } i \text{ points} = \pi r^2 \left[(I + 1 - i)^2 - (I - i)^2\right]$$

It would seem that in order to calculate the desired probabilities, one would need to divide such area by the total area of the dartboard – which equals $\pi r^2 I^2$ – however this is not quite enough, because one must take into account the event that a player fails to heat the board and scores 0 points, and that all the probabilities must sum up to one. Let the area outside the dartboard measure $T > 0$; an appropriate probability function would thus be given by $\mathbb{P}(0 \text{ points}) = T / (T + \pi I^2 r^2)$ and the following expression for $0 < i \leq I$.

$$\mathbb{P}(i \text{ points}) = \pi r^2 \left[(I + 1 - i)^2 - (I - i)^2\right] \left(T + \pi I^2 r^2\right)^{-1}$$

It is easy to verify that the three axioms by Kolmogorov are satisfied. ■

Some general properties of probability functions deserve to be discussed.

Theorem 1.2. Properties of Probability Functions (a). *If $\mathbb{P}$ is some probability function and $\mathbb{A}$ is a set in $\mathcal{B}$, the following properties hold:*

a. $\mathbb{P}(\varnothing) = 0$;

b. $\mathbb{P}(\mathbb{A}) \leq 1$;

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

Proof. The observation that $\mathbb{A}$ and $\mathbb{A}^c$ form a partition of $\mathbb{S}$ and thus it is $\mathbb{P}(\mathbb{A}) + \mathbb{P}(\mathbb{A}^c) = \mathbb{P}(\mathbb{S}) = 1$ proves **c.** – thus **a.** and **b.** follow from it. □

Theorem 1.3. Properties of Probability Functions (b). *If $\mathbb{P}$ is some probability function and $\mathbb{A}$, $\mathbb{B}$ are sets in $\mathcal{B}$, the following properties hold:*

a. $\mathbb{P}(\mathbb{B} \cap \mathbb{A}^c) = \mathbb{P}(\mathbb{B}) - \mathbb{P}(\mathbb{A} \cap \mathbb{B})$;

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

c. *if* $\mathbb{A} \subset \mathbb{B}$, *it is* $\mathbb{P}(\mathbb{A}) \leq \mathbb{P}(\mathbb{B})$.

Proof. To prove **a.** note that $\mathbb{B}$ can be expressed as the union of two disjoint sets $\mathbb{B} = \{\mathbb{B} \cap \mathbb{A}\} \cup \{\mathbb{B} \cap \mathbb{A}^c\}$, thus $\mathbb{P}(\mathbb{B}) = \mathbb{P}(\mathbb{B} \cap \mathbb{A}) + \mathbb{P}(\mathbb{B} \cap \mathbb{A}^c)$. To show **b.** decompose the union of $\mathbb{A}$ and $\mathbb{B}$ as $\mathbb{A} \cup \mathbb{B} = \mathbb{A} \cup \{\mathbb{B} \cap \mathbb{A}^c\}$, again two disjoint sets; hence:

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

where **a.** implies the second equality. Finally, **c.** follows from **a.** as $\mathbb{A} \subset \mathbb{B}$ implies that $\mathbb{P}(\mathbb{A} \cap \mathbb{B}) = \mathbb{P}(\mathbb{A})$, thus $\mathbb{P}(\mathbb{B} \cap \mathbb{A}^c) = \mathbb{P}(\mathbb{B}) - \mathbb{P}(\mathbb{A}) \geq 0$. □

Theorem 1.4. Properties of Probability Functions (c). *If $\mathbb{P}$ is some probability function, the following properties hold:*

- **a.** $\mathbb{P}(\mathbb{A}) = \sum_{i=1}^{\infty} \mathbb{P}(\mathbb{A} \cap \mathbb{C}_i)$ *for any $\mathbb{A} \in \mathcal{B}$ and any partition $\mathbb{C}_1, \mathbb{C}_2, \ldots$ of the sample space such that $\mathbb{C}_i \in \mathcal{B}$ for all $i \in \mathbb{N}$;*
- **b.** $\mathbb{P}(\cup_{i=1}^{\infty} \mathbb{A}_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(\mathbb{A}_i)$ *for any sets $\mathbb{A}_1, \mathbb{A}_2, \ldots$ such that $\mathbb{A}_i \in \mathcal{B}$ for all $i \in \mathbb{N}$.*

Proof. Regarding **a.** note that, by the Distributive Laws of events, it is

$$\mathbb{A} = \mathbb{A} \cap \mathbb{S} = \mathbb{A} \cap \left(\bigcup_{i=1}^{\infty} \mathbb{C}_i \right) = \bigcup_{i=1}^{\infty} (\mathbb{A} \cap \mathbb{C}_i)$$

where the intersection sets of the form $\mathbb{A} \cap \mathbb{C}_i$ are pairwise disjoint as the $\mathbb{C}_i$ sets are, hence:

$$\mathbb{P}(\mathbb{A}) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} (\mathbb{A} \cap \mathbb{C}_i) \right) = \sum_{i=1}^{\infty} \mathbb{P}(\mathbb{A} \cap \mathbb{C}_i)$$

as postulated. To establish **b.** it is useful to construct another collection of *pairwise disjoint* events $\mathbb{A}_1^*, \mathbb{A}_2^*, \ldots$ such that $\cup_{i=1}^{\infty} \mathbb{A}_i = \cup_{i=1}^{\infty} \mathbb{A}_i^*$ and

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} \mathbb{A}_i \right) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} \mathbb{A}_i^* \right) = \sum_{i=1}^{\infty} \mathbb{P}(\mathbb{A}_i^*) \leq \sum_{i=1}^{\infty} \mathbb{P}(\mathbb{A}_i)$$

where the second equality would follow from the pairwise disjoint property. Such additional collection of events can be obtained as:

$$\mathbb{A}_1^* = \mathbb{A}_1, \quad \mathbb{A}_i^* = \mathbb{A}_i \cap \left(\bigcup_{j=1}^{i-1} \mathbb{A}_j \right)^c = \mathbb{A}_i \cap \left(\bigcap_{j=1}^{i-1} \mathbb{A}_j^c \right) \text{ for } i = 2, 3, \ldots$$

which, by construction, are pairwise disjoint and satisfy $\cup_{i=1}^{\infty} \mathbb{A}_i = \cup_{i=1}^{\infty} \mathbb{A}_i^*$. Furthermore, by construction $\mathbb{A}_i^* \subset \mathbb{A}_i$ for every i, implying $\mathbb{P}(\mathbb{A}_i^*) \leq \mathbb{P}(\mathbb{A}_i)$ and thus the inequality $\sum_{i=1}^{\infty} \mathbb{P}(\mathbb{A}_i^*) \leq \sum_{i=1}^{\infty} \mathbb{P}(\mathbb{A}_i)$. □

1.2 Conditional Probability

That of **conditional probability** is a fundamental concept in probability, statistics and econometrics. It is a formalization of the idea of calculating probabilities within a more restricted set of events than the original sample space, and it allows to formulate arguments about probabilities for certain events *given* that some other events have been occurring alongside them.

Definition 1.7. Conditional Probability. Consider a sample space $\mathbb{S}$, an associated sigma algebra $\mathcal{B}$, and two events $\mathbb{A}, \mathbb{B} \in \mathcal{B}$ such that $\mathbb{P}(\mathbb{B}) > 0$. The **conditional probability** of $\mathbb{A}$ **given** $\mathbb{B}$ is written as $\mathbb{P}(\mathbb{A}|\,\mathbb{B})$ and is defined as follows.

$$\mathbb{P}(\mathbb{A}|\,\mathbb{B}) = \frac{\mathbb{P}(\mathbb{A} \cap \mathbb{B})}{\mathbb{P}(\mathbb{B})} \tag{1.1}$$

Example 1.2. Conditional Grades. Naturally, conditional probability is a moot concept for simple, binary scenarios like the coin tossing experiment. It is already a significant concept for the case of grades. Observe that the following two partitions of the grades sample space $\mathbb{S}$ belong to the same sigma algebra: the partition constituted by all singleton grades ($\mathbb{A}_A = \{A\}$, $\mathbb{A}_B = \{B\}$ etc.) and the "passing vs. fail" (where $\mathbb{A}_{passing} = \{A, B, C\}$ and $\mathbb{A}_{failing} = \{D, E, F\}$) partition.[4] Thus it make sense to talk about the probability for a student of scoring a specific grade *given* passing (or failing) an exam. Suppose for example that the probability function for individual grades is as follows:

$$\begin{aligned} \mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) &= 0.3 \\ \mathbb{P}(D) = \mathbb{P}(F) &= 0.05 \\ \mathbb{P}(E) &= 0 \end{aligned}$$

(nobody uses E anymore). Simple additions thus give $\mathbb{P}(passing) = 0.9$ and $\mathbb{P}(failing) = 0.1$. Also note that $A \subset \mathbb{A}_{passing}$, so $\mathbb{P}(A \cap passing) = \mathbb{P}(A)$. Consequently, the probability of getting an A *given* that a student passes the exam can be expressed as the following conditional probability:

$$\mathbb{P}(A|\,passing) = \frac{\mathbb{P}(A \cap passing)}{\mathbb{P}(passing)} = \frac{0.3}{0.9} = \frac{1}{3}$$

and similarly $\mathbb{P}(D|\,failing) = \mathbb{P}(F|\,failing) = 0.5$ express the odds that a student who fails the exams gets either D or F. ■

[4] Consider the sigma algebra for $\mathbb{S}$ based upon the maximal partition including all singleton grade sets; the two "passing" and "failing" sets are obtained by taking appropriate unions of the singleton grade sets, hence must be part of the same sigma algebra.

Example 1.3. Conditional Probabilities on a Dartboard. Endowed with the concept of conditional probability, let us further expand Example 1.1 about dartboard play. Suppose that a player has learned never to miss the dartboard, but still is not skilled enough to effectively target the center, so that his or her darts still hit the dartboard quite randomly (this would be an unusual scenario, but it serves for illustration). The probability that this player scores a given number of points can be obtained from the probability function from Example 1.1, *conditional* upon hitting the board. Remember that failure to hit the dartboard (and scoring zero points) has a probability of $\mathbb{P}(i=0) = T/(T+\pi I^2 r^2)$ for a completely naïve player, meaning that hitting the board has a probability of $\mathbb{P}(i>0) = \pi I^2 r^2/(T+\pi I^2 r^2)$. Thus,

$$\mathbb{P}(i \text{ points}|\, i>0) = \frac{\mathbb{P}(i \cap i>0)}{\mathbb{P}(i>0)} = I^{-2}\left[(I+1-i)^2-(I-i)^2\right]$$

is a conditionally probability function for the general case that can be interpreted as the probability function for the slightly more experienced player. Suppose next that this player trains more and learns how to score at least $3 < I$ points with every dart, but still without becoming more effective at actually approaching the center. Therefore:

$$\mathbb{P}(i \text{ points}|\, i>2) = \frac{\mathbb{P}(i \cap i>2)}{\mathbb{P}(i>2)} = \frac{\left[(I+1-i)^2-(I-i)^2\right]}{(I-2)^2}$$

is the corresponding, more restrictive (conditional) probability function. ■

Example 1.4. Preemptive Medical Treatment. All these examples are relatively trivial, as they are all based on nested sets so that the numerator of (1.1) coincides with either set under consideration. To better appreciate conditional probability, suppose that a population at risk for some kind of illness is offered a preemptive medical treatment that is not always 100% effective. Not all subjects at risk take up the treatment: if some are "takers," others are "hesitant." In both groups, some individuals eventually become sick while others stay healthy. The probability measures associated with all possible intersections between these simple partitions are as follows.

$$\begin{aligned}\mathbb{P}(taker \cap healthy) &= 0.40\\ \mathbb{P}(taker \cap sick) &= 0.20\\ \mathbb{P}(hesitant \cap healthy) &= 0.15\\ \mathbb{P}(hesitant \cap sick) &= 0.25\end{aligned}$$

In terms of the original partitions, $\mathbb{P}(taker) = 0.60$, $\mathbb{P}(hesitant) = 0.40$, $\mathbb{P}(healthy) = 0.55$ and $\mathbb{P}(sick) = 0.45$.

By looking at the elementary probabilities expressed by the intersected sets, it might seem that it is more likely to take the treatment and falling sick rather than not taking it while still falling sick. This conclusion is fallacious though: these concepts must be expressed as conditional probabilities:

$$\mathbb{P}\left(\, sick \middle|\, taker\right) = \frac{\mathbb{P}\left(taker \cap sick\right)}{\mathbb{P}\left(taker\right)} = \frac{1}{3}$$
$$\mathbb{P}\left(\, sick \middle|\, hesitant\right) = \frac{\mathbb{P}\left(hesitant \cap sick\right)}{\mathbb{P}\left(hesitant\right)} = \frac{5}{8}$$

that is, the truth is actually opposite to the original suggestion! ■

An important result based on conditional probabilities is **Bayes' Rule**. Note from inspecting (1.1) that the roles of events $\mathbb{A}$ and $\mathbb{B}$ can be reversed so long as $\mathbb{P}\left(\mathbb{A}\right) > 0$. This observation allows to rewrite the expression in a way that relates the two "reverse" conditional probabilities.

$$\mathbb{P}\left(\, \mathbb{A} \middle|\, \mathbb{B}\right) = \frac{\mathbb{P}\left(\, \mathbb{B} \middle|\, \mathbb{A}\right) \mathbb{P}\left(\mathbb{A}\right)}{\mathbb{P}\left(\mathbb{B}\right)} \tag{1.2}$$

The above expression is a simple version of Bayes' Rule: a powerful result used to calculate conditional probabilities or statistical estimators in a wide variety of settings. The more general version is given as follows.

Theorem 1.5. Bayes' Rule. *Let $\mathbb{A}_1, \mathbb{A}_2, \ldots$ be a partition of the sample space $\mathbb{S}$, and $\mathbb{B}$ some event $\mathbb{B} \subset \mathbb{S}$. For $i = 1, 2, \ldots$ the following holds.*

$$\mathbb{P}\left(\, \mathbb{A}_i \middle|\, \mathbb{B}\right) = \frac{\mathbb{P}\left(\, \mathbb{B} \middle|\, \mathbb{A}_i\right) \mathbb{P}\left(\mathbb{A}_i\right)}{\sum_{j=1}^{\infty} \mathbb{P}\left(\, \mathbb{B} \middle|\, \mathbb{A}_j\right) \mathbb{P}\left(\mathbb{A}_j\right)}$$

Proof. This follows from (1.2) for $\mathbb{A} = \mathbb{A}_i$ and by observing that:

$$\mathbb{P}\left(\mathbb{B}\right) = \sum_{j=1}^{\infty} \mathbb{P}\left(\mathbb{B} \cap \mathbb{A}_j\right) = \sum_{j=1}^{\infty} \mathbb{P}\left(\, \mathbb{B} \middle|\, \mathbb{A}_j\right) \mathbb{P}\left(\mathbb{A}_j\right)$$

from Theorem 1.4 and the definition of conditional probability. □

Example 1.5. Imperfect Medical Treatment, continued. Let us continue with example 1.4, and suppose that one is interested in the conditional probability of finding a taker among the sick, knowing the conditional probability of getting sick after taking the treatment, the total probability measure of takers, and the total probability measure corresponding with getting sick. This is easily calculated from (1.2).

$$\mathbb{P}\left(\, taker \middle|\, sick\right) = \mathbb{P}\left(\, sick \middle|\, taker\right) \frac{\mathbb{P}\left(taker\right)}{\mathbb{P}\left(sick\right)} = \frac{4}{9}$$

Note: this number says little about the effectiveness of the treatment! ■

On occasion, conditional probabilities do not differ from unconditional probabilities, e.g. for two events $\mathbb{A}$ and $\mathbb{B}$, it is $\mathbb{P}(\mathbb{A}|\,\mathbb{B}) = \mathbb{P}(\mathbb{A})$. Intuitively, this is so because the two events are completely unrelated, and knowing that one of the two (say $\mathbb{B}$) occurs does not change the odd of the other event ($\mathbb{A}$) to happen. In such cases, it is said that the two events are **independent**.

Definition 1.8. Statistical independence (*two events*). Two events $\mathbb{A}$ and $\mathbb{B}$ are **statistically independent** if the following holds.

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

This intuition is easily extended to groups (collections) of events.

Definition 1.9. Mutual statistical independence (*multiple events*). The events of any collection $\mathbb{A}_1, \mathbb{A}_2, \ldots, \mathbb{A}_N$ are **mutually independent** if, for any subcollection $\mathbb{A}_{i_1}, \mathbb{A}_{i_2}, \ldots, \mathbb{A}_{i_{N'}}$ with $N' \leq N$, the following holds.

$$\mathbb{P}\left(\bigcap_{j=1}^{N'} \mathbb{A}_{i_j}\right) = \prod_{j=1}^{N'} \mathbb{P}\left(\mathbb{A}_{i_j}\right)$$

The notion of independence is also extended to the complements of the events involved.

Theorem 1.6. Independence and Complementary Events. *Consider any two independent events $\mathbb{A}$ and $\mathbb{B}$. It can be concluded that the following pairs of events are independent too:*

a. *$\mathbb{A}$ and $\mathbb{B}^c$;*

b. *$\mathbb{A}^c$ and $\mathbb{B}$;*

c. *$\mathbb{A}^c$ and $\mathbb{B}^c$.*

Proof. Case **a.** is easily shown as follows:

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

where the second equality follows from the definition of independence. Cases **b.** and **c.** are analogous. □

As it is elaborated in later chapters, the primitive concept of statistical independence and the associated mathematical definitions extend one-to-one to probability distributions of random variables; as such they are crucial for characterizing the properties of statistical samples and estimators.

1.3 Probability Distributions

The general concept of probability function applies to any suitable set which can be expressed as a sample space and endowed with a sigma algebra. In statistics and econometrics, however, it is often convenient to re-formulate probability functions so that the domain of interest are real numbers. This requires handling two fundamental probabilistic concepts: that of **random variables** and that of **probability distribution functions**.

Definition 1.10. Random Variables. A **random variable** X is a function from the the sample space $\mathbb{S}$ onto the set of real numbers $X : \mathbb{S} \to \mathbb{R}$.

Conventionally, random variables are denoted by slanted capital letters, while their **realizations**, corresponding to specific outcomes in the original sample space(s) that are hypothesized to occur, are denoted by italic lower case letters. For example, the random variable X_{coin} for the coin experiment is expressed as:

$$x_{coin} = \begin{cases} 1 & \text{if } Tail \\ 0 & \text{if } Head \end{cases}$$

or vice versa. In more extended experiments where *many* (e.g. $n \geq 1$) coins are tossed, the outcome of interest is the total count $X_{n.coins} \in \{0, 1, \ldots, n\}$ of heads (or tails): $X_{n.coins}$ is another random variable. A random variable for letter grades – X_{grade} – can correspond to each grade's weight for calculating the GPA, such as $x_{grade} = 4$ for A, $x_{grade} = 3$ for B, *et cetera*. In the case of sample spaces like the number of received emails, individual income and individual wealth, the corresponding random variables X typically map onto the original sample spaces: $\mathbb{N}_0$, $\mathbb{R}_+$ and $\mathbb{R}$ respectively. However, the mapping itself may be non-trivial for reasons of interpretation.[5]

Definition 1.11. Probability Distribution. Given a random variable X, a **cumulative (probability) distribution function** (often abbreviated as **c.d.f.**) is a function $F_X(x)$ which is defined as follows.

$$F_X(x) = \mathbb{P}(X \leq x) \quad \text{for all } x \in \mathbb{R}$$

A cumulative probability distribution function is the mathematical object that allows to reformulate a primitive probability function, as expressed in a sample space, into a function which takes real numbers as arguments, yet conveys the same information about probability as the original function $\mathbb{P}$. The subscript X associates a c.d.f. to a specific random variable X, and is often omitted when discussing generic distributions or their properties.

[5]For example, one may want to convert heterogeneous monetary values expressed in different currencies into standardized monetary units.

Example 1.6. Tossing Two Coins. Consider the "extended" coin experiment for $n = 2$. The sample space for this scenario is the set

$$\mathbb{S}_{2.coins} = \{Head\,\&\,Head,\ Head\,\&\,Tail,\ Tail\,\&\,Head,\ Tail\,\&\,Tail\}$$

where for each element, the two terms before and after the '&' sign represent the outcome for the first and second coin respectively. The random variable of interest takes values in the set $X_{2.coins} \in \{0, 1, 2\} \in \mathbb{R}$, and the mapping X equals the number of tails in each attempt.

$$\begin{aligned} X\,(Head\,\&\,Head) &= 0 \\ X\,(Head\,\&\,Tail) &= 1 \\ X\,(Tail\,\&\,Head) &= 1 \\ X\,(Tail\,\&\,Tail) &= 2 \end{aligned}$$

Assuming that the coins in the experiment are "balanced" (that is, there are equal chances to obtain heads or tails), and given that clearly the outcome of either coin cannot be predicted by the other (if treated as separate events, they would be *independent*), the probability associated with each element of $\mathbb{S}_{2.coins}$ is 0.25, meaning that $\mathbb{P}\,(X_{2.coins} = 0) = \mathbb{P}\,(X_{2.coins} = 2) = 0.25$ while $\mathbb{P}\,(X_{2.coins} = 1) = 0.50$. This results in the following cumulative probability distribution:

$$F_{X_{2.coins}}\,(x) = \begin{cases} 0 & \text{if } x \in (-\infty, 0) \\ .25 & \text{if } x \in [0, 1) \\ .75 & \text{if } x \in [1, 2) \\ 1 & \text{if } x \in [2, \infty) \end{cases}$$

which is easily represented graphically as in in Figure 1.2 below.

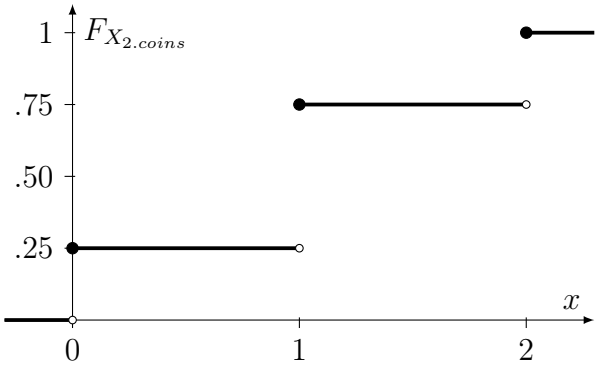


Figure 1.2: Probability Distribution for the Two Coins Experiment

Figure 1.2 displays a step function which is non-decreasing, with image in $[0, 1]$, and which is right-continuous (but left-discontinuous) at every step. As it is discussed next, this property is not unique to this example. ■

Theorem 1.7. Properties of Probability Distribution Functions. *A function $F(x)$ can be a probability distribution function if and only if the following three conditions hold:*

a. $\lim_{x\to-\infty} F(x) = 0$ *and* $\lim_{x\to\infty} F(x) = 1$*;*

b. $F(x)$ *is a nondecreasing function of* x*;*

c. $F(x)$ *is right-continuous, that is* $\lim_{x\downarrow x_0} F(x) = F(x_0)\ \ \forall x_0 \in \mathbb{R}$.

Proof. (Outline.) Necessity follows easily from the definition of Probability Functions. Sufficiency requires some reverse engineering, that would show how for each Probability Distribution Function with the above properties, one can find an appropriate sample space $\mathbb{S}$, an associated probability function $\mathbb{P}$ and a relative random variable X. □

While all probability distributions must conform to the conditions established by Theorem 1.7, not all of them take the shape of step functions. In fact, the latter is true only for a certain class of distributions: those for *discrete* – as opposed to *continuous* – random variables.

Definition 1.12. Classes of Random Variables. A random variable X is **continuous** if $F_X(x)$ is a continuous function of x, while it is **discrete** if $F_X(x)$ is a step function of x.

The following example provides a discussion of the two most exemplifying continuous probability distributions.

Example 1.7. Standard Logistic and Normal Distributions. While perhaps not the most frequently found, the **standard logistic** probability distribution

$$F_X(x) = \Lambda(x) = \frac{1}{1+\exp(-x)}$$

is arguably the continuous probability distribution taking positive values on the entire real line with the simplest mathematical expression. It is depicted in Figure 1.3 (continuous line) and it is easy to verify that it satisfies the conditions of Theorem 1.7 (note that its derivative is always positive). The most important continuous distribution, however, is certainly the **standard normal** or **Gaussian distribution**, a more complex function:

$$F_X(x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$$

while also taking positive values on the entire real line and complying with Theorem 1.7. Even this distribution is displayed in Figure 1.3 (dashed line); observe how relative to the standard logistic, the standard normal implies a higher probability associated to the values of x closer to zero.

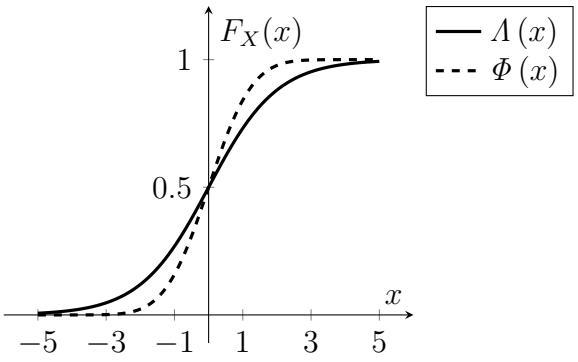


Figure 1.3: Standard Logistic and Normal Probability Distributions

As elaborated later in this chapter, the standard logistic and the standard normal are both specialized cases of more flexible specifications of the logistic and normal distributions, which allow for **parameters** that determine their exact shape (however, the particular notation $\Lambda(x)$ and $\Phi(x)$ is typically reserved for the "standard" versions of both distributions). Both distributions are often used to represent real world scenarios that are best represented on the entire real line, like the deviations of a certain variable of interest (say, human height) from some focal point (say, a group-specific average). The predominance of the standard normal is motivated by a fundamental result in asymptotic probability theory, the Central Limit Theorem, which is discussed at length in Lecture 6. ■

Cumulative probability distributions are seldom handled directly; it is usually more convenient to manipulate some associated mathematical objects that more directly relate to the underlying probability measures. Such objects, the probability **mass** and **density** functions, are defined differently for discrete and continuous distributions, respectively. These two concepts make it easier to also characterize the **support** of a random variable, which intuitively is the subset of $\mathbb{R}$ where all the probability is concentrated.

Definition 1.13. Probability Mass Function. Given a *discrete* random variable X, its probability **mass** function $f_X(x)$ (which is often abbreviated as **p.m.f.**) is defined as follows.

$$f_X(x) = \mathbb{P}(X = x) \quad \text{for all } x \in \mathbb{R}$$

Definition 1.14. Probability Density Function. Given a *continuous* random variable X, its probability **density** function $f_X(x)$ (which is often abbreviated as **p.d.f.**) is defined as the function that satisfies the following relationship.

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt \quad \text{for all } x \in \mathbb{R}$$

If a continuous cumulative distribution is differentiable everywhere in $\mathbb{R}$, the associated density function is simply its first derivative: $f_X(t) = \frac{\partial F_X(t)}{\partial x}$.

Definition 1.15. Support of a random variable. Given some random variable X which is either discrete or continuous, its **support** $\mathbb{X}$ is defined as the following set

$$\mathbb{X} \equiv \{x : x \in \mathbb{R}, f_X(x) > 0\}$$

where $f_X(x)$ is the probability mass *or* density function associated with X, as appropriate.

Clearly, the support of a discrete random variable is a countable set, thus a probability mass function has an easy interpretation as a transposition of the underlying probability function, implying for instance that:

$$\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a) = \sum_{t=a}^{b} f_X(t)$$

hence:

$$\mathbb{P}(X \leq b) = F_X(b) = \sum_{t=\inf \mathbb{X}}^{b} f_X(t)$$

and:

$$\mathbb{P}(X \in \mathbb{X}) = \sum_{t \in \mathbb{X}} f_X(t) = 1$$

which connects directly with the cumulative probability distribution $F_X(x)$.

Example 1.8. Tossing Two Coins, Revisited. Consider the cumulative distribution function for the experiment about "tossing two coins" described in Example 1.6. The associated probability mass function is obtained from the original probability function:

$$f_{X_{2.coins}}(x) = \begin{cases} .25 & \text{if } x = 0 \\ .50 & \text{if } x = 1 \\ .25 & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases}$$

and it is visually represented in Figure 1.4, as displayed next. ■

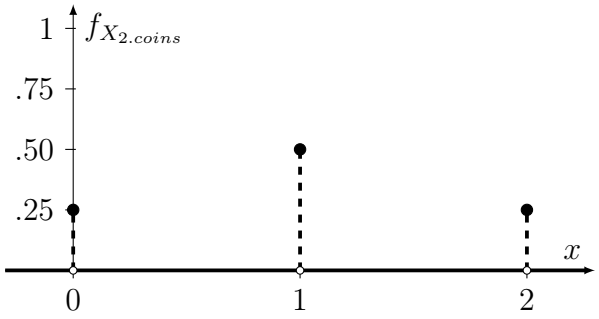


Figure 1.4: Probability Mass Function for the Two Coins Experiment

For density functions instead the support is an uncountable set, and the interpretation of the quantity $f_X(x) \geq 0$ is subtler: it cannot be interpreted as a probability because x has measure zero in the support. However, when X is continuous the definition of cumulative distribution functions implies that:

$$\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(t)\, dt$$

hence:

$$\mathbb{P}(X \in \mathbb{X}) = \int_{\mathbb{X}} f_X(t)\, dt = 1$$

hence density functions bear a probabilistic interpretation for *segments* of $\mathbb{R}$. Also observe that unlike in the case of mass functions, density functions can generally take values larger than one, since their probabilistic interpretation is based on the above integral formulations.

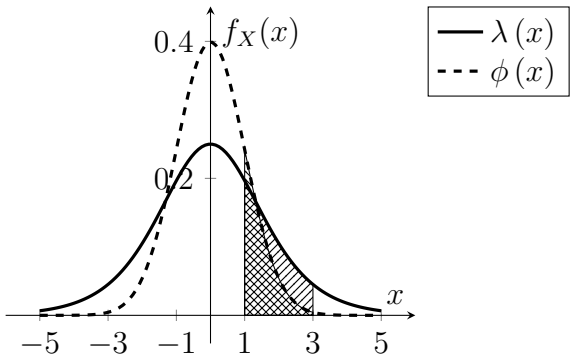
Example 1.9. Standard Logistic and Normal Density Functions. Let us expand Example 1.7. The density function associated with the standard logistic distribution $\Lambda(x)$ is:

$$\lambda(x) = \frac{d\Lambda(x)}{dx} = \frac{\exp(-x)}{[1 + \exp(-x)]^2}$$

while the density function of the standard normal distribution is as follows.

$$\phi(x) = \frac{d\Phi(x)}{dx} = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

Both functions are displayed in Figure 1.5 below. The graphical comparison between the two density functions highlights again that the standard logistic is "thicker in the tails" relative to the standard normal, that is, the standard



Note: the shaded areas represent the probability that x falls in the $[1, 3]$ interval for either distribution

Figure 1.5: Standard Logistic and Normal Probability Densities

logistic probability is more dispersed towards the outer values of the support $\mathbb{R}$. To better exemplify the probabilistic interpretation of density functions, the Figure also displays – by means of distinct shaded areas – the probability that x occurs between 1 and 3 for either distribution. ■

One can summarize the properties of mass and density function through the following statement.

Theorem 1.8. Properties of mass and density functions. *A function $f_X(X)$ is an appropriate probability mass or density function of a random variable X if and only if:*

a. $f_X(X) \geq 0$ *for all* $x \in \mathbb{R}$*;*

b. $\sum_{x \in \mathbb{X}} f_X(x) = 1$ *or* $\int_{\mathbb{X}} f_X(x)\, dx = 1$ *for mass and density functions respectively.*

Proof. (*Outline.*) Necessity follows by the definitions of cumulative distribution, mass and density functions. Sufficiency follows by Theorem 1.7 after having constructed the associated cumulative distribution $F_X(X)$. □

It must be noted at this point that not all random variables are either exclusively discrete, or exclusively continuous. In numerous situations of interest, a random variable appears continuous only on a subset of the support and discrete in other points. In such cases, the definition of cumulative probability distribution is still valid, however those of mass and density functions are only valid upon a subset of the support. It is possible to describe these mixed cases by using a generalized density which is formulated in terms of a Lebesgue integral, but this is beyond the scope of this treatment.

Example 1.10. Truncated Standard Normal Distribution. Suppose that a real world phenomenon is distributed according to a standard normal distribution, but is not actually observed for negative values. An example is that of an electronic detector of potential power overload, which would measure any positive deviation from some optimal "average" (which is established at zero) but would not detect negative deviations, which are recorded simply as $x = 0$. In such a case, one typically says that the distribution in question (here, the standard normal) is **truncated** at zero. The cumulative distribution function would read here as:

$$\Phi_{\geq 0}(x) = \begin{cases} 0 & \text{if } x < 0 \\ \Phi(x) & \text{if } x \geq 0 \end{cases}$$

and would be drawn as in Figure 1.6 below.

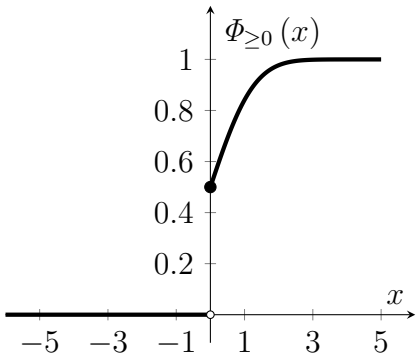


Figure 1.6: Cumulative standard normal distribution truncated at zero

In this case, it is sensible to characterize the density function only for the nonnegative part of the distribution's support:

$$\phi_{\geq 0}(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \text{ if } x \geq 0$$

which allows to calculate the probabilities about observations falling within specified intervals in $\mathbb{R}_+$ (observe that $\int_0^\infty \phi(x)\, dx = 0.5$). The description of this particular truncated distribution is completed by specifying that the rest of the probability *mass* is found at zero:

$$\mathbb{P}(X = 0) = 0.5$$

since obviously $\mathbb{P}(X < 0) = 0$. ■

1.4 Relating Distributions

Many random variables are related to one another, in the sense that they convey similar and sometimes identical probabilistic information for certain events. It is useful to characterize when two random variables are **identical**, and when they can be expressed as a **transformation** from one to another.

Definition 1.16. Identically Distributed Random Variables. Any two random variables X and Y sharing a sample space $\mathbb{S}$ and an associated sigma algebra $\mathcal{B}$ are said to be **identically distributed** if for every event $\mathbb{A} \in \mathcal{B}$, it is $\mathbb{P}\left(X \in X\left(\mathbb{A}\right)\right) = \mathbb{P}\left(Y \in Y\left(\mathbb{A}\right)\right)$.

The definition is quite straightforward: two identically distributed random variables express the same information about the probability of fundamental events in the sample space. Observe that, however, two identically distributed random variables need not be equal mappings. Suppose that, for example, one defines X as the count of "Heads" and Y as the count of "Tails" in the simple coin experiment (even if repeated n times). Clearly X and Y are identically distributed even if $X\left(\mathbb{A}\right) \neq Y\left(\mathbb{A}\right)$ for all $\mathbb{A} \in \mathcal{B}$. If the domain of the original mapping is a subset of $\mathbb{R}$, the following holds.

Theorem 1.9. Identical Distribution. *Given two random variables X and Y whose primitive sample space is a subset of the real numbers $\mathbb{S} \subseteq \mathbb{R}$, the following two statements are identical:*

a. *X and Y are identically distributed;*

b. *$F_X\left(x\right) = F_Y\left(x\right)$ for every x in the relevant support.*

Proof. (Outline.) Clearly **a.** implies **b.** by construction. The converse – that **b.** implies **a.** – requires showing that if the two distributions are identical, they share a probability function defined for some sigma algebra $\mathcal{B}$ of $\mathbb{S}$. □

Two random variables also convey similar information if they are related through some functional dependence. Suppose that given random variable X, another random variable Y is defined as $Y = g\left(X\right)$, where $g\left(\cdot\right)$ is some function – one would commonly say that Y is a *transformation* of X. Thus, for any two real numbers $a \leq b$, it is:

$$\mathbb{P}\left(Y \in [a, b]\right) = \mathbb{P}\left(g\left(X\right) \in [a, b]\right) \tag{1.3}$$

and similarly for intervals of the form $[a, b)$, $(a, b]$ and (a, b). The probabilistic relationship between X and Y (and their distributions) becomes more apparent when the function g is invertible on the interval of interest, hence:

$$\mathbb{P}\left(Y \in [a, b]\right) = \mathbb{P}\left(X \in g^{-1}\left([a, b]\right)\right) \tag{1.4}$$

where $g^{-1}([a,b]) \in \mathbb{R}$ is the subset of real numbers that are mapped by the inverse function $g^{-1}(\cdot)$.[6] Also note that in general, a transformed random variable Y has a a support $\mathbb{Y}$ which differs from the support $\mathbb{X}$ of the original random variable X; an obvious example is $Y = \exp(X)$ whereby if $\mathbb{X} = \mathbb{R}$, it is $\mathbb{Y} = \mathbb{R}_{++}$; conversely if $Y = \log(X)$ and $\mathbb{X} = \mathbb{R}_{++}$, it is $\mathbb{Y} = \mathbb{R}$.

A relevant question is about how to calculate the distribution and the mass or density functions of Y starting from those of X. If X is discrete also Y is, and the calculation of mass functions is straightforward.

$$f_Y(y) = f_X\left(g^{-1}(y)\right) \tag{1.5}$$

Thus, the cumulative distribution for Y can be obtained by summing all the mass points for preceding values in the support.

$$F_Y(y) = \sum_{\inf \mathbb{Y}}^{y} f_Y(y)$$

For continuous random variables, things are slightly more complicated. Let us start from the following result about cumulative distributions.

Theorem 1.10. Cumulative Distribution of Transformed Random Variables. *Let X and $Y = g(X)$ be two random variables that are related by a transformation $g(\cdot)$, $\mathbb{X}$ and $\mathbb{Y}$ their respective supports, and $F_X(x)$ the cumulative distribution of X.*

- **a.** *If $g(\cdot)$ is increasing in $\mathbb{X}$, it is $F_Y(y) = F_X(g^{-1}(y))$ for all $y \in \mathbb{Y}$.*
- **b.** *If $g(\cdot)$ is decreasing in $\mathbb{X}$ and X is a continuous random variable, it is $F_Y(y) = 1 - F_X(g^{-1}(y))$ for all $y \in \mathbb{Y}$.*

Proof. This is almost tautological: **a.** is shown as:

$$F_Y(y) = \int_{-\infty}^{g^{-1}(y)} f_X(x)\,dx = F_X\left(g^{-1}(y)\right)$$

where the first equality is motivated on (1.4) and the fact that an increasing function applied upon some interval preserves its order. The demonstration of **b.** is symmetric:

$$F_Y(y) = \int_{g^{-1}(y)}^{\infty} f_X(x)\,dx = 1 - F_X\left(g^{-1}(y)\right)$$

since a decreasing function upon an interval inverts the order and because $\int_{-\infty}^{a} f_X(x)\,dx + \int_{a}^{\infty} f_X(x)\,dx = 1$ if $f_X(x)$ is a density function. □

[6] Note that this subset may not equal $\left[g^{-1}(a), g^{-1}(b)\right]$ because the inverse mapping $g^{-1}(\cdot)$ may not preserve the order or the connectedness of the original interval.

To calculate the transformed density, another theorem – building on the previous one – comes to rescue.

Theorem 1.11. Density of Transformed Random Variables (1). *Let X and $Y = g(X)$ be two random variables related by a transformation $g(\cdot)$, $\mathbb{X}$ and $\mathbb{Y}$ their respective supports, and $f_X(x)$ the probability density function of X, which is continuous on $\mathbb{X}$. If the inverse of the transformation function, $g^{-1}(\cdot)$, is continuously differentiable on $\mathbb{Y}$, the probability density function of Y can be calculated as follows.*

$$f_Y(y) = \begin{cases} f_X\left(g^{-1}(y)\right) \left|\frac{d}{dy} g^{-1}(y)\right| & \text{if } y \in \mathbb{Y} \\ 0 & \text{if } y \notin \mathbb{Y} \end{cases}$$

Proof. Both increasing and decreasing functions are monotone; hence, since $g^{-1}(\cdot)$ is continuously differentiable on $\mathbb{Y}$, for all $y \in \mathbb{Y}$:

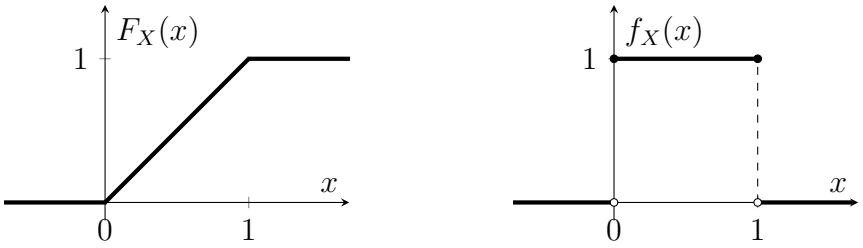
$$f_Y(y) = \frac{d}{dy} F_Y(y) = \begin{cases} f_X\left(g^{-1}(y)\right) \frac{d}{dy} g^{-1}(y) & \text{if } g(\cdot) \text{ is increasing} \\ -f_X\left(g^{-1}(y)\right) \frac{d}{dy} g^{-1}(y) & \text{if } g(\cdot) \text{ is decreasing} \end{cases}$$

by Theorem 1.10 and the chain rule. □

Example 1.11. Uniform to Exponential Transformation. Let X be a random variable with a **uniform distribution** on the **unit interval**: a random variable with support $\mathbb{X} = [0, 1]$, cumulative distribution

$$F_X(x) = \begin{cases} 0 & \text{if } x \in (-\infty, 0] \\ x & \text{if } x \in (0, 1) \\ 1 & \text{if } x \in [1, \infty) \end{cases}$$

and density function $f_X(x) = \mathbb{1}[x \in [0, 1]]$, as depicted in Figure 1.7.



Note: cumulative distribution function $F_X(x)$ on the left, density function $f_X(x)$ on the right

Figure 1.7: Uniform distribution on the unit interval

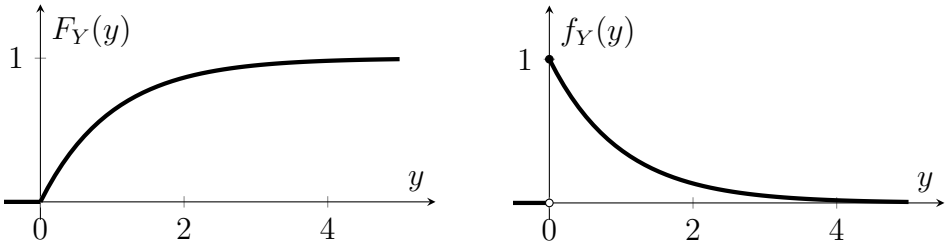
Now consider the transformed random variable $Y = -\log X$, which is obtained by applying a decreasing monotone function to X. Observe how the support of Y is the set of all nonnegative real numbers $\mathbb{Y} = \mathbb{R}_+$, and that the transformation can be inverted: $X = \exp(-Y)$. By applying Theorem 1.10, the cumulative distribution of the transformed random variable is:

$$F_Y(y) = 1 - \exp(-y)$$

while its density function is:

$$f_Y(y) = \exp(-y)$$

as it is easily verified through Theorem 1.11.



Note: cumulative distribution function $F_Y(y)$ on the left, density function $f_Y(y)$ on the right

Figure 1.8: Exponential distribution with unit parameter

As it shall be expanded later, the transformed random variable Y follows a particular case of the **exponential distribution**, that with unit parameter. Figure 1.8 shows its cumulative distribution and density function. ■

Theorem 1.11 is restricted to monotone transformations. However, it can be extended to a more general class of transformations as follows.

Theorem 1.12. Density of Transformed Random Variables (2). *Let X and $Y = g(X)$ be two random variables related by some transformation $g(\cdot)$, $\mathbb{X}$ and $\mathbb{Y}$ their respective supports, and $f_X(x)$ the probability density function of X. Suppose further that there exists a partition of the support of X, $\mathbb{X}_0, \mathbb{X}_1, \ldots, \mathbb{X}_K$ such that $\cup_{i=0}^{k} \mathbb{X}_i = \mathbb{X}$, $\mathbb{P}(x \in \mathbb{X}_0) = 0$, and $f_X(x)$ is continuous on each $\mathbb{X}_i$. Finally, suppose that there is a sequence of functions $g_1(x), \ldots, g_k(x)$, each associated with a corresponding set in $\mathbb{X}_1, \ldots, \mathbb{X}_K$, satisfying the following conditions for $i = 1, \ldots, K$:*

i. *$g(x) = g_i(x)$ for every $x \in \mathbb{X}_i$;*

ii. *$g_i(x)$ is monotone in $\mathbb{X}_i$;*

iii. $\mathbb{Y} = \{y : y = g_i(x) \text{ for some } x \in \mathbb{X}_i\}$, *that is the image of* $g_i(x)$ *is always equal to the support of* Y*;*

iv. $g_i^{-1}(y)$ *exists and is continuously differentiable in* $\mathbb{Y}$.

Then the density of Y *can be calculated as follows.*

$$f_Y(y) = \begin{cases} \sum_{i=1}^{K} f_X\left(g_i^{-1}(y)\right) \left|\frac{d}{dy} g_i^{-1}(y)\right| & \text{if } y \in \mathbb{Y} \\ 0 & \text{if } y \notin \mathbb{Y} \end{cases}$$

Proof. (Outline.) The logic of this result is that if $g(\cdot)$ is not monotone, but it can be separated into a sequence of monotone subfunctions over different intervals of the support of $\mathbb{X}$, then the result of Theorem 1.11 can be applied to each interval, and thus the density for each point $y \in \mathbb{Y}$ can be obtained as the sum of the transformed densities associated with all points in $x \in \mathbb{X}$ that map to y (note that this allows $g(\cdot)$ not to be invertible over the entire support of X, it suffices that it is invertible on each interval in the partition). The "dummy" set $\mathbb{X}_0$ with zero probability allows for discontinuity or even saddle points separating the K subfunctions in the partition. □

Example 1.12. Normal to Chi-squared Transformation. Let X be a random variable that follows the standard normal distribution $\Phi(x)$; the support of X is thus $\mathbb{X} = \mathbb{R}$. Consider the transformation $Y = X^2$: function $g(x) = x^2$ is obviously not monotone over all $\mathbb{R}$. However, it is respectively decreasing in $\mathbb{R}_-$ and increasing in $\mathbb{R}_+$; and it is easy to verify that it satisfies the requirements of Theorem 1.12 for the following sets.

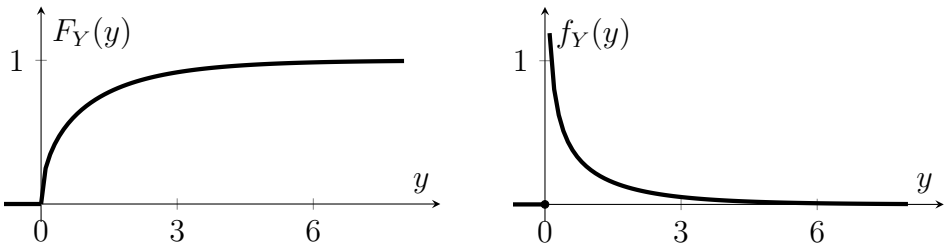
$$\begin{aligned} \mathbb{X}_0 &= \{0\} \\ \mathbb{X}_1 &= \mathbb{R}_{--} \\ \mathbb{X}_2 &= \mathbb{R}_{++} \end{aligned}$$

Therefore, the density of Y is obtained, for $y \in \mathbb{Y} = \mathbb{R}_{++}$, as:

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(-\sqrt{y}\right)^2}{2}\right) \left|-\frac{1}{2\sqrt{y}}\right| + \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(\sqrt{y}\right)^2}{2}\right) \left|\frac{1}{2\sqrt{y}}\right| \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} \exp\left(-\frac{y}{2}\right) \end{aligned}$$

as $g_1^{-1}(y) = -\sqrt{y}$ and $g_2^{-1}(y) = \sqrt{y}$. Its cumulative distribution is obtained by integrating the density.

$$F_Y(y) = \int_0^y \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{t}} \exp\left(-\frac{t}{2}\right) dt$$



Note: cumulative distribution function $F_Y(y)$ on the left, density function $f_Y(y)$ on the right.

Figure 1.9: Chi-squared distribution (χ^2) with one degree of freedom

This turns out to be a particular case of the **chi-squared** (χ^2) distribution, that with one "degree of freedom" – its cumulative distribution and density function are displayed in Figure 1.9. The general version of the chi-squared distribution, its relationship with other distributions besides the standard normal, and its role in statistical inference are discussed later at length. ■

It is worth to conclude this general discussion of probability distributions by introducing the concept of a random variable's **quantile function**.

Definition 1.17. Quantile Function. The *quantile* function associated with a random variable X is the following function with argument $p \in (0, 1)$.

$$Q_X(p) = \inf\{x \in \mathbb{X} : p \leq F_X(x)\}$$

Observe that $Q_X(p)$ corresponds to the inverse of $F_X(x)$ if the latter is strictly increasing; otherwise – if $F_X(x)$ is flat on segments of the support of X – the quantile function returns a "pseudo-inverse" with the property

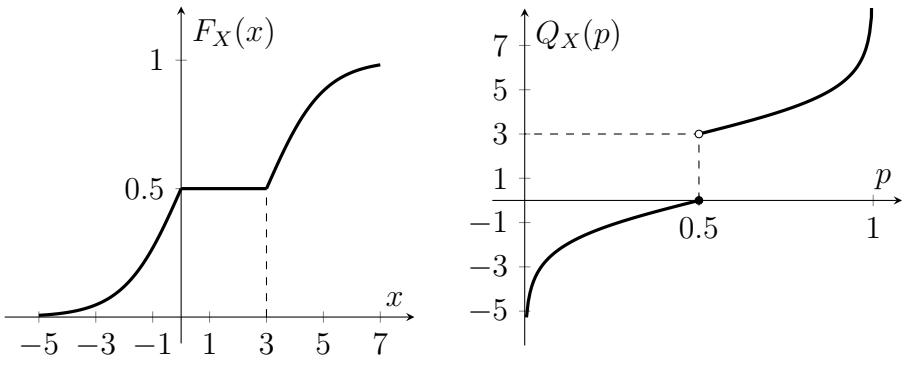
$$\mathbb{P}(Q_X[F_X(X)] \leq Q_X(p)) = \mathbb{P}(X \leq Q_X(p))$$

for all $p \in (0, 1)$, by construction. This is best illustrated via an example.

Example 1.13. The quantile function of a non-strictly-monotonic distribution. Consider a random variable X with the following cumulative distribution function.

$$F_X(x) = \begin{cases} [1 + \exp(-x)]^{-1} & \text{if } x \in (-\infty, 0] \\ 0.5 & \text{if } x \in (0, 3) \\ [1 + \exp(-x + 3)]^{-1} & \text{if } x \in [3, \infty) \end{cases}$$

The shape of this distribution closely mimics that of the standard logistic from Example 1.7, but with a crucial difference that makes it "non-strictly" monotonic: over the $(0, 3)$ interval this distribution is flat, and it resumes a "standard logistic" behavior only for $x \geq 3$, as if the support is shifted by three units of measurement associated with the random variable X.



Note: cumulative distribution function $F_X(X)$ on the left, quantile function $Q_X(p)$ on the right.

Figure 1.10: Quantile function for a non-strictly-monotonic distribution

According to the definition, the quantile function of X is as follows.

$$Q_X(p) = \begin{cases} \log(x) - \log(1-x) & \text{if } x \in (0, 0.5] \\ \log(x) - \log(1-x) + 3 & \text{if } x \in (0.5, 1) \end{cases}$$

This "pseudo-inverse" is visibly discontinuous at $p = 0.5$, which is the value that the cumulative distribution takes for $x \in [0, 3]$ – an interval whose infimum is the value that the quantile function itself takes at the discontinuity, $Q_X(0.5) = 0$. The cumulative distribution as well as the quantile function of X are both displayed in Figure 1.10. ■

Quantile functions have several applications; a typical one is grounded on the result that is expressed and demonstrated next.

Theorem 1.13. Cumulative Transformation. *For any continuous random variable X with cumulative distribution denoted as $F_X(x)$, the transformation $P = F_X(X)$ follows a uniform distribution on the unit interval.*

Proof. By the properties of quantile functions (including the fact that they are monotone increasing by definition), for all $p \in (0, 1)$ it holds that:

$$\begin{aligned} \mathbb{P}(P \leq p) &= \mathbb{P}(F_X(X) \leq p) \\ &= \mathbb{P}(Q_X[F_X(X)] \leq Q_X(p)) \\ &= \mathbb{P}(X \leq Q_X(p)) \\ &= F_X(Q_X(p)) \\ &= p \end{aligned}$$

where the fourth and fifth lines follow from the definition and continuity of $F_X(x)$. Since $F_P(p) = 0$ for $p \leq 0$ and $F_P(p) = 1$ for $p \geq 0$ by construction, P follows a uniform distribution on the interval $(0, 1)$. □

This theorem motivates the use of the uniform distribution for generating random draws from any distribution $F_X(x)$. Given that it is easier to obtain actual random draws from the uniform distribution, it is convenient to do so and then apply the quantile function $Q_X(p)$ in order to obtain the desired random draws from $F_X(x)$, whatever the random variable X of interest.

1.5 Moments of Distributions

The **moments** of a probability distribution are quantities that summarize some of its properties. Moments can be either **uncentered** or **centered**.

Definition 1.18. Uncentered Moments. The r-th uncentered moment of a random variable X with support $\mathbb{X}$, denoted as $\mathbb{E}[X^r]$, is defined for some positive integer r and for discrete random variables as

$$\mathbb{E}[X^r] = \sum_{x \in \mathbb{X}} x^r f_X(x)$$

while it is defined as follows in the case of continuous random variables.

$$\mathbb{E}[X^r] = \int_{\mathbb{X}} x^r f_X(x)\, dx$$

The zero-th centered moment $\mathbb{E}[X^0]$ is defined as the unweighted sum of the probability mass function or the unweighted integral of the probability density function over the support of X, and is thus by definition $\mathbb{E}[X^0] = 1$.

The most important uncentered moment is that for $r = 1$, $\mathbb{E}[X]$; it is called **mean** or **expected value** (or even **expectation**). The mean is a measure of the "central" value of the distribution of X in a probabilistic sense, and it is obtained by summing or integrating all the elements of the support, weighted by their probability mass or density. The mean is instrumental in the definition of the centered moments.

Definition 1.19. Centered Moments. The r-th centered moment of a random variable X with support $\mathbb{X}$, denoted as $\mathbb{E}[(X - \mathbb{E}[X])^r]$, is defined for some positive integer r and for discrete random variables as:

$$\mathbb{E}[(X - \mathbb{E}[X])^r] = \sum_{x \in \mathbb{X}} (x - \mathbb{E}[X])^r f_X(x)$$

while it is defined as follows in the case of continuous random variables.

$$\mathbb{E}[(X - \mathbb{E}[X])^r] = \int_{\mathbb{X}} (x - \mathbb{E}[X])^r f_X(x)\, \mathrm{d}x$$

The most important centered moment is that for $r = 2$, the **variance**: it is always nonnegative and is denoted as follows.

$$\mathbb{V}\mathrm{ar}\left[X\right] = \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right] \geq 0 \tag{1.6}$$

The variance provides a measure of the "dispersion" of the distribution of X around the mean.[7] Higher centered moments provide the basis for other relevant measures. For example, the centered moment for $r = 3$, when it is standardized by the variance up to the power of 1.5, delivers the so-called **skewness**:

$$\mathbb{S}\mathrm{kew}\left[X\right] = \frac{\mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^3\right]}{\left(\mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right]\right)^{\frac{3}{2}}} \gtreqless 0$$

a measure of the degree of *asymmetry* of a distribution. The skewness is a positive number for asymmetric distributions that are "skewed" to the right of the mean – and vice versa; it is equal to zero only for distributions that are symmetric around the mean. The centered moment for $r = 4$ instead is functional for calculating the **kurtosis**, which is defined through another standardization – now, by the square of the variance:

$$\mathbb{K}\mathrm{urt}\left[X\right] = \frac{\mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^4\right]}{\left(\mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right]\right)^2} \geq 0$$

and it is once again an always nonnegative number; indeed a measure of the overall "thickness" of the distribution – the relative frequency of realizations of X that are distant from the mean. Centered moments can be conveniently expressed as functions of uncentered moments only. In the variance case, for example:

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[X\right] &= \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right] \\ &= \mathbb{E}\left[X^2\right] - 2\,\mathbb{E}\left[X\right]\mathbb{E}\left[X\right] + \mathbb{E}\left[X\right]^2 \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2\end{aligned}$$

while in general – for higher centered moments – a variation of the binomial formula applies.

$$\mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^r\right] = \sum_{i=0}^{r} \binom{r}{i} \left(-1\right)^{r-i} \mathbb{E}\left[X^i\right] \mathbb{E}\left[X\right]^{r-i}$$

This is a useful fact since, as it is discussed later, the uncentered moments of a distribution can be calculated through the associated *moment generating* or *characteristic* functions.

[7] Clearly, the variance is equal to zero only in the case of degenerate discrete distributions where the the entire probability mass is concentrated in one single point.

Example 1.14. Mean and variance for coin experiments. Consider the usual simple example about tossing a coin whereby $x_{coin} = 1$ if *Head* and $x_{coin} = 0$ if *Tail*. Suppose however that the coin is "unfair:" in particular, *Head* occurs with probability 0.6 and Tail with probability 0.4. Thus:

$$\begin{aligned}\mathbb{E}\left[X_{coin}\right] &= 1 \cdot f_{X_{coin}}\left(1\right) + 0 \cdot f_{X_{coin}}\left(0\right) \\ &= 1 \cdot 0.6 + 0 \cdot 0.4 \\ &= 0.6\end{aligned}$$

whereas the variance can be calculated as follows.

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[X_{coin}\right] &= \left(1 - \mathbb{E}\left[X_{coin}\right]\right)^2 \cdot f_{X_{coin}}\left(1\right) + \left(0 - \mathbb{E}\left[X_{coin}\right]\right)^2 \cdot f_{X_{coin}}\left(0\right) \\ &= \left(0.4\right)^2 \cdot 0.6 + \left(-0.6\right)^2 \cdot 0.4 \\ &= 0.24\end{aligned}$$

Now consider the more complex case of the random variable $X_{n.coins}$ which counts the heads (or tails) out of n iterations of the simple coin experiment. In total, there are 2^n possible sequences of outcomes; however, it is quite easy to enumerate those corresponding to exactly x heads (or tails) by using the binomial coefficient $\binom{n}{x}$. Because every single *sequence* counting exactly x heads (or tails) occurs with probability $0.6^x \cdot 0.4^{n-x}$, the probability mass function for this specific random variable can be written as:

$$f_{X_{n.coins}}\left(x\right) = \binom{n}{x} \cdot 0.6^x \cdot 0.4^{n-x}$$

and its expected value can be calculated as follows, for $y = x - 1$:

$$\begin{aligned}\mathbb{E}\left[X_{n.coins}\right] &= \sum_{x=0}^{n} x \binom{n}{x} \cdot 0.6^x \cdot 0.4^{n-x} \\ &= \sum_{x=1}^{n} n \binom{n-1}{x-1} \cdot 0.6^x \cdot 0.4^{n-x} \\ &= 0.6 \cdot n \sum_{x=1}^{n} \binom{n-1}{x-1} \cdot 0.6^{x-1} \cdot 0.4^{n-1-x+1} \\ &= 0.6 \cdot n \underbrace{\sum_{y=0}^{n} \binom{n-1}{y} \cdot 0.6^{y} \cdot 0.4^{n-1-y}}_{=1} \\ &= 0.6 \cdot n\end{aligned}$$

where the simplification in the second-to-last line occurs because the summation therein is recognized as the total probability mass of an analogous, hypothetical experiment with $n - 1$ attempts and y successes.

The second uncentered moment is calculated similarly.

$$\begin{aligned}
\mathbb{E}\left[X^2_{n.coins}\right] &= \sum_{x=0}^{n} x^2 \binom{n}{x} \cdot 0.6^x \cdot 0.4^{n-x} \\
&= \sum_{x=1}^{n} xn \binom{n-1}{x-1} \cdot 0.6^x \cdot 0.4^{n-x} \\
&= n \sum_{y=0}^{n} (y+1) \binom{n-1}{y} \cdot 0.6^{y+1} \cdot 0.4^{n-y-1} \\
&= 0.6 \cdot n \sum_{y=0}^{n} y \binom{n-1}{y} \cdot 0.6^y \cdot 0.4^{n-y} + \\
&\quad + 0.6 \cdot n \sum_{y=0}^{n} \binom{n-1}{y} \cdot 0.6^y \cdot 0.4^{n-y} \\
&= 0.6 \cdot n \cdot [0.6 \cdot (n-1) + 1]
\end{aligned}$$

Here, the two summations in the second-to-last line also simplify, since they correspond respectively to the mean – equaling $0.6 \cdot (n-1)$ – and the total probability mass – equaling 1 – of the hypothetical experiment discussed earlier. Exploiting the fact that $\mathbb{V}\text{ar}\left[X_{n.coins}\right] = \mathbb{E}\left[X^2_{n.coins}\right] - \mathbb{E}\left[X_{n.coins}\right]^2$, it is easy to verify that the variance of $X^2_{n.coins}$ equals $0.24 \cdot n$. ■

Example 1.15. Mean and variance of the uniform distribution. Let again some continuous random variable X follow the uniform distribution on the $(0, 1)$ interval. Its mean is given simply by:

$$\mathbb{E}\left[X\right] = \int_0^1 x dx = \left.\frac{x^2}{2}\right|_0^1 = \frac{1}{2}$$

whereas, since:

$$\mathbb{E}\left[X^2\right] = \int_0^1 x^2 dx = \left.\frac{x^3}{3}\right|_0^1 = \frac{1}{3}$$

the variance is calculated as $\mathbb{V}\text{ar}\left[X\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$. ■

Example 1.16. Mean and variance of the exponential distribution. Consider some random variable Y that follows the exponential distribution with unit parameter, which is connected with the uniform distribution on the $[0, 1]$ interval through the transformation $Y = -\log X$ as per Example 1.11. Its mean is calculated through integration by parts:

$$\mathbb{E}\left[Y\right] = \int_0^\infty y \exp\left(-y\right) dy = \left. -y \exp\left(-y\right)\right|_0^\infty + \int_0^\infty \exp\left(-y\right) dy = 1$$

since $\lim_{M\to\infty} -y\exp(-y)\big|_0^M = 0$ and $\int_0^\infty \exp(-y)\,dy = 1$. The variance is calculated by noting that, integrating by parts again:

$$\mathbb{E}\left[Y^2\right] = \int_0^\infty y^2\exp(-y)\,dy = -y^2\exp(-y)\big|_0^\infty + 2\int_0^\infty y\exp(-y)\,dy = 2$$

hence $\mathbb{V}\mathrm{ar}[Y] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = 2 - 1 = 1$. ■

One is often interested in the analysis of the moments of a transformed random variable $Y = g(X)$ in terms of the moments of the original random variable X. By applying the standard linear properties of summations and integration, it is quite easy to see that if $Y = a + bX$, then:

$$\mathbb{E}[Y] = a + b\,\mathbb{E}[X]$$

and, since $\mathbb{E}\left[(Y - \mathbb{E}[Y])^2\right] = \mathbb{E}\left[b^2(X - \mathbb{E}[X])^2\right]$, the following holds too.

$$\mathbb{V}\mathrm{ar}[Y] = b^2\,\mathbb{V}\mathrm{ar}[X]$$

For non-linear functions $g(X)$, **Jensen's Inequality** can be extended to probability distributions to show that, if $g(\cdot)$ is a concave function,

$$\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$$

while if $g(\cdot)$ is a convex function instead, the converse applies.

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$$

This shows that, in general, a non-linear function does not pass through the expectation operator. In addition, a first-order approximation of $g(X)$ based on a Taylor expansion around $\mathbb{E}[X]$:

$$g(X) \approx g(\mathbb{E}[X]) + g'(\mathbb{E}[X])[X - \mathbb{E}[X]]$$

shows that $\mathbb{E}[g(X)] \approx g(\mathbb{E}[X])$ is hardly an acceptable approximation. However, one can rewrite the expansion as:

$$g(X) \approx [g(\mathbb{E}[X]) - g'(\mathbb{E}[X])\,\mathbb{E}[X]] + g'(\mathbb{E}[X])\,X$$

showing that the approximation:

$$\mathbb{V}\mathrm{ar}[g(X)] \approx [g'(\mathbb{E}[X])]^2\,\mathbb{V}\mathrm{ar}[X]$$

is actually a decent one, as it accounts for the first order term of the series.

The next two properties of the mean and the variance are instrumental in establishing some important results of asymptotic theory.

Theorem 1.14. Markov's Inequality. *Given a nonnegative random variable $X \in \mathbb{R}_+$ and a constant $k > 0$, it must be $\mathbb{P}[X \geq k] \leq \mathbb{E}[X]/k$.*

Proof. Apply the decomposition

$$\mathbb{E}[X] = \int_0^{+\infty} xf(x)\,dx \geq \int_k^{+\infty} xf(x)\,dx \geq k\int_k^{+\infty} f(x)\,dx = k\,\mathbb{P}[X \geq k]$$

with the first equality requiring X to be nonnegative. □

Theorem 1.15. Čebyšëv's Inequality. *Given a random variable $Y \in \mathbb{R}$ and a number $\delta > 0$, it must be $\mathbb{P}[|Y - \mathbb{E}[Y]| \geq \delta] \leq \mathbb{V}\mathrm{ar}[Y]/\delta^2$.*

Proof. Rephrase Markov's inequality setting $X = (Y - \mathbb{E}[Y])^2$ and $k = \delta^2$, and notice that:

$$\mathbb{P}[|Y - \mathbb{E}[Y]| \geq \delta] \leq \mathbb{P}\left[(Y - \mathbb{E}[Y])^2 \geq \delta^2\right] \leq \frac{\mathbb{E}\left[(Y - \mathbb{E}[Y])^2\right]}{\delta^2} = \frac{\mathbb{V}\mathrm{ar}[Y]}{\delta^2}$$

as postulated. □

An additional important relationship about the mean and the variance, which helps build intuition about these two moments and interpreting them, relates to the property of the mean as being the "best predictor" of a random variable X under a quadratic criterion of distance. More concretely, suppose that one is aiming to *guess* an unknown realization of X, whose distribution is known, by solving the following problem:

$$\min_{\widehat{X}} \mathbb{E}\left[\left(X - \widehat{X}\right)^2\right]$$

which is intuitively appealing, since unexpected deviations that are larger in magnitude, when squared, count more towards the evaluation of the above expectation. It is easy to see that the solution is found at $\widehat{X} = \mathbb{E}[X]$:

$$\begin{aligned}
\mathbb{E}\left[\left(X - \widehat{X}\right)^2\right] &= \mathbb{E}\left[\left(X - \mathbb{E}[X] + \mathbb{E}[X] - \widehat{X}\right)^2\right] \\
&= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] + \mathbb{E}\left[\left(\mathbb{E}[X] - \widehat{X}\right)^2\right] \\
&\quad + 2\underbrace{\mathbb{E}\left[(X - \mathbb{E}[X])\left(\mathbb{E}[X] - \widehat{X}\right)\right]}_{=0} \\
&= \mathbb{V}\mathrm{ar}[X] + \mathbb{E}\left[\left(\mathbb{E}[X] - \widehat{X}\right)^2\right]
\end{aligned} \tag{1.7}$$

where the third term in the second line is zero because:

$$\mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)\left(\mathbb{E}\left[X\right] - \widehat{X}\right)\right] = \left(\mathbb{E}\left[X\right] - \widehat{X}\right)\mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)\right] = 0$$

since neither $\mathbb{E}\left[X\right]$ nor $\widehat{X}$ are random and can be taken out of the expectation operator, while $\mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)\right] = 0$ by definition. Of the two remaining terms in the last line of (1.7), the first one – the variance of X – is constant, while the second is shrunk to zero when $\widehat{X} = \mathbb{E}\left[X\right]$. Thus, in addition to the interpretation of the mean as "best predictor" under quadratic distances, the variance is intuitively interpreted as the prediction error that "cannot be removed." Later in Lecture 7 this property of the mean and variance is generalized in a setting where multiple random variables are used to jointly predict the realization of some other random variable of interest.

The last concept covered in this lecture is about classes of functions that are most useful to calculate the moments of a distribution.

Definition 1.20. Moment generating function. Given a random variable X with support $\mathbb{X}$, the **moment-generating** function $M_X\left(t\right)$ is defined, for $t \in \mathbb{R}$, as the expectation of the transformation $g\left(X\right) = \exp\left(tX\right)$, so long as it exists; for discrete random variables this is:

$$M_X\left(t\right) = \mathbb{E}\left[\exp\left(tX\right)\right] = \sum_{x \in \mathbb{X}} \exp\left(tx\right) f_X\left(x\right)$$

while for continuous random variables, it is as follows.

$$M_X\left(t\right) = \mathbb{E}\left[\exp\left(tX\right)\right] = \int_{\mathbb{X}} \exp\left(tx\right) f_X\left(x\right) dx$$

Moment generating functions draw their name from the following result.

Theorem 1.16. Moment generation. *If a random variable X has an associated moment generating function $M_X\left(t\right)$, its r-th uncentered moment can be calculated as the r-th derivative of the moment generating function evaluated at $t = 0$.*

$$\mathbb{E}\left[X^r\right] = \left.\frac{d^r M_X\left(t\right)}{dt^r}\right|_{t=0}$$

Proof. Note that, for all $r = 1, 2, \ldots$:

$$\frac{d^r M_X\left(t\right)}{dt^r} = \frac{d^r}{dt^r}\,\mathbb{E}\left[\exp\left(tX\right)\right] = \mathbb{E}\left[\frac{d^r}{dt^r}\exp\left(tX\right)\right] = \mathbb{E}\left[X^r \exp\left(tX\right)\right]$$

so long as the r-th derivative with respect to t can pass through the expectation operator. If so, it is $\mathbb{E}\left[X^r \exp\left(tX\right)\right] = \mathbb{E}\left[X^r\right]$ for $t = 0$. □

Obviously, this allows to obtain centered moments as well, by the earlier observation that all centered moments can be expressed as functions of the uncentered ones. It is a useful exercise to calculate both the mean and the variance of the distributions from Examples 1.14, 1.15 and 1.16 by using the respective moment generating functions, which are provided in the following additional list of examples.

Example 1.17. Moment generating function for coin experiments. Let us return to the coin experiments from example 1.14. In the one attempt case, the moment generating function is simple to calculate.

$$\begin{aligned} M_{X_{coin}}(t) &= \exp(t \cdot 1) \cdot f_{X_{coin}}(1) + \exp(t \cdot 0) \cdot f_{X_{coin}}(0) \\ &= 0.6 \cdot \exp(t) + 0.4 \end{aligned}$$

With n attempts instead:

$$\begin{aligned} M_{X_{n.coins}}(t) &= \sum_{x=0}^{n} \binom{n}{x} \exp(tx) \cdot 0.6^x \cdot 0.4^{n-x} \\ &= \sum_{x=0}^{n} \binom{n}{x} [0.6 \cdot \exp(t)]^x \cdot 0.4^{n-x} \\ &= [0.6 \cdot \exp(t) + 0.4]^n \end{aligned}$$

the result is obtained with another application of the binomial formula. ■

Example 1.18. Moment generating function of the uniform distribution. The moment generating function of a random variable X which is uniformly distributed on the unit segment is simple to calculate.

$$M_X(t) = \int_0^1 \exp(tx)\, dx = \frac{1}{t} \exp(tx) \Big|_0^1 = \frac{1}{t} [\exp(t) - 1]$$

Note: it might seem that this function is ill-defined at $t = 0$, but applying the Taylor expansion $\exp(t) = \sum_{n=0}^{\infty} t^n/n!$ for $t = 0$ into the above formula would reveal that actually, $M_X(0) = 1$. ■

Example 1.19. Moment generating function of the exponential distribution. Calculations are not that more difficult in the case of a random variable Y following the exponential distribution with unit parameter, but it must be noted that *it only exists for* $t < 1$.

$$M_Y(t) = \int_0^{\infty} \exp((t-1)\, y)\, dy = \lim_{M \to \infty} -\frac{1}{1-t} \exp(-(1-t)\, y) \Big|_0^M = \frac{1}{1-t}$$

In fact, it is easy to see that the above integral diverges if $t \geq 1$. ■

The moment generating functions of two random variables Y and X that are related by a linear transformation, say $Y = a + bX$, are also themselves related through a simple formula. In fact:

$$M_Y(t) = \exp(at) M_X(bt)$$

where $M_X(t)$ and $M_Y(t)$ are the moment generating functions of X and Y respectively. This is fact is easily shown as follows.

$$\mathbb{E}[\exp(tY)] = \mathbb{E}[\exp(ta + tbX)] = \exp(at)\,\mathbb{E}[\exp(btX)]$$

A fundamental property of moment generating functions is that they **uniquely characterize a distribution**, in the sense that each distribution $F_X(x)$ has its own distinct moment generating function $M_X(t)$. The proof of this result requires more involved mathematics and is not developed here. Observe that in general a unique *sequence of moments* does not identify a unique distribution, or vice versa. It can be shown that this is only the case for a specific subset of distributions: those with *bounded support*. In other words, there exist different distributions, whose support is unbounded, that have different moment generating functions but share identical moments.

As hinted, sometimes a moment-generating function does not exist, or it is not defined within any open interval around $t = 0$. In such a case, the alternative **characteristic function** $\varphi_X(t)$ is guaranteed to always exist, and to be unique for each distribution.

Definition 1.21. Characteristic function. For a given random variable X with support $\mathbb{X}$, the characteristic function $\varphi_X(t)$ is defined, for $t \in \mathbb{R}$ and for discrete random variables:

$$\varphi_X(t) = \mathbb{E}[\exp(itX)] = \sum_{x \in \mathbb{X}} \exp(itx) f_X(x)$$

while for continuous random variables it is:

$$\varphi_X(t) = \mathbb{E}[\exp(itX)] = \int_{\mathbb{X}} \exp(itx) f_X(x)\, dx$$

where i is the imaginary unit.

A result analogous to Theorem 1.16 shows that it is possible to calculate all the moments of a distribution by using the characteristic functions instead.

$$\mathbb{E}[X^r] = \frac{1}{i^r} \cdot \left. \frac{d^r \varphi_X(t)}{dt^r} \right|_{t=0} \quad \text{for } r \in \mathbb{N}$$

As it involves complex numbers, the characteristic function is a much more difficult object to handle than the moment-generating function. The characteristic function is most useful to prove fundamental results of asymptotic probability theory, such as the Central Limit Theorem.

Lecture 2

Common Distributions

This lecture is essentially an annotated list of probability distributions that are most frequently encountered in practice. The list is organized by type of distribution (e.g. discrete vs. continuous) and especially in the case of the continuous ones, by "family" – a term that denotes groups of distributions sharing similar characteristics. The objective of this lecture is not simply to familiarize with important and frequently found distributions, but also to highlight those relationships between distributions that are especially useful towards statistical inference and econometric analysis.

2.1 Discrete Distributions

Some examples of discrete probability distributions are developed in Lecture 1 starting from the simple coin experiment. In the first part of this section these examples are more rigorously generalized, allowing for a *parametrization* of the hypothetical experiment. Subsequently, the discussion of other discrete distributions follows. Starting from this section, some distributions are associated with a specific notation, so that it is easier to interpret – for example – the following conventional expression:

$$X \sim \text{Be}(p)$$

as "the random variable X is distributed according to the Bernoulli distribution with parameter p." While allowing for parameters, it can be useful to indicate them in the expression of mass or density functions, cumulative distributions, *et cetera*. Therefore, a conventional expression such as:

$$f_X(x; p) = px + (1 - p)(1 - x)$$

means that X follows the above probability mass function *given* a parameter p. Note how a semicolon separates realizations and parameters in $f_X(x; p)$.

Bernoulli distribution

The Bernoulli distribution is the one describing dichotomous events akin to those of the coin experiment. In general, one must allow for the two events under consideration to occur with different probabilities (for example, coins may not be "fair" or "balanced"). One writes $X \sim \text{Be}(p)$ if $\mathbb{X} = \{0, 1\}$ and:

$$\begin{aligned} \mathbb{P}(X = 1) &= p \\ \mathbb{P}(X = 0) &= 1 - p \end{aligned}$$

implying a probability mass function that can be written as in the above example about notation, or equivalently – but more elegantly – as:

$$f_X(x; p) = p^x (1 - p)^{1-x} \tag{2.1}$$

for $x \in \{0, 1\}$ and $p \in [0, 1]$. The cumulative distribution writes:

$$F_X(x; p) = (1 - p) \cdot \mathbb{1}[x \in [0, 1)] + \mathbb{1}[x \in [1, \infty)] \tag{2.2}$$

its moment generating function is:

$$M_X(t; p) = p \exp(t) + (1 - p) \tag{2.3}$$

and this lets obtain the mean and the variance easily as follows.

$$\mathbb{E}[X] = p \tag{2.4}$$

$$\mathbb{V}\text{ar}[X] = p(1 - p) \tag{2.5}$$

The Bernoulli distribution is elementary; thus, it forms the basis for several other discrete distributions.

Binomial distribution

The binomial distribution characterizes a random variable defined on a sample space constituted by all possible recombinations of n Bernoulli (binary) events with probability p, and that measures the probability for the number x and $n - x$ of realizations of each alternative. Thus, this distribution corresponds to the hypothetical experiment from Lecture 1 about tossing several, (say n) possibly unbalanced coins. Conventionally, the outcomes counted as x are defined as "successes" and those counted as $n - x$ as "failures;" for this reason, it is common to verbally describe the binomial distribution as the one that measures the "probability of x successes of a binary phenomenon out of n attempts." In less verbal terms, a random variable X that follows the binomial distribution is typically denoted as follows.

$$X \sim \text{Bn}(p, n)$$

The above expression highlights the two parameters that characterize the binomial distribution: the probability of a single trial $p \in [0,1]$ and the number of trials $n \in \mathbb{N}$. It is helpful to appreciate the following facts about random variables X that follow a binomial distribution:

- a possible reformulation of the sample space is the set $\mathbb{S} = \{0,1\}^n$;
- the support of the random variable is $\mathbb{X} = \{0, 1, \ldots, n\}$;
- if the probability of all Bernoulli realizations is p, all the events that define the binomial distribution are mutually independent.

The binomial distribution owes its name to the fact that its probability mass function is described through the formula:

$$\mathbb{P}(X = x; p, n) = f_X(x; p, n) = \binom{n}{x} p^x (1-p)^{n-x} \tag{2.6}$$

as argued earlier in the particular context of Example 1.14, this is due to the fact that the count of possible outcomes featuring exactly x successes is given by the binomial coefficient: $\binom{n}{x}$. The cumulative probability function writes, for $x \in [0, n]$, and given $\lfloor x \rfloor$ the largest integer smaller than x, as:

$$\mathbb{P}(X \leq x; p, n) = F_X(x; p, n) = \sum_{i=0}^{\lfloor x \rfloor} \binom{n}{i} p^i (1-p)^{n-i} \tag{2.7}$$

with $\mathbb{P}(X = n; p, n) = F_X(n; p, n) = 1$ per the binomial formula.

$$\begin{aligned} F_X(n; p, n) &= \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \\ &= (p + (1-p))^n \\ &= 1 \end{aligned}$$

The moment generating function is similarly obtained (see Example 1.17):

$$\begin{aligned} M_X(t; p, n) &= \sum_{x=0}^{n} \binom{n}{x} \exp(tx)\, p^x (1-p)^{n-x} \\ &= [p \exp(t) + (1-p)]^n \end{aligned} \tag{2.8}$$

while the mean and variance are as follows (one can calculate them through the approach in Example 1.14, or with the moment generation function).

$$\mathbb{E}[X] = np \tag{2.9}$$

$$\mathbb{V}\mathrm{ar}[X] = np(1-p) \tag{2.10}$$

The two distributions discussed next are variations on the idea of multiple Bernoulli events or, as these are usually referred to, "Bernoulli trials."

Geometric distribution

Consider the sample space constructed out of all combinations of *an infinite number* of Bernoulli trials with identical probability p. Suppose that these trials are ordered; for example, the order might correspond to a sequence in time when the trials are realized. Rather than defining a random variable X that counts the number of successes, let $X \in \mathbb{N}$ denote the index of the first Bernoulli trial in the sequence for which a "success" is observed. It is:

$$\mathbb{P}(X = x; p) = f_X(x; p) = p(1-p)^{x-1} \tag{2.11}$$

because for a success with probability p to happen in the x-th trial, $x-1$ failures must first occur, an event with probability $(1-p)^{x-1}$. The probability mass function in (2.11) is characterized by a *geometric* series, which motivates the name for the distribution associated with X. By the properties of the geometric series, for $x \in \mathbb{R}$ the cumulative distribution function of X is obtained as:

$$\begin{aligned} \mathbb{P}(X \leq x; p) = F_X(x; p) &= \sum_{i=0}^{\lfloor x \rfloor - 1} p(1-p)^i \\ &= \frac{1-(1-p)^{\lfloor x \rfloor}}{1-(1-p)} p \\ &= 1-(1-p)^{\lfloor x \rfloor} \end{aligned} \tag{2.12}$$

which converges to 1 as $x \to \infty$; while it is $F_X(x; p) = 0$ for $x < 1$. Similarly, the moment generating function is obtained, for $t < -\log(1-p)$, as:

$$\begin{aligned} M_X(t; p) &= \lim_{M \to \infty} \sum_{x=0}^{M} \exp(tx) \cdot p(1-p)^{x-1} \\ &= p \exp(t) \cdot \lim_{M \to \infty} \sum_{x=0}^{M} [(1-p) \cdot \exp(t)]^{x-1} \\ &= p \exp(t) \cdot \lim_{M \to \infty} \frac{1-[(1-p) \cdot \exp(t)]^M}{1-(1-p) \cdot \exp(t)} \\ &= \frac{p \exp(t)}{1-(1-p)\exp(t)} \end{aligned} \tag{2.13}$$

allowing to derive the mean and variance following tedious calculations.

$$\mathbb{E}[X] = \frac{1}{p} \tag{2.14}$$

$$\mathbb{V}\mathrm{ar}[X] = \frac{1-p}{p^2} \tag{2.15}$$

A very important feature of the geometric distribution is its *memoryless* property, which is defined as follows for any two integers s, t with $s > t$:

$$\mathbb{P}(X > s | X > t) = \mathbb{P}(X > s - t) \tag{2.16}$$

that is, the probability that success occurs at the s-th trial conditional on t trials having already occurred (implicitly, with failure) is equal to the *ex ante* probability that success occurs with exactly $s-t$ trials. In other words – keeping with the time interpretation of the sequence of Bernoulli trials – every failure is uninformative about future odds of success (or failure): it is as if the calculation of future odds "forgets" past realizations. This property of the geometric distribution is easily proved as follows.

$$\begin{aligned}
\mathbb{P}(X > s | X > t) &= \frac{\mathbb{P}(X > s \cap X > t)}{\mathbb{P}(X > t)} \\
&= \frac{\mathbb{P}(X > s)}{\mathbb{P}(X > t)} \\
&= (1-p)^{s-t} \\
&= \mathbb{P}(X > s - t)
\end{aligned}$$

Because of the memoryless property, the geometric distribution is used to model the "waiting count" of homogeneous Bernoulli trials that occur before some events of interest, on the assumption that the passing of time does not affect their probabilities. It is *not* an adequate distribution to describe, say, the probability that some physical objects stop to function or some living beings die if it is likely that some aging process – of either physical objects or living beings – may affect these probabilities.[1]

Negative binomial distribution

Consider the same setup as that of the geometric distribution: all the possible combinations of an infinite number of ordered Bernoulli trials. However, let the outcome of interest be not the number of trials it takes to achieve *one* success, but instead the number of trials $X \in \mathbb{N}$ it takes to get a *generic number* of successes $r \in \mathbb{N}$. This probability is defined as:

$$\mathbb{P}(X = x; p, r) = f_X(x; p, r) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \tag{2.17}$$

which is the probability mass function of the *negative* binomial distribution; one would typically express this as follows.

$$X \sim \mathrm{NB}(p, r)$$

[1] In these cases, the expression "success" of a Bernoulli trial is certainly a misnomer.

The motivation for (2.17) is readily explained: every unique combination of x Bernoulli trials featuring r successes and that *terminates with a success* in the x-th trial must contain exactly $r-1$ successes in the previous $x-1$ trials; the number of those combinations is counted through the binomial coefficient $\binom{x-1}{r-1}$, and each of them occurs with probability $p^r (1-p)^{x-r}$.

Observation 2.1. The geometric distribution is a special case of the negative binomial distribution, with $r = 1$; thus it is denoted as $X \sim \mathrm{NB}(p, 1)$.

The negative binomial distribution can also be expressed by the alternative random variable $Y = X - r$ which counts the number of *failures* that occur before the r-th success; by analogous reasoning:

$$\mathbb{P}(Y = y; p, r) = f_Y(y; p, r) = \binom{r+y-1}{y} p^r (1-p)^y \tag{2.18}$$

which, by the properties of the binomial coefficients with negative integers, can also be written as:

$$\mathbb{P}(Y = y; p, r) = f_Y(y; p, r) = (-1)^y \binom{-r}{y} p^r (1-p)^y$$

which motivates the term *negative binomial*, thanks to the resemblance with (2.6) if not for the (possibly) negative multiplicative term. The cumulative distribution is obtained by appropriately summing over (2.17) – or (2.18) – and it can be shown that, when its argument x (or y) goes to infinity, it has limit one. The moment generating function is, for $t < -\log(1-p)$:

$$M_X(t; p, r) = \left(\frac{p \exp(t)}{1 - (1-p)\exp(t)} \right)^r \tag{2.19}$$

while the key moments of X are as follows (those of Y are derived easily).

$$\mathbb{E}[X] = \frac{r}{p} \tag{2.20}$$

$$\mathbb{V}\mathrm{ar}[X] = \frac{r(1-p)}{p^2} \tag{2.21}$$

The negative binomial distribution is frequently used in both statistics and econometrics in order to model countable events of interest.

Poisson distribution

The Poisson distribution, which is presented next, is an important discrete distribution with numerous applications. Like the other distributions presented thus far, also the Poisson is related to the concept of Bernoulli trials, although the connection is less immediate to intuitively appreciate. To that end, it is helpful to provide first a formal description of the distribution.

The support of the Poisson distribution is the set of nonnegative integers $\mathbb{X} = \mathbb{N}_0 = \{0, 1, 2, \dots\}$; the distribution has one parameter – a nonnegative real number – which is usually called "intensity" and is denoted by $\lambda \in \mathbb{R}_+$; while the notation indicating that a random variable X follows the Poisson distribution with intensity parameter λ is the following.

$$X \sim \text{Pois}(\lambda)$$

The probability mass function of the Poisson distribution is:

$$\mathbb{P}(X = x; \lambda) = f_X(x; \lambda) = \frac{\exp(-\lambda) \cdot \lambda^x}{x!} \tag{2.22}$$

its cumulative distribution is:

$$\mathbb{P}(X \leq x; \lambda) = F_X(x; \lambda) = \exp(-\lambda) \sum_{i=0}^{\lfloor x \rfloor} \frac{\lambda^i}{i!} \tag{2.23}$$

and by exploiting the Taylor expansion of the exponential function, it is:

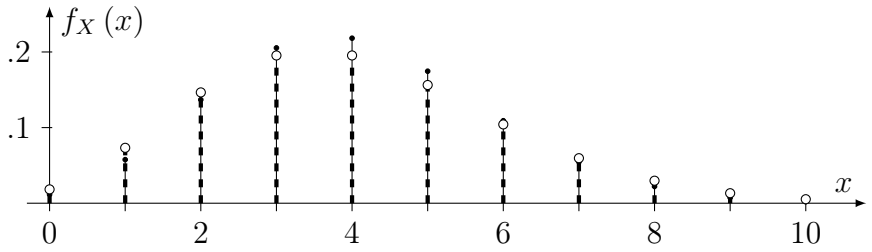
$$\begin{aligned} \lim_{M \to \infty} F_X(M; \lambda) &= \exp(-\lambda) \cdot \lim_{M \to \infty} \sum_{x=0}^{M} \frac{\lambda^x}{x!} \\ &= \exp(-\lambda) \cdot \exp(\lambda) \\ &= 1 \end{aligned}$$

showing compliance with the definition of probability distribution.

An important and well-known feature of the Poisson distribution is that it resembles – i.e. it mathematically approximates – a binomial distribution where n is "large" and p is "small." To show this, suppose that $X \sim \text{Bn}(p, n)$ and let $\lambda = np$ be some fixed number. Consider the limit of the probability mass function for some $X = x$ as n goes to infinity; for λ fixed, this implies that p goes to zero, therefore it is more convenient to express the limit solely in terms of n and λ, as in the following derivation.

$$\begin{aligned} \lim_{n \to \infty} f_X(x; p, n) &= \lim_{n \to \infty} \binom{n}{x} p^x (1-p)^{n-x} \\ &= \lim_{n \to \infty} \frac{n!}{x!\,(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\ &= \frac{\lambda^x}{x!} \lim_{n \to \infty} \underbrace{\left(\frac{\prod_{k=1}^{x}(n-k+1)}{n^x}\right)}_{\to 1} \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-x}}_{\to 1} \underbrace{\left(1 - \frac{\lambda}{n}\right)^{n}}_{\to \exp(-\lambda)} \\ &= \frac{\lambda^x}{x!} \exp(-\lambda) \\ &= f_X(x; \lambda) \end{aligned}$$

The approximation is quite good even for moderately large values of n and moderately small values of p, as shown in the example from Figure 2.1.



Note: binomial probabilities are denoted with solid thin lines, smaller full points; Poisson probabilities are denoted with dashed thicker lines, larger hollow points. All probabilities for $x > 10$ are negligible.

Figure 2.1: Binomial vs. Poisson comparison with $n = 20$, $p = 0.2$, $\lambda = 4$

This mathematical relationship provides both intuition and motivation to support the use of the Poisson distribution as a model for the probability that a number X of events occurs in a specified (usually small) interval of time or space. Common examples are the number of phone calls (or emails) received in a time interval of interest, or the number of sewing imperfections found upon a uniform section of textile. The hypotheses that underpin the use of the Poisson distribution in situation of this kind are that:

- the events of interest happen independently, they are all equally likely, and they cannot overlap (in time, in space, *et cetera*);
- the larger the interval under examination, the higher the probability to encounter a single event.

The Taylor expansion of the exponential functions is also useful to derive the moment generating function of the Poisson distribution:

$$\begin{aligned} M_X(t;\lambda) &= \lim_{M\to\infty} \sum_{x=0}^{M} \exp(tx) \cdot \frac{\exp(-\lambda)\cdot\lambda^x}{x!} \\ &= \exp(-\lambda) \cdot \lim_{M\to\infty} \sum_{x=0}^{M} \frac{[\lambda\cdot\exp(t)]^x}{x!} \\ &= \exp(-\lambda)\cdot\exp(\lambda\cdot\exp(t)) \\ &= \exp(\lambda(\exp(t)-1)) \end{aligned} \tag{2.24}$$

from which one can calculate the mean and variance as follows.

$$\mathbb{E}[X] = \lambda \tag{2.25}$$

$$\mathbb{V}\mathrm{ar}[X] = \lambda \tag{2.26}$$

Both moments are equal to one another, as well as to the parameter λ! This motivates the name "intensity" for the latter: λ represents the average rate at which the events occur; the more frequent they are, the larger is the average number of "successes" as well as their dispersion on the support $\mathbb{N}_0$.

Discrete uniform distribution

There are certainly discrete distribution unrelated to the notion of Bernoulli trials. Perhaps the simplest example of such a distribution is the "discrete uniform" one, where *discrete* marks a difference with the analogous continuous distribution already encountered in several examples from Lecture 1. This simple distribution has *equal probability mass* over N mass points:

$$\mathbb{P}(X = x; N) = f_X(x; N) = \frac{1}{N} \tag{2.27}$$

where $|\mathbb{X}| = N$. Naturally, the other features of this distribution depend on the specific points in the support $\mathbb{X}$. At one extreme, these points are the first N integers: $\mathbb{X} = \{1, 2, \ldots, N\}$, in this case, calculating the cumulative distribution and the moments of X is a simple exercise; at the other extreme, these can be irregularly selected points, and those calculations may be non-trivial and context-dependent.

Between these two extremes, a quite frequent scenario is that where $\mathbb{X}$ contains all the integers that belong to some interval $[a, b]$ of length $N - 1$: $\mathbb{X} = \{x \in \mathbb{Z} : x = a, \ldots, b\}$, $b - a = N - 1$. In this case, one usually writes:

$$X \sim \mathcal{U}\{a, b\}$$

denoting that the distribution has two parameters, a and b, which together determine $N = b - a + 1$. The cumulative distribution is:

$$\mathbb{P}(X \leq x; a, b) = F_X(x; a, b) = \frac{\lfloor x \rfloor - a + 1}{b - a + 1} \cdot \mathbb{1}[a \leq x \leq b] + \mathbb{1}[b < x] \tag{2.28}$$

the moment generating function is:

$$M_X(t; a, b) = \frac{\exp(at) - \exp((b + 1)t)}{N(1 - \exp(t))} \tag{2.29}$$

while the mean and variance are as follows.

$$\mathbb{E}[X] = \frac{a + b}{2} \tag{2.30}$$

$$\mathbb{V}\text{ar}[X] = \frac{N^2 - 1}{12} \tag{2.31}$$

In this case, it may be easier to derive the two moments by direct application of their definition than by using the moment generating function.

Hypergeometric distribution

The last discrete distribution considered here is a variation of the binomial experiment of obtaining x successes out of n attempts of a binary outcome, but in an environment where the probability of each success is *not* fixed, and thus the correspondence with n independent Bernoulli trials *cannot* be established. Specifically, the *hypergeometric* distribution is modeled on the idea of randomly selecting $n \in \mathbb{N}$ objects out of a population that contains $N \in \mathbb{N}$ of them in total, of which $K \in \mathbb{N}$ presents a certain binary feature that is called a "success," and where $n < N$ and $K < N$. The two most common concrete representations of this mental experiment are:

- the *urn model* about the extractions of certain n items ("balls") from a container ("urn") – the items-balls are N in total and K of them present a certain feature ("color");
- the concept of *sampling without replacement* from a finite population of size N, K of whose elements present a certain feature; this corresponds to the case of a statistical sample which is obtained by randomly drawing n individual units from the population, one at a time, and excluding them from additional draws once they are selected.

The random variable X that represents the number of "successes" x out of the n "extractions" is said to follow the hypergeometric distribution, also written as:

$$X \sim \mathcal{H}\left(N, K, n\right)$$

which presents three parameters: N, K, and n. Its support $\mathbb{X}$ must satisfy the following four conditions: $x \geq 0$, $x \leq n$, (these two are obvious) $x \leq K$ and $n - x \leq N - K$ (actual successes and failures cannot exceed the possible maxima). Combining all these inequalities together, the support is written as follows.

$$\mathbb{X} = \left\{\max\left(0, n + K - N\right), \ldots, \min\left(n, K\right)\right\}$$

The hypergeometric probability mass function presents an expression composed by three binomial coefficients:

$$\mathbb{P}\left(X = x; N, K, n\right) = f_X\left(x; N, K, n\right) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} \tag{2.32}$$

which, however, has an intuitive explanation. There are in total $\binom{N}{n}$ ways to extract a sample of size n out of a population of size N; under the hypotheses

underpinning the hypergeometric distribution, each of these has an identical probability to occur. Therefore, in order to calculate the probability that x successes are achieved, one must multiply the total number of combinations that x successes are obtained out of the K units with that defining feature, times the number of ways that $n-x$ failures realize out of the residual units $N-K$ lacking that feature. This is what is expressed in (2.32).

Yet manipulating (2.32) is complicated and it requires familiarity with combinatorics. The cumulative distribution function is obtained by appropriately summing over successive elements of the support and can expressed in terms of the hypergeometric function, which gives its name to the distribution itself. Furthermore, it can be shown that this function equals one upon summing over all the elements of $\mathbb{X}$. Likewise, the moment generating function is also expressed in terms of the hypergeometric function. The mean and the variance are best obtained through direct application of the definition; for example:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x\in\mathbb{X}} x \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} \\
&= \sum_{x\in\mathbb{X}} \frac{K\binom{K-1}{x-1}\binom{N-K}{n-x}}{\frac{N}{n}\binom{N-1}{n-1}} \\
&= n\frac{K}{N}\sum_{y\in\mathbb{Y}} \frac{\binom{K-1}{y}\binom{(N-1)-(K-1)}{n-1-y}}{\binom{N-1}{n-1}} \\
&= n\frac{K}{N}
\end{aligned} \tag{2.33}$$

where the summation in the third line equals one because it corresponds to the total probability mass of another hypergeometric distribution with parameters $(N-1, K-1, n-1)$, with $y = x-1$ and where $\mathbb{Y}$ is the "rescaled" support of $Y = X-1$. Intuitively, the mean of X equals is proportional to the ratio of potential successes in the population, K/N, multiplied by the number of attempts n. To calculate the variance:

$$\mathbb{V}\text{ar}[X] = n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1} \tag{2.34}$$

it is convenient to compute the second uncentered moment $\mathbb{E}[X^2]$ first.

2.2 Continuous Distributions I

The treatment of continuous distributions starts in this section with the general, parametric version of the normal distribution, and proceeds – still in this section – with other distributions that are similar to the normal, in the sense that they all present two key parameters: a **location parameter** and a **scale parameter**. These two parameters define two key features of each distribution: the location parameter characterizes the relative *position* of the distribution on $\mathbb{R}$ (intuitively, the area where most of the probability is found) while the shape parameter defines the actual form of the density function and the overall probability dispersion around the mean (which may or may not linked to specific moments of the distribution). These concepts are worth of a formal treatment.

Definition 2.1. Location and scale families. Let $f_Z(z)$ be a probability density function associated with some random variable Z. For any $\mu \in \mathbb{R}$ and any $\sigma \in \mathbb{R}_{++}$, the family of probability density functions of the form

$$f_X(x) = \frac{1}{\sigma} f_Z\left(\frac{x-\mu}{\sigma}\right)$$

for a generic random variable X is called the **location-scale family** with **standard probability density function** $f_Z(z)$; μ is called the **location parameter** while σ is called the **scale parameter**.

The distributions belonging to a location-scale family are inextricably connected, in the sense expressed by the following theorem.

Theorem 2.1. Standardization of densities. *Let $f(\cdot)$ be any probability density function, $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_{++}$. Then, a random variable X follows a probability distribution with density function:*

$$f_X(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$$

if and only if there exists a continuous random variable Z whose probability density function is $f_Z(z) = f(z)$ and $X = \sigma Z + \mu$.

Proof. Necessity is shown by noting that $X = g(Z) = \sigma Z + \mu$, where $g(\cdot)$ is a monotone transformation with $g^{-1}(x) = (x-\mu)/\sigma$, $\left|\frac{d}{dx} g^{-1}(x)\right| = \sigma^{-1}$ and thus the conditions of Theorem 1.11 apply. Sufficiency is shown through the converse exercise: define $Z = g(X) = (X-\mu)/\sigma$, again a monotone transformation with $g^{-1}(z) = \sigma z + \mu$, $\left|\frac{d}{dz} g^{-1}(z)\right| = \sigma$ and again one can extend the theorem for monotone transformations of density functions. □

Some very useful implications of this result are that:

- a standard density function (with associated distribution) exists for every location-scale family;
- given that all the distributions in a location-scale family are all linked through a linear transformation, their mean, variance, other moments and moment generating functions are related via simple functions;
- all probabilities specific to a distribution from a location-scale family can be expressed with reference to the standard distribution.

$$\begin{aligned}\mathbb{P}(a \leq X \leq b) &= \mathbb{P}\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) \\ &= \mathbb{P}\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right)\end{aligned}$$

The usefulness of these facts is best understood when put in context.

Normal distribution

The all-important location-scale family of normal distributions, sometimes called *Gaussian*, includes all continuous distributions with support on the whole of $\mathbb{R}$, density function of the form

$$f_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{2.35}$$

and cumulative distribution obtained by appropriate integration as follows.

$$F_X(x; \mu, \sigma^2) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt \tag{2.36}$$

These distributions present two parameters, μ for location and σ for shape, although the latter is typically replaced with its square σ^2 for an immediate interpretation as variance (see below). The expression

$$X \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

indicates that X follows the normal distribution with the specified parameters. In the case of the *standard normal distribution* from Examples 1.7 and 1.9, with density $\phi(x)$ and cumulative distribution $\Phi(x)$, these parameters are equal to 0 and 1 respectively, hence the above writes:

$$X \sim \mathcal{N}(0, 1)$$

although X is often replaced with Z for clearer indication of standardization – e.g. $Z \sim \mathcal{N}(0, 1)$.

Figure 2.2 shows three examples of normal density functions. The continuous line represents the familiar standard version. The dashed line displays a density with $\mu = 2$ and $\sigma^2 = 1$: note how an increase of the location parameter produces a "shift to the right" of the distribution (if the location parameter is decreased, one would obtain a "shift to the left"). The dotted line depicts a density with $\mu = 0$ and $\sigma^2 = 4$: while still centered at zero, an increase in the shape parameter produces a more "flattened out," dispersed distribution (conversely, one would obtain a more concentrated distribution if the shape parameter is decreased).

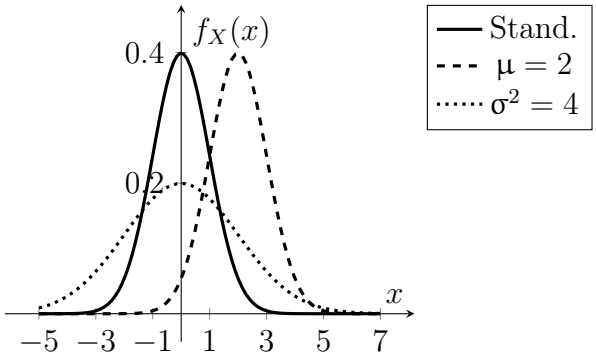


Figure 2.2: Different normal probability densities

Occasionally, the normal distribution is expressed through the following alternative parametrization:

$$f_X(x; \mu, \phi^2) = \sqrt{\frac{\phi^2}{2\pi}} \exp\left(-\frac{\phi^2 (x-\mu)^2}{2}\right) \tag{2.37}$$

where $\phi^2 = \sigma^{-2}$ is called the *precision parameter*. In Figure 2.2, the dotted density has $\phi^2 = 0.25$.

Showing that the density function of a normal distribution integrates to one is not an immediate task; thanks to Theorem 2.1, it is enough to show that the result holds in the standardized case, that is:

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz = 1$$

or, exploiting symmetry around zero, that the following holds.

$$\int_{0}^{\infty} \exp\left(-\frac{z^2}{2}\right) dz = \frac{\sqrt{2\pi}}{2} = \sqrt{\frac{\pi}{2}} \tag{2.38}$$

This is accomplished by squaring the left-hand side of (2.38), manipulating it, and showing that it equals to $\pi/2$:

$$\begin{aligned}\left(\int_0^\infty \exp\left(-\frac{z^2}{2}\right) dz\right)^2 &= \left(\int_0^\infty \exp\left(-\frac{t^2}{2}\right) dt\right)\left(\int_0^\infty \exp\left(-\frac{u^2}{2}\right) du\right) \\ &= \int_0^\infty \int_0^\infty \exp\left(-\frac{t^2+u^2}{2}\right) dtdu \\ &= \int_0^\infty \int_0^{\frac{\pi}{2}} r \cdot \exp\left(-\frac{r^2}{2}\right) d\theta dr \\ &= \frac{\pi}{2}\int_0^\infty r \cdot \exp\left(-\frac{r^2}{2}\right) dr \\ &= \frac{\pi}{2}\left[-\exp\left(-\frac{r^2}{2}\right)\Big|_0^\infty\right] \\ &= \frac{\pi}{2}\end{aligned}$$

where the third line implements a change of variables in polar coordinates, $t = r \cdot \cos(\theta)$ and $u = r \cdot \sin(\theta)$.

The derivation of the moment generating function is fortunately way easier. It is again best to work with the standardized random variable Z:

$$\begin{aligned}M_Z(t) &= \int_{-\infty}^{+\infty} \exp(tz) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz \\ &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2 - 2zt + t^2 - t^2}{2}\right) dz \\ &= \exp\left(\frac{t^2}{2}\right) \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-t)^2}{2}\right) dz \\ &= \exp\left(\frac{t^2}{2}\right)\end{aligned}$$

where the integral in the third line vanishes because it corresponds to the density function of a normal distribution with $\mu = t$ and $\sigma^2 = 1$ integrated over its entire support, and hence equal to one. By the properties of moment generating functions, for any normally distributed random variable obtained as $X = \sigma Z + \mu$:

$$M_X\left(t; \mu, \sigma^2\right) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right) \tag{2.39}$$

which shows that:

$$\mathbb{E}[X] = \mu \tag{2.40}$$

$$\mathbb{V}\mathrm{ar}[X] = \sigma^2 \tag{2.41}$$

that is, the two parameters of the normal distribution have an immediate interpretation in terms of fundamental moments, a quite convenient fact! In addition, it can be shown that:

$$\mathbb{S}\text{kew}[X] = 0 \tag{2.42}$$

$$\mathbb{K}\text{urt}[X] = 3 \tag{2.43}$$

that is, all normal distributions have zero skewness (they are all perfectly symmetric) and kurtosis equal to three, which makes this number a reference value for evaluating the kurtosis of other distributions.[2]

The normal distribution has ubiquitous applications that are motivated by its numerous relationships with other probability distributions and, most importantly, by the asymptotic result known as Central Limit Theorem, in its various versions (Lecture 6). For these reasons, the normal distribution is central in statistical inference as well as in econometric analysis.

Lognormal distribution

The *lognormal* distribution has the following probability density function, for $y \in \mathbb{R}_{++}$, $\mu \in \mathbb{R}$, and $\sigma^2 \in \mathbb{R}_{++}$.

$$f_Y(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{y} \exp\left(-\frac{(\log y - \mu)^2}{2\sigma^2} \right) \tag{2.44}$$

Observation 2.2. By Theorem 1.11, one can observe that the lognormal distribution is evidently obtained through the transformation $Y = \exp(X)$, where $X \sim \mathcal{N}(\mu, \sigma^2)$. Conversely, $X = \log(Y)$: the logarithm of a random variable Y which follows the lognormal distribution is normally distributed, hence the former distribution's name.

The cumulative distribution of the lognormal distribution is:

$$F_Y(y; \mu, \sigma^2) = \int_0^y \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{t} \exp\left(-\frac{(\log t - \mu)^2}{2\sigma^2} \right) dt \tag{2.45}$$

and the density must clearly integrate to one since the normal distribution does. In order to specify that some random variable Y follows the lognormal distribution, the most convenient way is surely as follows.

$$\log(Y) \sim \mathcal{N}(\mu, \sigma^2)$$

[2] In fact, the fourth standardized central moment of some random variable X is often expressed in terms of "excess kurtosis" $\mathbb{K}\text{urt}[X] - 3$ that is, as the differences between the kurtosis of X and that of the normal distribution.

Figure 2.3 displays example of lognormal density functions that are analogous to those from Figure 2.2, highlighting how lognormal densities can assume different shapes, always asymmetric by some degree; for example, the standard version is recognizable by its characteristic hump. These characteristics make the lognormal distribution well suited to describe phenomena that only take positive values and that are characterized by apparent "inequality," such as the income distribution or that of firm size.

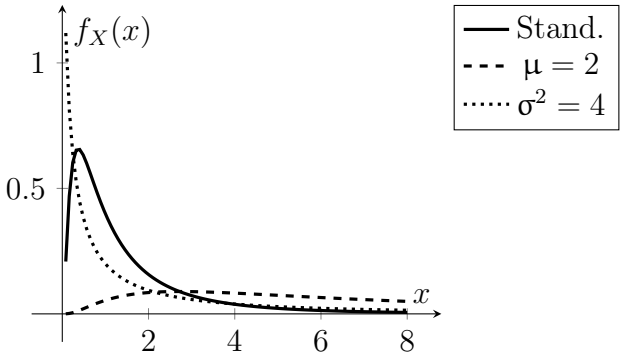


Figure 2.3: Different lognormal probability densities

The lognormal distribution is one for which *the moment generating function does not exist* (the integral that defines it diverges), and even the characteristic function takes a very complex expression. Fortunately, uncentered moments can be calculated easily:

$$\begin{aligned}\mathbb{E}\left[Y^r\right] &= \mathbb{E}\left[\left(\exp\left(X\right)\right)^r\right] \\ &= \mathbb{E}\left[\exp\left(Xr\right)\right] \\ &= \exp\left(\mu r + \frac{\sigma^2 r^2}{2}\right)\end{aligned}$$

because the second equality corresponds with the definition of moment generating function of $X \sim \mathcal{N}(\mu, \sigma^2)$ given $r = t$. The mean and variance, for example, are obtained as follows.

$$\mathbb{E}\left[Y\right] = \exp\left(\mu + \frac{\sigma^2}{2}\right) \tag{2.46}$$

$$\mathbb{V}\mathrm{ar}\left[Y\right] = \left[\exp\left(\sigma^2\right) - 1\right]\exp\left(2\mu + \sigma^2\right) \tag{2.47}$$

The lognormal distribution is obviously asymmetric and clearly, it always shows a positive skewness which is a function of the scale parameter.

$$\mathbb{S}\mathrm{kew}\left[Y\right] = \left[\exp\left(\sigma^2\right) + 2\right] \cdot \sqrt{\exp\left(\sigma^2\right) - 1} > 0 \tag{2.48}$$

Logistic distribution

Even the standard logistic distribution introduced with Examples 1.7 and 1.9 has a full-fledged location-scale family. With support $\mathbb{X} = \mathbb{R}$, a location parameter $\mu \in \mathbb{R}$, and a scale parameter $\sigma \in \mathbb{R}_{++}$, the probability density function of a generic logistic distribution is written as follows.

$$f_X(x; \mu, \sigma) = \frac{1}{\sigma} \exp\left(-\frac{x-\mu}{\sigma}\right) \left[1 + \exp\left(-\frac{x-\mu}{\sigma}\right)\right]^{-2} \tag{2.49}$$

Also the logistic cumulative distribution has a closed form expression, which can be obtained by an exercise in integrating the density above.

$$F_X(x; \mu, \sigma) = \left[1 + \exp\left(-\frac{x-\mu}{\sigma}\right)\right]^{-1} \tag{2.50}$$

making it obvious that $\lim_{x\to\infty} F_X(x; \mu, \sigma) = 1$. Expression (2.50) is easy to manipulate and invert; consequently the logistic distribution has a simple expression for its quantile function.

$$Q_X(p; \mu, \sigma) = \mu + \sigma \log\left(\frac{p}{1-p}\right) \quad \text{for } p \in (0, 1) \tag{2.51}$$

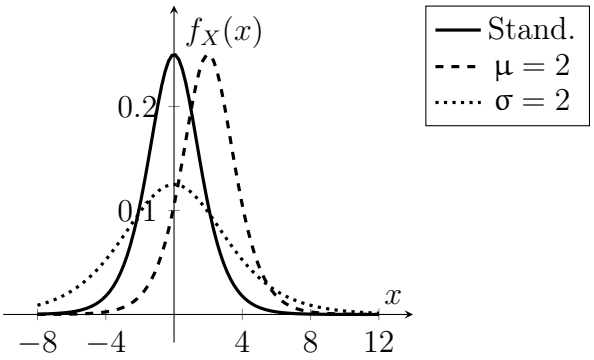


Figure 2.4: Different logistic probability densities

As shown above in Figure 2.4, and as already observed with reference to the respective standardized distributions, the logistic density function has a symmetric bell shape similar to the normal's. However, if the two distributions have equal variances the logistic is "thinner" than the normal around the mean, and has "thicker" outer tails. A random variable X following the logistic distribution can be denoted with the following notation.

$$X \sim \text{Logistic}(\mu, \sigma)$$

The moment generating function of the *standard* logistic distribution is obtained via a change of variable in the integral that defines the function, where the function being applied is the cumulative distribution itself.

$$u = \frac{1}{1 + \exp(-z)}$$

This implies a range of integration restricted to $(0, 1)$, that:

$$\exp(tz) = \left(\frac{1}{\exp(-z)}\right)^t = \left(\frac{u}{1-u}\right)^t = u^t (1-u)^{-t}$$

and that the differentials *and the density function* are related as follows.

$$du = \frac{\exp(-z)}{(1 + \exp(-z))^2} dz$$

All this implies that:

$$\begin{aligned} M_Z(t) &= \int_{-\infty}^{\infty} \exp(tz) \frac{\exp(-z)}{(1 + \exp(-z))^2} dz \\ &= \int_0^1 u^t (1-u)^{-t} du \\ &= \mathrm{B}(1+t, 1-t) \end{aligned}$$

as the second line is recognized as a particular case of the **Beta Function**, an important mathematical function with two arguments $a > 0$ and $b > 0$.

$$\mathrm{B}(a, b) \equiv \int_0^1 u^{a-1} (1-u)^{b-1} du$$

Clearly, here the correspondence is obtained for $a = 1 + t$ and $b = 1 - t$. The Beta function is notoriously symmetric in its two arguments, thus the moment generating function of the standard logistic is also (perhaps more commonly) written as:

$$M_Z(t) = \int_0^1 u^{-t} (1-u)^t du = \mathrm{B}(1-t, 1+t)$$

and since the Beta generating function is defined only for positive values of a and b, the moment generating function is only defined for $t \in (-1, 1)$. For a general logistic distribution such that $X = \sigma Z + \mu$, by the properties of moment generating functions the one of X is:

$$M_X(t; \mu, \sigma) = \exp(\mu t) \cdot \int_0^1 u^{-\sigma t} (1-u)^{\sigma t} du \tag{2.52}$$

or, in terms of the Beta Function:

$$M_X(t;\mu,\sigma) = \exp(\mu t)\cdot \mathrm{B}(1-\sigma t, 1+\sigma t)$$

and its domain is thus restricted to $t \in (-\sigma^{-1}, \sigma^{-1})$.

By the properties of the Beta function, the mean and the variance of a generic logistic distribution are obtained as follows.

$$\mathbb{E}[X] = \mu \tag{2.53}$$

$$\mathbb{V}\mathrm{ar}[X] = \frac{\sigma^2\pi^2}{3} \tag{2.54}$$

Observe that unlike the standard normal distribution, the standard logistic distribution with $\mu = 0$ and $\sigma = 1$ has variance greater than one. In order to appropriately the variance so as to equal 1, a reparametrization:

$$\sigma^* = \frac{\sqrt{3}}{\pi}\sigma$$

is occasionally employed. One can also demonstrate that:

$$\mathbb{S}\mathrm{kew}[X] = 0 \tag{2.55}$$

$$\mathbb{K}\mathrm{urt}[X] = \frac{21}{5} \tag{2.56}$$

showing that the logistic distribution has always a kurtosis higher than the normal's, a fact that matches some previous observations about the shape of the two distributions. Thanks to its similarity with the normal, the logistic distribution is typically used as an approximation of the former when easy analytical manipulation of the cumulative distribution is necessary.

Cauchy distribution

The Cauchy distribution is an interesting "pathological" case of a location-scale family of symmetric, bell-shaped distributions having $\mathbb{X} = \mathbb{R}$ as their support. Its density functions writes, given a location parameter $\mu \in \mathbb{R}$ and a scale parameter $\sigma \in \mathbb{R}_{++}$, as:

$$f_X(x;\mu,\sigma) = \frac{1}{\pi\sigma}\left[1 + \left(\frac{x-\mu}{\sigma}\right)^2\right]^{-1} \tag{2.57}$$

while the cumulative distribution has the following closed form expression, displaying the necessary property $\lim_{x\to\infty} F_X(x;\mu,\xi) = 1$.

$$F_X(x;\mu,\sigma) = \frac{1}{\pi}\arctan\left(\frac{x-\mu}{\sigma}\right) + \frac{1}{2} \tag{2.58}$$

Since the above is an invertible function, also the quantile function of the Cauchy distribution has a simple closed form expression, for $p \in (0, 1)$.

$$Q_X(p; \mu, \sigma) = \mu + \sigma \tan\left(\pi\left(p - \frac{1}{2}\right)\right) \tag{2.59}$$

A random variable X which is distributed according to the Cauchy distribution is denoted as follows.

$$X \sim \text{Cauchy}(\mu, \sigma)$$

The Cauchy probability density functions displayed in Figure 2.5 below appear not too dissimilar from the shapes of the normal and logistic distributions, however the Cauchy is an interesting pathological case because *its moment generating functions, and its moments themselves, do not exist* in the sense that the integrals that characterize them have no defined solution.

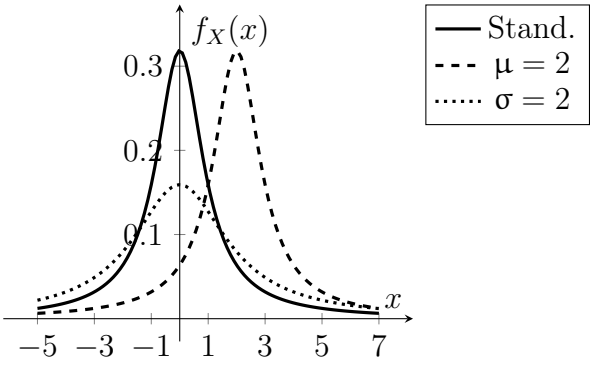


Figure 2.5: Different Cauchy probability densities

To show this property, one typically works with the standardized case $Z = (X - \mu)/\sigma$, whose expectation is rewritten as follows:

$$\mathbb{E}[Z] = \int_{-\infty}^{0} \frac{1}{\pi} \frac{z}{1+z^2}\, dz + \int_{0}^{+\infty} \frac{1}{\pi} \frac{z}{1+z^2}\, dz$$

that is, the integral that defines the mean is split between two parts that, since the distribution is symmetric around $z = 0$, must have opposite signs but equal absolute value. Consider the integral defined on $\mathbb{R}_{++}$; note that:

$$\int_{0}^{+\infty} \frac{z}{1+z^2}\, dz = \lim_{M \to \infty} \frac{\log(1+z^2)}{2}\bigg|_{0}^{M} = \infty$$

the integral does not converge and lacks a finite solution. Because the same applies to the other half of the partition above, the value of $\mathbb{E}[Z]$ remains undefined. Similar arguments also apply to the non-standardized versions of the distribution, the higher moments, and the moment generating function. However, the Cauchy distributions – like all distributions – always has a characteristic function, which can be shown to be the following.

$$\varphi_X(t; \mu, \sigma) = \exp(i\mu t - \sigma|t|) \tag{2.60}$$

Note that this particular characteristic function is not differentiable at $t = 0$, thus it cannot help derive the moments of the distribution. As it lacks the moments, the Cauchy distribution has limited practical applications; it is however notable for its links – to be illustrated later – with distributions of more practical use such as the normal and the Student's t-distribution.

Laplace distribution

The last location-scale family discussed here is that of the Laplace distribution, characterized by support $\mathbb{X} = \mathbb{R}$ and the following density function.

$$f_X(x; \mu, \sigma) = \frac{1}{2\sigma} \exp\left(-\frac{|x - \mu|}{\sigma}\right) \tag{2.61}$$

Note that this density function is discontinuous at $x = \mu$, but despite this, the above density is easy to integrate on the two complementary subsets of $\mathbb{R}$ that are split at the discontinuity point. Hence, the cumulative density function has two distinct closed form expressions:

$$F_X(x; \mu, \sigma) = \begin{cases} \frac{1}{2}\exp\left(\frac{x-\mu}{\sigma}\right) & \text{if } x < \mu \\ 1 - \frac{1}{2}\exp\left(-\frac{x-\mu}{\sigma}\right) & \text{if } x \geq \mu \end{cases} \tag{2.62}$$

and is notably continuous at $x = \mu$. This expression clearly conforms to the properties of a probability distribution and it integrates to 1 over $\mathbb{R}$. Since the distribution is symmetric, the quantile for $p = 0.5$ corresponds to the mean; hence the quantile function is also split in two expressions for values of p that are either below or above $p = 0.5$.

$$Q_X(p; \mu, \sigma) = \begin{cases} \mu + \sigma\log(2p) & \text{if } p \in \left(0, \frac{1}{2}\right] \\ \mu - \sigma\log(2 - 2p) & \text{if } p \in \left[\frac{1}{2}, 1\right) \end{cases} \tag{2.63}$$

Should a random variable X follow the Laplace distribution, this is usually expressed with the following notation.

$$X \sim \text{Laplace}(\mu, \sigma)$$

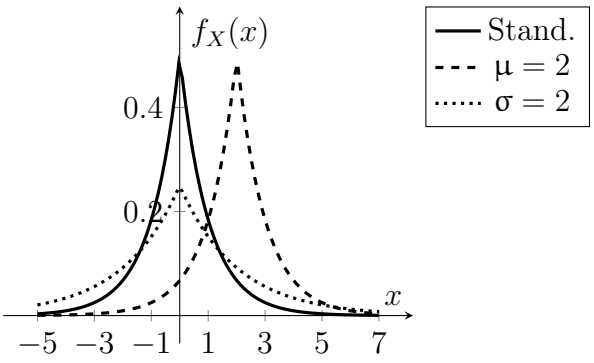


Figure 2.6: Different Laplace probability densities

The three Laplace densities displayed in Figure 2.6 all take the typical "tent shape" that characterizes this distribution. Unlike in the Cauchy case, this distribution has well-defined moments; to see this, it is easiest to start from calculating the moment generating function in the standardized case:

$$\begin{aligned} M_Z(t) &= \int_{-\infty}^{+\infty} \frac{1}{2} \exp(tz - |z|)\, dz \\ &= \frac{1}{2} \int_{-\infty}^{0} \exp((1+t)z)\, dz + \frac{1}{2} \int_{0}^{+\infty} \exp(-(1-t)z)\, dz \\ &= \frac{1}{2} \left[\frac{1}{1+t} + \frac{1}{1-t} \right] \\ &= \frac{1}{1-t^2} \end{aligned}$$

which is only defined for $|t| < 1$ (or else the two integrals in the second line diverge). By the properties of the moment generating functions for linearly transformed random variables, it is easy to see that in the general case:

$$M_X(t; \mu, \sigma) = \frac{\exp(\mu t)}{1 - \sigma^2 t^2} \tag{2.64}$$

where, for analogous reasons, this moment generating function is defined only for $|t| < \sigma^{-1}$. By (2.64) it is possible to obtain all the other moments; the mean and variance for example are as follows.

$$\mathbb{E}[X] = \mu \tag{2.65}$$

$$\mathbb{V}\mathrm{ar}[X] = 2\sigma^2 \tag{2.66}$$

The Laplace distribution has a limited number of applications (it can be used, for example, to model the *growth rates* of firms) but is mostly known

for its relationships with other distributions – in particular, the exponential distribution. In an anticipation of a discussion developed in the next section, if a random variable X follows the Laplace distribution, the transformation $Y = |X - \mu|$ (which is obtained by "mirroring" X around its mean) follows the exponential distribution, a fact that gained the alternative name *double exponential distribution* to the Laplace distribution.

2.3 Continuous Distributions II

The location-scale families of continuous distributions do not certainly exhaust the set of relevant continuous probability distributions. The distributions that follow next in the discussion are all continuous, do not generally follow in that category, yet they have theoretical or practical importance. Some of these distributions have a special place in the theory and practice of statistical inference. In the ensuing discussion, some emphasis is placed on those relationships between distributions *that can be described in terms of univariate transformations*. Other relationships, including some that are especially relevant for statistical inference, require the development of concepts such as those of joint probability distributions, independent random variables, random samples, and convergence in distribution; for this reason, the illustration of these relationships is postponed to later lectures.

Continuous uniform distribution

Lecture 1 already provides ample discussion of the uniform distribution on the unit interval, including its moments and moment generating function. Similarly to its discrete analogue, the continuous uniform distribution can be generalized to have support on any segment $\mathbb{X} = [a, b]$ of the real line, where the extremes a and b of the interval are expressed as two *parameters*. The general form of the density function is:

$$f_X(x; a, b) = \frac{1}{b-a} \cdot \mathbb{1}[x \in [a, b]] \tag{2.67}$$

while that of the cumulative distribution is as follows.

$$F_X(x; a, b) = \frac{x-a}{b-a} \cdot \mathbb{1}[x \in (a, b)] + \mathbb{1}[x \in [b, \infty)] \tag{2.68}$$

Obviously, the two parameters a and b (where $b > a$) characterize both the position and the overall "spread" of the distribution; for this reason, the family of uniform distributions is occasionally classified among the location-scale families. However, the two parameters a and b here play a symmetric

role; neither of them is more characteristic of the location or the scale of the distribution (like μ and σ do for the distributions examined previously). If a random variable X follows the uniform distribution on the $[a, b]$ interval, one usually writes:

$$X \sim \mathcal{U}(a, b)$$

where $X \sim \mathcal{U}(0, 1)$ is just but a special case. By generalizing the examples given in the previous lecture, it is straightforward to verify that the moment generating function of a generic uniform distribution are:

$$M_X(t; a, b) = \begin{cases} \frac{1}{t(b-a)} \left[\exp(bt) - \exp(at)\right] & \text{if } t \neq 0 \\ 1 & \text{if } t = 0 \end{cases} \tag{2.69}$$

while the mean and variance are as follows.

$$\mathbb{E}[X] = \frac{a+b}{2} \tag{2.70}$$

$$\mathbb{V}\text{ar}[X] = \frac{1}{12}(b-a)^2 \tag{2.71}$$

Depending on the context of interest, all uniform distributions can be alternatively defined on the open interval (a, b); the analysis of the distribution is largely unaffected whether the support is an open or a closed set.

Beta distribution

With the expression *Beta distributions* one usually refers to a family of distributions that, like the uniform distribution, have support on a segment of the real line, but unlike the uniform distribution, can take varying shapes. The starting point in their description is the *standard* family of Beta distributions with support on the unit interval, $\mathbb{X} = (0, 1)$. These particular distributions are defined by two positive parameters $\alpha \in \mathbb{R}_{++}$ and $\beta \in \mathbb{R}_{++}$ that jointly define the **shape** of the density function, which reads:

$$f_X(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du} = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha, \beta)} \quad \text{for } x \in (0, 1) \tag{2.72}$$

where $\mathrm{B}(\alpha, \beta)$ is a Beta function with the parameters α and β as arguments and serves as a normalization constant to ensure that the density integrates to one on the unit interval. The Beta function is also related to the so-called **Gamma function** $\Gamma(\gamma)$, a function with one argument $\gamma \in \mathbb{R}_{++}$:[3]

$$\Gamma(\gamma) = \int_0^\infty u^{\gamma-1} \exp(-u)\,du$$

[3]Specifically, it can be shown that $\mathrm{B}(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$.

and therefore (2.72) can be alternatively written as follows.

$$f_X(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot x^{\alpha-1}(1-x)^{\beta-1} \quad \text{for } x \in (0,1) \tag{2.73}$$

The cumulative distribution is:

$$F_X(x;\alpha,\beta) = \frac{\int_0^x t^{\alpha-1}(1-t)^{\beta-1}\,dt}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du} = \frac{\mathrm{B}(x;\alpha,\beta)}{\mathrm{B}(\alpha,\beta)} \tag{2.74}$$

where $\mathrm{B}(x;a,b) \equiv \int_0^x t^{a-1}(1-t)^{b-1}\,dt$ is the so-called **lower incomplete Beta function**; it is thus obvious that the cumulative distribution function integrates to 1 over the $(0,1)$ interval. To express that a random variable X follows a *standard* Beta distribution it is common to write as follows.

$$X \sim \mathrm{Beta}(\alpha,\beta)$$

It must be observed that for many Beta distributions, the support can be defined in terms of the closed interval $[0,1]$ without affecting the analysis.

Observation 2.3. $X \sim \mathrm{Beta}(1,1)$ is equivalent to $X \sim \mathcal{U}(0,1)$, that is, the uniform distribution on the unit interval is a special case of the standard Beta distribution, for parameters $\alpha = 1$ and $\beta = 1$.

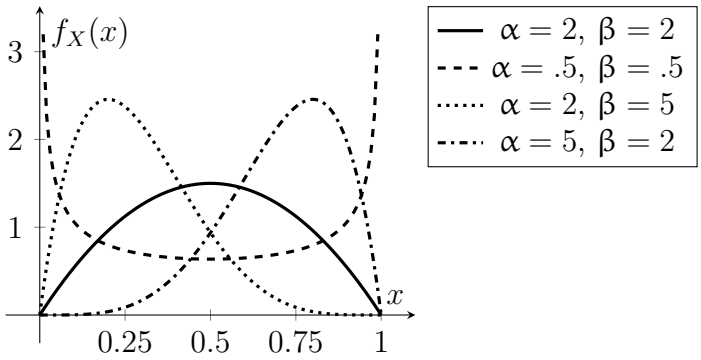


Figure 2.7: Different Beta probability densities

Figure 2.7 displays some examples of the standard Beta distribution that are different from the uniform case. The shapes assumed by the different density functions in the figure are wildly different, and more can be obtained with different configurations of the parameters. It is easy to calculate the moment of this distribution directly. Observe that, because of (2.72):

$$\mathbb{E}[X^r] = \frac{1}{\mathrm{B}(\alpha,\beta)} \int_0^1 x^{r+\alpha-1}(1-x)^{\beta-1}\,dx = \frac{\mathrm{B}(r+\alpha,\beta)}{\mathrm{B}(\alpha,\beta)}$$

which can be alternatively expressed in terms of the Gamma function.

$$\mathbb{E}[X^r] = \frac{\Gamma(r+\alpha)\,\Gamma(\alpha+\beta)}{\Gamma(r+\alpha+\beta)\,\Gamma(\alpha)}$$

Thanks to the property of the Gamma function that $\Gamma(\gamma+1) = \gamma\Gamma(\gamma)$, one can show that:

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta} \tag{2.75}$$

$$\mathbb{V}\text{ar}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{2.76}$$

and that the distribution is asymmetric if $\alpha \neq \beta$. This is actually an easier way to proceed for calculating the moments, since the moment generating function of the standard Beta distribution is particularly involved.

$$M_X(t;\alpha,\beta) = 1 + \sum_{q}^{\infty}\left(\prod_{k=0}^{q-1}\frac{\alpha+k}{\alpha+\beta+k}\right)\frac{t^q}{q!} \tag{2.77}$$

The standard family of Beta distributions can be easily generalized to different segments in $\mathbb{R}$. The *nonstandard* family of Beta distributions has support $\mathbb{X} = (a,b)$ – or in many cases, $\mathbb{X} = [0,1]$ – and the following density function.

$$f_X(x;\alpha,\beta,a,b) = \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)\cdot(b-a)^{\alpha+\beta-1}} \quad \text{for } x \in (a,b) \tag{2.78}$$

The analysis of this distribution proceeds similarly as in the standard case; when $\alpha = \beta = 1$, a nonstandard Beta distribution coincides with a uniform distribution on the (a,b) interval. The Beta distribution is a useful one for modeling the probability of events that are defined on a bounded interval of the real line but are not uniform. Thanks to its relationships with other distributions, the Beta distribution has a number of other applications in statistical inference, one of which is mentioned later in Lecture 5.

Exponential distribution

Like the uniform distribution, even the *exponential* distribution has already been introduced in Lecture 1 through a special case, defined as that with "unit parameter." The larger family of exponential distributions takes its support on the set of nonnegative real numbers $\mathbb{X} = \mathbb{R}_+$, and it allows for different values of the parameter $\lambda \in \mathbb{R}_{++}$ (where $\lambda = 1$ is clearly the "unit parameter" case). The probability density function reads generally as:

$$f_X(x;\lambda) = \frac{1}{\lambda}\exp\left(-\frac{x}{\lambda}\right) \quad \text{for } x \geq 0 \tag{2.79}$$

while the cumulative distribution function is:

$$F_X(x;\lambda) = 1 - \exp\left(-\frac{x}{\lambda}\right) \tag{2.80}$$

which obviously integrates to one. The functional form of these distributions is regular and simple, as shown below in Figure 2.8.

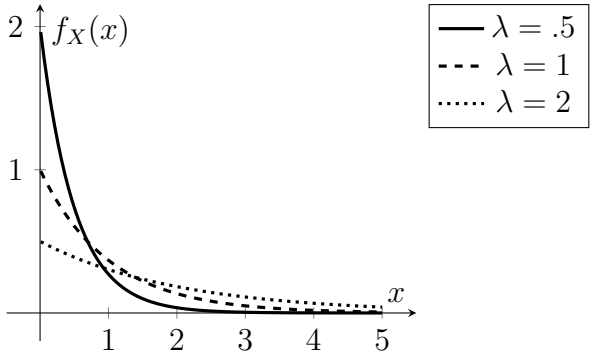


Figure 2.8: Different exponential probability densities

By extending the examples from Lecture 1, it is straightforward to calculate the moment generating function (for $t \leq \lambda^{-1}$) as:

$$M_X(t;\lambda) = \frac{1}{1-\lambda t} \tag{2.81}$$

and the mean and variance as:

$$\mathbb{E}[X] = \lambda \tag{2.82}$$

$$\mathbb{V}\text{ar}[X] = \lambda^2 \tag{2.83}$$

which loads λ with interpretation. The exponential distribution is typically utilized to model the "waiting time" before the occurrence of some particular event, if time is measured as a positive real number $X > 0$. In this sense, the exponential distribution is a continuous analogue of the discrete geometric distribution. Therefore, the parameter λ is both a measure of the average waiting time, and of its variation. Even the exponential distribution features the *memoryless* property (2.16), that is for any two integers s, t with $s > t$:

$$\mathbb{P}(X > s | X > t) = \mathbb{P}(X > s - t)$$

which is easy to show by following the same procedure as in the case of the geometric distribution. Intuitively, this means that "waiting times" do not depend upon the passing of time, and that the expectations are continuously reset so long as the event of interest does not occur.

If a random variable X follows the exponential distribution λ, this is usually written as follows.

$$X \sim \text{Exp}(\lambda)$$

It is simple to verify that if a random variable $Y = K \cdot X$ is obtained by rescaling an exponentially distributed random variable X by some constant K, Y is also exponentially distributed with a rescaled parameter, as follows.

$$Y \sim \text{Exp}(K\lambda)$$

The exponential distribution has numerous relationships with other distributions. The ones that relate it to distributions that were already discussed in this Lecture are summarized next.

Observation 2.4. If $X \sim \mathcal{U}(0,1)$ and $Y = -\lambda \log(X)$ it is $Y \sim \text{Exp}(\lambda)$.

Observation 2.5. If $X \sim \text{Exp}(\lambda)$ and $Y = \exp(-X)$ it is $Y \sim \text{Beta}\left(\frac{1}{\lambda}, 1\right)$.

Observation 2.6. As it was already anticipated, if $X \sim \text{Laplace}(\mu, \sigma)$ and $Y = |X - \mu|$ it is $Y \sim \text{Exp}(\sigma)$.

Observation 2.7. If $X \sim \text{Exp}(1)$ and

$$Y = \mu - \sigma \log\left(\frac{\exp(-X)}{1 - \exp(-X)}\right)$$

it is $Y \sim \text{Logistic}(\mu, \sigma)$. The interpretation of this result is as follows: the logistic distribution can model a linear function of the logarithm of an *odds ratio* between two probabilities – that is, the probabilities that the waiting time for some event which can be modeled by the exponential distribution with unit parameter is either longer or shorter than some given number X.

Important relationships between the exponential distribution and other distributions not discussed yet are illustrated as these are introduced.

Gamma distribution

The *Gamma* family is central in the taxonomy of continuous distributions, since it relates directly or indirectly to many other such families. Its support equates the set of positive real numbers: $\mathbb{X} = \mathbb{R}_{++}$; like the Beta family, it is identified by two positive parameters $\alpha \in \mathbb{R}_{++}$ and $\beta \in \mathbb{R}_{++}$ (but several different parametrizations are possible, here the focus is on a specific one). The name of this distribution derives from the fact that its density function

features the Gamma function; according to the parametrization adopted in these lectures, the density function reads as follows.

$$f_X(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)}\beta^{\alpha}x^{\alpha-1}\exp(-\beta x) \quad \text{for } x>0 \tag{2.84}$$

The cumulative distribution lacks a closed form expression, but similarly as in the Beta case, it is often expressed in terms of an ancillary function called **lower incomplete Gamma function**: $\gamma(a,b) \equiv \int_0^b u^{a-1}\exp(-u)\,du$.

$$F_X(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)}\int_0^x \beta^{\alpha}t^{\alpha-1}\exp(-\beta t)\,dt = \frac{\gamma(\alpha,\beta x)}{\Gamma(\alpha)} \tag{2.85}$$

A property, which is easy to show, of the lower incomplete Gamma function is that $\lim_{b\to\infty}\gamma(a,b) = \Gamma(a)$; clearly, all Gamma distributions integrate to 1. If a random variable X follows a Gamma distribution with parameters α and β, this is generally written in one of two ways.

$$X \sim \Gamma(\alpha,\beta)$$
$$X \sim \text{Gamma}(\alpha,\beta)$$

Observation 2.8. $X \sim \text{Gamma}\left(1,\frac{1}{\lambda}\right)$ is equivalent to $X \sim \text{Exp}(\lambda)$, that is, exponential distributions are all special cases of the Gamma family.

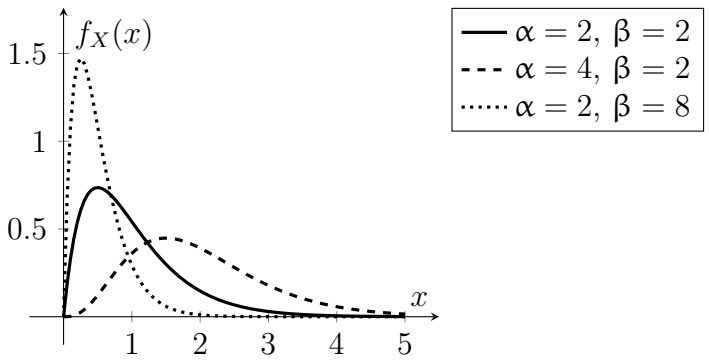


Figure 2.9: Different Gamma probability densities

Figure 2.9 displays different Gamma density functions for $\alpha > 1$, they all display the asymmetric humped shape that usually characterizes these distributions. Similarly as with the Beta distributions, the direct calculation of the Gamma centered moments is easy. By inspecting (2.84), it is:

$$\mathbb{E}[X^r] = \frac{1}{\Gamma(\alpha)}\frac{1}{\beta^r}\int_0^{\infty}\beta^{r+\alpha}x^{r+\alpha-1}\exp(-\beta x)\,dx = \frac{\Gamma(r+\alpha)}{\Gamma(\alpha)\,\beta^r}$$

since the expression inside the integral, divided by $\Gamma(r+\alpha)$, is the density function of yet another Gamma distribution with parameters $r+\alpha$ and β. All the moments are thus easily obtained thanks to the Gamma function's property that $\Gamma(\gamma+1) = \gamma\Gamma(\gamma)$; for example, the mean and the variance are expressed as follows.

$$\mathbb{E}[X] = \frac{\alpha}{\beta} \tag{2.86}$$

$$\mathbb{V}\mathrm{ar}[X] = \frac{\alpha}{\beta^2} \tag{2.87}$$

Alternatively, one could have used the moment generating function, which is calculated in analogy with the centered moments.

$$\begin{aligned} M_X(t;\alpha,\beta) &= \frac{1}{\Gamma(\alpha)} \int_0^\infty \exp(tx)\, \beta^\alpha x^{\alpha-1} \exp(-\beta x)\, dx \\ &= \frac{\beta^\alpha}{(\beta-t)^\alpha} \frac{1}{\Gamma(\alpha)} \int_0^\infty (\beta-t)^\alpha\, x^{\alpha-1} \exp(-(\beta-t)\, x)\, dx \\ &= \left(\frac{\beta}{\beta-t}\right)^\alpha \\ &= \left(1-\frac{t}{\beta}\right)^{-\alpha} \end{aligned} \tag{2.88}$$

Note that the integral in the second line is easily related to a density function of a Gamma function with parameters α and $\beta - t > 0$, which again helps simplify the accounts. This implies that the moment generating function is only defined for $t < \beta$.

The Gamma distribution has numerous applications in several branches of science; but its direct applications in the social sciences are quite scant. As mentioned, the main importance of this distribution lies in its relationship with other distributions.

Chi-squared distribution

The family of *chi-squared* distributions is central in the theory of statistical inference. It has its support on the positive real numbers $\mathbb{X} = \mathbb{R}_{++}$, a single positive parameter $\kappa \in \mathbb{R}_{++}$, and its density function is as follows.

$$f_X(x;\kappa) = \frac{1}{\Gamma\left(\frac{\kappa}{2}\right) \cdot 2^{\frac{\kappa}{2}}}\, x^{\frac{\kappa}{2}-1} \exp\left(-\frac{x}{2}\right) \quad \text{for } x > 0 \tag{2.89}$$

If a random variable X follows the chi-squared distribution, this is written in the following way, which might slightly differ in the details.

$$X \sim \chi^2(\kappa) \quad \text{or} \quad X \sim \chi^2_\kappa$$

Notably, when κ is an integer it is called the *number of degrees of freedom* of the chi-squared distribution, for reasons that relate to the interpretation of the latter as the distribution of a sample variance obtained from normally distributed, independent random variables (see Lecture 5).[4]

Observation 2.9. $X \sim \text{Gamma}\left(\frac{\kappa}{2}, \frac{1}{2}\right)$ is equivalent to $X \sim \chi^2(\kappa)$, that is, chi-squared distributions are all special cases of the Gamma family.

Observation 2.10. $X \sim \chi^2(2)$ is equivalent to $X \sim \text{Exp}\left(\frac{1}{2}\right)$.

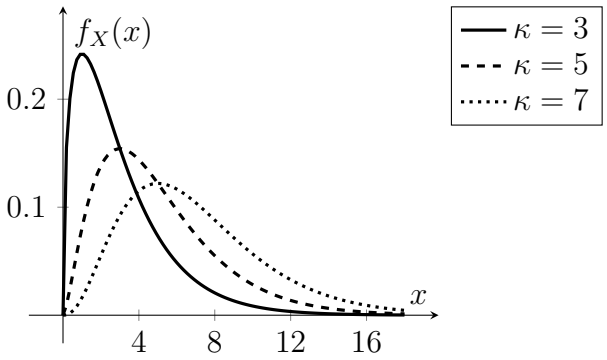


Figure 2.10: Different chi-squared probability densities

The two observations above clarify that the chi-squared distributions are a subfamily of the Gamma family, and are consequently related to the exponential distribution as well. Unsurprisingly, the chi-squared distributions typically have humped shapes similar to the Gamma ones, as displayed in Figure 2.10. For low values of κ however, as for Gamma distributions with specific combinations of α and β, this is not the case; recall for example Figure 1.9 which shows that $f_X(x, \kappa = 1)$ approaches the y-axis asymptotically as $x \to 0$, and lacks a maximum.[5] Since every chi-squared distribution is a particular Gamma distribution, the analysis of the former follows that of the latter. For example, the mean and variance are as follows.

$$\mathbb{E}[X] = \kappa \tag{2.90}$$

$$\mathbb{V}\text{ar}[X] = 2\kappa \tag{2.91}$$

The moment generating function, defined for $t < 0.5$, is instead given below.

$$M_X(t; \kappa) = (1 - 2t)^{-\frac{\kappa}{2}} \tag{2.92}$$

[4] In this case one says that some random variable X follows the chi-squared distribution "with κ degrees of freedom."

[5] The chi-squared density function with one degree of freedom given in Example 1.12 is reconciled with (2.89) by noting that $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

The most important property of the chi-squared distribution is certainly its relationship with the standard normal distribution, as already introduced in Example 1.12. This is reiterated here.

Observation 2.11. If $X \sim \mathcal{N}(0,1)$ and $Y = X^2$, it is $Y \sim \chi^2(1)$.

This fact plays a fundamental role in statistical inference, as elaborated in Lecture 5 and – for the asymptotic case – Lecture 6.

Snedecor's *F*-distribution

The distribution named after Ronald Fisher and George W. Snedecor – for brevity, Snedecor's F-distribution or just the F-distribution, is yet another quite involved family of distributions with support restricted to the positive set of real numbers, $\mathbb{X} = \mathbb{R}_{++}$ which is defined by two positive parameters $(\nu_1, \nu_2) \in \mathbb{R}^2_{++}$; its density is given as:

$$f_X(x; \nu_1, \nu_2) = \frac{1}{\mathrm{B}\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2}-1} \left(1 + \frac{\nu_1}{\nu_2} x\right)^{-\frac{\nu_1+\nu_2}{2}} \quad \text{for} \quad x > 0 \tag{2.93}$$

where $\mathrm{B}\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)$ is a Beta function of the (halved) parameters. Its cumulative distribution can be expressed in a compact way by using the previously introduced incomplete Beta function:

$$F_X(x; \nu_1, \nu_2) = \frac{\mathrm{B}\left(x, \frac{\nu_1}{2}, \frac{\nu_2}{2}\right)}{\mathrm{B}\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \tag{2.94}$$

showing that the distribution integrates to 1. Some random variable X that follows the Fisher-Snedecor distribution is generally indicated with one of two slightly different notations.

$$X \sim \mathcal{F}(\nu_1, \nu_2) \quad \text{or} \quad X \sim \mathcal{F}_{\nu_1, \nu_2}$$

Observation 2.12. If $X \sim \mathcal{F}(\nu_1, \nu_2)$ and $Y = X^{-1}$, it is $Y \sim \mathcal{F}(\nu_2, \nu_1)$.

Observation 2.13. If $X \sim \mathcal{F}(\nu_1, \nu_2)$ and $Y \sim \text{Beta}\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)$, the random variables X and Y are related by the following reciprocal transformations: $Y = (\nu_1 X/\nu_2) / (1 - \nu_1 X/\nu_2)$ and $X = \nu_2 Y/\nu_1 (1 - Y)$.

Selected density functions conforming to (2.93) are displayed in Figure (2.11) below. For values of ν_1 and ν_2 around or smaller than 2, the shape of the F-distribution resembles that of the chi-squared for values of κ around or smaller than 1, respectively. In a similar vein, values of the parameters

larger than 2 lead to a typical hump-shaped density. Like in the lognormal and Cauchy cases, also the F-distribution lacks a moment generating function (and even its characteristic function is quite involved). Fortunately, its moments can be calculated by direct integration as in the Beta and Gamma cases. This allows to derive the mean and variance as:

$$\mathbb{E}[X] = \frac{\nu_2}{\nu_2 - 2} \tag{2.95}$$

$$\mathbb{V}\mathrm{ar}[X] = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)} \tag{2.96}$$

but these two moments are only defined for $\nu_2 > 2$ and $\nu_2 > 4$ respectively (or else the integral that defines them diverges).

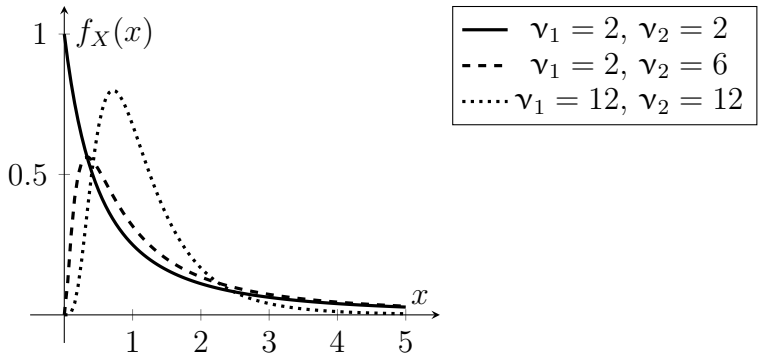


Figure 2.11: Different F probability densities

The Fisher-Snedecor distribution has multiple relationships with other distributions; its various links to the chi-squared distribution are especially relevant and are discussed at length in later lectures; the importance of these links is due to the F-distribution's role in statistical inference about samples drawn from the normal distribution. In those contexts, the parameters ν_1 and ν_2 are integers, and are also referred to – as in the chi-squared case – as two *degrees of freedom*.[6]

Student's t-distribution

The distribution named after the pseudonym of William Sealy Gosset, that is *Student* (while 't' derives from the distribution's use in statistical *tests*), is the only one analyzed in this section which has support on the entire set of

[6] In such settings, one would say that X follows the F-distribution "with ν_1 and ν_2 degrees of freedom."

real numbers, $\mathbb{X} = \mathbb{R}$; it is symmetric and defined by one positive parameter $\nu \in \mathbb{R}_{++}$. Its density function can be expressed in two alternative ways, with the aid of either the Gamma or the Beta function. In the former case, one has:

$$f_X(x;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\pi\nu}} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{2.97}$$

while in the latter it is as follows.

$$f_X(x;\nu) = \frac{1}{\mathrm{B}\left(\frac{1}{2}, \frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu}} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{2.98}$$

Its cumulative distribution can be expressed through the incomplete Beta function:

$$F_X(x;\nu) = \begin{cases} \frac{1}{2}\mathrm{B}\left(\frac{\nu}{x^2+\nu}, \frac{1}{2}, \frac{\nu}{2}\right) & \text{if } x \leq 0 \\ 1 - \frac{1}{2}\mathrm{B}\left(\frac{\nu}{x^2+\nu}, \frac{1}{2}, \frac{\nu}{2}\right) & \text{if } x > 0 \end{cases} \tag{2.99}$$

this shows that the distribution must integrate to 1 over the real line. A random variable X that follows the Student's t-distribution can be indicated in two similar ways.

$$X \sim \mathcal{T}(\nu) \quad \text{or} \quad X \sim \mathcal{T}_\nu$$

Observation 2.14. $X \sim \mathcal{T}(1)$ is equivalent to $X \sim \text{Cauchy}(0, 1)$.

The above observation states that for $\nu = 1$, the Student's t is the standard Cauchy distribution. For moderate values of ν, however, the Student's t appears almost identical to the standard normal distribution, up to the point that the two distributions "coincide" in an asymptotic sense as shown later in Lecture 6. For small values of ν that yet are larger than 1 however, the Student's t-distribution can be thought as a sort of "middle ground" between the standard Cauchy and the standard Normal distributions. This is illustrated in Figure 2.12, which displays the density function of the latter two distributions along with the density of the Student's t for $\nu = 3$.

It is easy to guess that because of the similarity with the Cauchy, the moments of the Student's t-distribution may not always be defined. In fact, the distribution lacks a moment generating function and its characteristic function is also hard to manipulate, while given a value of the parameter ν, the moments of order $r \geq \nu$ do not exist. Yet one can show that:

$$\mathbb{E}[X^r] = \begin{cases} \dfrac{\Gamma\left(\dfrac{r+1}{2}\right)\Gamma\left(\dfrac{\nu-r}{2}\right)}{\Gamma\left(\dfrac{\nu}{2}\right)} \cdot \sqrt{\dfrac{\nu^r}{\pi}} & \text{if } r \text{ is even and } 0 < r < \nu \\ 0 & \text{if } r \text{ is odd and } 0 < r < \nu \end{cases}$$

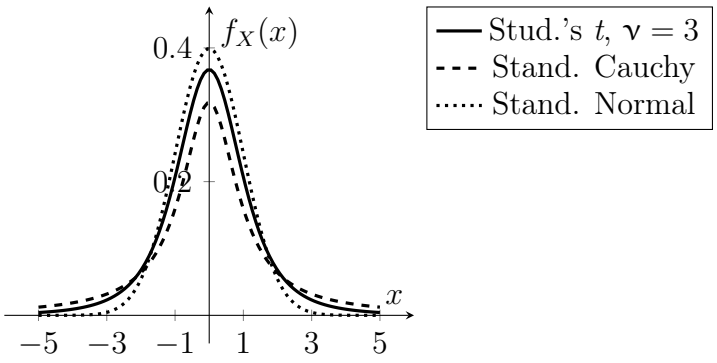


Figure 2.12: Student's t, standard Cauchy, and standard normal densities

which implies that, for $\nu > 1$, the mean is defined as:

$$\mathbb{E}[X] = 0 \tag{2.100}$$

for $\nu > 2$, the variance is defined as:

$$\mathbb{V}\mathrm{ar}[X] = \frac{\nu}{\nu - 2} \tag{2.101}$$

for $\nu > 3$, the skewness is defined as:

$$\mathbb{S}\mathrm{kew}[X] = 0 \tag{2.102}$$

for $\nu > 4$, the kurtosis is defined as:

$$\mathbb{K}\mathrm{urt}[X] = \frac{3\nu - 6}{\nu - 4} \tag{2.103}$$

and so on and so forth with the higher moments.

The Student's t-distribution is another important distribution in the theory of statistical inference, as it is used to model the distribution of a sample mean drawn from normally distributed random variables. In these contexts, the parameter ν is an integer and takes the not-too-original name of *number of degrees of freedom*.[7] The Student's t-distribution is also related to the Snedecor's F-distribution through some non-monotone transformations, as expressed in the following two observations.

Observation 2.15. If $X \sim \mathcal{T}(\nu)$ and $Y = X^2$, it is $Y \sim \mathcal{F}(1, \nu)$.

Observation 2.16. If $X \sim \mathcal{T}(\nu)$ and $Y = X^{-2}$, it is $Y \sim \mathcal{F}(\nu, 1)$.

Even these transformations play a role in statistical inference and are worth to keep in mind.

[7] Similarly to the chi-squared and F cases, in these contexts one would thus say that X follows the Student's t-distribution "with ν degrees of freedom."

Pareto distribution

This section is concluded with the analysis of the distribution named after Vilfredo Pareto, a famous distribution with support defined on a subset of the set of positive real numbers, $\mathbb{X} = [\alpha, \infty)$. Here, $\alpha \in \mathbb{R}_{++}$ is a parameter of the family of Pareto distributions; this role is shared with another positive parameter $\beta \in \mathbb{R}_{++}$. Given two such parameters, the density function of a particular Pareto distribution is:

$$f_X(x; \alpha, \beta) = \frac{\beta \alpha^\beta}{x^{\beta+1}} \quad \text{for } x \geq \alpha \tag{2.104}$$

and the cumulative distribution is:

$$F_X(x; \alpha, \beta) = 1 - \left(\frac{\alpha}{x}\right)^\beta \quad \text{for } x \geq \alpha \tag{2.105}$$

and zero otherwise ($x < \alpha$). Its cumulative distribution clearly tends to 1 as X tends to infinity. The shape of the distribution is displayed in Figure 2.13 below for a fixed value $\alpha = 1$ and three different values of β. Clearly, the parameter β affects the *shape* of the distribution, with lower values making the distribution flatter, and vice versa (similarly as, but contrarily to, λ in the exponential distribution's case). Instead, parameter α affects both the *location* of the distribution (as it defines the support) and the overall *scale*. A random variable X distributed according to some Pareto distribution is denoted as follows.

$$X \sim \text{Pareto}(\alpha, \beta)$$

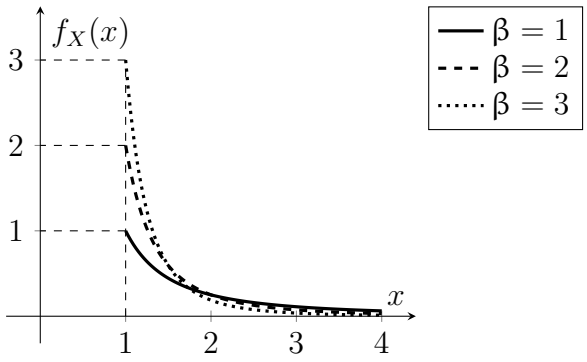


Figure 2.13: Different Pareto probability densities, $\alpha = 1$

Observation 2.17. If $X \sim \text{Pareto}(\alpha, \beta)$ and $Y \sim \text{Exp}(\beta^{-1})$, the two random variables are related through the two symmetric transformations $X = \alpha \exp(Y)$ and $Y = \log(X/\alpha)$.

The moment generating function of the Pareto distribution is:

$$M_X(x;\alpha,\beta) = \int_\alpha^\infty \exp(tx)\cdot\frac{\beta\alpha^\beta}{x^{\beta+1}}\,dx = \beta(-\alpha t)^\beta\cdot\Gamma(-\beta,-\alpha t) \quad (2.106)$$

where $\Gamma(a,b) = \int_b^\infty u^{a-1}\exp(-u)\,du$ is the so-called **upper incomplete Gamma function**, another ancillary function; the equivalence between the two expressions in (2.106) is obtained through the simple change of variable $u = -tx$ and since $\Gamma(a,b)$ is only defined for $b > 0$, the moment generating functions of Pareto distributions are only defined for $t < 0$. To calculate the moments of this distribution, it is easiest to proceed by direct integration of the density, which results in finite solutions only for restricted values of the "shape" parameter β. Specifically, the mean is such that:

$$\mathbb{E}[X] = \begin{cases} \infty & \text{for } \beta \leq 1 \\ \dfrac{\alpha\beta}{\beta-1} & \text{for } \beta > 1 \end{cases} \quad (2.107)$$

while the variance is as follows.

$$\mathbb{V}\text{ar}[X] = \begin{cases} \infty & \text{for } \beta \leq 2 \\ \dfrac{\alpha^2\beta}{(\beta-1)^2(\beta-2)} & \text{for } \beta > 2 \end{cases} \quad (2.108)$$

Intuitively, the Pareto distribution should not be too flat, or else its right tail becomes "too heavy" causing the relevant moments to diverge. This is a well known property of the Pareto distribution.

Like the lognormal and the Gamma distributions, the Pareto distribution is typically used to model asymmetric phenomena that are associated with positive real numbers; it may be more motivated than its competing alternatives when it is necessary to study phenomena that display "fat tails" (i.e. notable inequality), such as the distribution of wealth or that of cities' size. Another attractive feature of the Pareto distribution is that the density function (2.104) satisfies a mathematical **power law**, which is easy to describe as a linear function upon applying a logarithmic transformation.

$$\log f_X(x;\alpha,\beta) = \log\left(\beta\alpha^\beta\right) - (\beta+1)\log x \quad \text{for } x \geq \alpha$$

Furthermore, the power law implies that the cumulative distribution (2.105) can be easily inverted, resulting in a quantile function of conveniently simple manipulation.

$$Q_X(p;\alpha,\beta) = \alpha(1-p)^{-\frac{1}{\beta}}$$

It must be mentioned that the family of distributions discussed so far is just a particular case of a more general family called **generalized Pareto distribution**, whose cumulative distribution function reads:

$$F_X(x;\beta,\gamma,\mu,\sigma) = 1 - \left[1 + \left(\frac{x-\mu}{\sigma}\right)^{\frac{1}{\gamma}}\right]^{-\beta} \quad \text{for } x \geq \mu \tag{2.109}$$

with parameters $\mu \in \mathbb{R}$, $(\beta,\gamma,\sigma) \in \mathbb{R}^3_{++}$ and support $\mathbb{X} = [\mu,\infty)$. In this context, the particular subfamily defined by (2.104) and (2.105) is referred to as "Type I" Pareto distribution.

2.4 Continuous Distributions III

This lecture is completed with the separate treatment of a specific wider family of continuous distributions, all encompassed under the umbrella of the **Generalized Extreme Value** (**GEV**) distribution. As the name suggests, these distributions are especially well suited to model the probability about "extreme" realizations of events, which are defined (depending on the context) as the maxima or the minima of a certain sequence of realizations. The understanding about this specific interpretation of GEV distributions is better developed through a result of asymptotic theory known under various names, including that of "Extreme Value Theorem," and discussed in Lecture 6. The analysis in this section is limited to that of the mathematical properties of the GEV distribution, and their mutual relationships.

A GEV distribution depends on three parameters: a **location** parameter $\mu \in \mathbb{R}$, a **scale** parameter $\sigma \in \mathbb{R}_{++}$, and a so-called **shape** parameter $\xi \in \mathbb{R}$. The support of a GEV distribution depends on the value of ξ. In particular, if $\xi > 0$, it is:

$$\mathbb{X}_{\xi>0} = \left[\mu - \frac{\sigma}{\xi}, \infty\right) \tag{2.110}$$

if $\xi < 0$, it is:

$$\mathbb{X}_{\xi<0} = \left(-\infty, \mu - \frac{\sigma}{\xi}\right] \tag{2.111}$$

while if $\xi = 0$, the support $\mathbb{X}_{\xi=0} = \mathbb{R}$ is more simply equal to the entire set of real numbers. The general expression of the density function of the GEV distribution is best defined through the *standardized value* $Z = (X-\mu)/\sigma$; as a function of ξ, the density function of Z is:

$$f_Z(z;\xi) = \begin{cases} (1+\xi z)^{-\frac{1}{\xi}-1} \exp\left(-(1+\xi z)^{-\frac{1}{\xi}}\right) & \text{for } \xi \neq 0 \text{ and } \xi z > -1 \\ \exp(-z)\exp(-\exp(-z)) & \text{for } \xi = 0 \end{cases} \tag{2.112}$$

and the corresponding cumulative distribution is:

$$F_Z(z;\xi) = \begin{cases} \exp\left(-(1+\xi z)^{-\frac{1}{\xi}}\right) & \text{for } \xi \neq 0 \text{ and } \xi z > -1 \\ \exp(-\exp(-z)) & \text{for } \xi = 0 \end{cases} \tag{2.113}$$

and one can easily verify that in all cases, the distribution integrates to 1. Observe that for both the density function and the cumulative distribution, the expression for $\xi = 0$ corresponds to the limit case of the expression for $\xi \neq 0$. An important property of GEV distributions is that (2.113) is easy to invert in all cases, so that the quantile function can be written as follows.

$$Q_Z(p;\xi) = \begin{cases} \dfrac{(-\log(p))^{-\xi} - 1}{\xi} & \text{for } \xi > 0,\ p \in [0,1);\ \text{or } \xi < 0,\ p \in (0,1] \\ -\log(-\log(p)) & \text{for } \xi = 0 \text{ and } p \in (0,1) \end{cases} \tag{2.114}$$

Note that the restrictions in the domain of the quantile function correspond to the restrictions on the support of X.

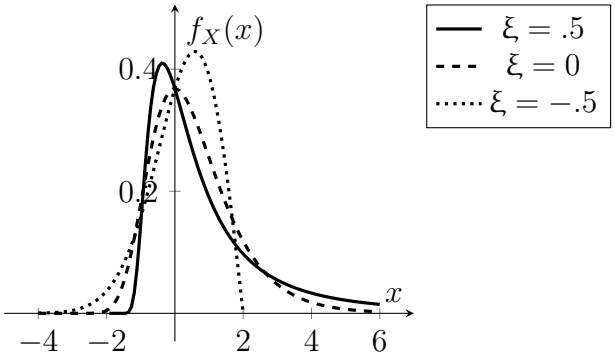


Figure 2.14: Different GEV probability densities with $\mu = 0$ and $\sigma = 1$

Some examples of GEV density functions, for constant parameters $\mu = 0$ and $\sigma = 1$ but different values of ξ, are displayed in Figure 2.14 above. All these examples of distribution are clearly symmetric; also observe that for $\xi \neq 0$, the support of the distribution is bounded accordingly. In order to denote that a random variable X follows a GEV distribution with specified parameters μ, σ and ξ, one usually writes as follows.

$$X \sim \text{GEV}(\mu, \sigma, \xi)$$

A general expression for the moment generating (or characteristic) function of GEV distributions is difficult to derive, but the moments of interest can

be obtained by direct integration; their definition (or lack thereof) depends again the value of the parameter ξ. The mean, for example, is given as:

$$\mathbb{E}[X] = \begin{cases} \mu + \dfrac{\sigma}{\xi}[\Gamma(1-\xi) - 1] & \text{if } \xi \neq 0, \xi < 1 \\ \mu + \sigma\gamma & \text{if } \xi = 0 \\ \infty & \text{if } \xi \geq 1 \end{cases} \tag{2.115}$$

where $\Gamma(\cdot)$ is the Gamma function and $\gamma \simeq 0.577$ is the Euler-Mascheroni constant, while the variance is as follows.

$$\mathbb{V}\text{ar}[X] = \begin{cases} \dfrac{\sigma^2}{\xi^2}\left[\Gamma(1-2\xi) - (\Gamma(1-\xi))^2\right] & \text{if } \xi \neq 0, \xi < \frac{1}{2} \\ \sigma^2\dfrac{\pi^2}{6} & \text{if } \xi = 0 \\ \infty & \text{if } \xi \geq \frac{1}{2} \end{cases} \tag{2.116}$$

The remainder of this section (and of this lecture alike) discusses in more detail three particular cases of GEV distributions. These are of particular interest for economists and econometricians, as they feature prominently in both theoretical economic models with a stochastic component and in the closely related structural econometric models. These restricted subfamilies of the larger GEV family are typically named according to their discoverers, but are also distinguished by a number from a classification between *types*.

Type I GEV distribution ($\xi = 0$): Gumbel

The GEV subfamily restricted to $\xi = 0$ is the simplest one, and its members are said to be following the *Gumbel* or *Type I GEV* distribution. Its support $\mathbb{X} = \mathbb{R}$ is the entire set of real numbers and it is defined only in terms of the location and scale parameters μ and σ respectively. In the non-standardized cases, the Gumbel density function is given by:

$$f_X(x; \mu, \sigma) = \frac{1}{\sigma}\exp\left(-\frac{x-\mu}{\sigma}\right)\exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right) \tag{2.117}$$

its cumulative distribution is:

$$F_X(x; \mu, \sigma) = \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right) \tag{2.118}$$

and its quantile function is simply as follows, for $p \in (0, 1)$.

$$Q_X(p; \mu, \sigma) = \mu - \sigma\log(-\log(p)) \tag{2.119}$$

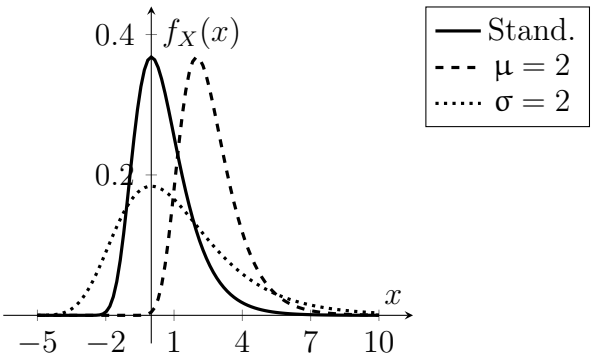


Figure 2.15: Different Gumbel probability densities

Figure 2.15 shows the "standardized" version of the Gumbel density function along with versions for different values of μ and σ; their respective roles as location and scale parameters are well evident. The mean and the variance of the Gumbel distribution are given as the central cases for $\xi = 0$ in (2.115) and (2.116) respectively.[8] If some random variable X follows the Gumbel distribution, this is indicated in one of two ways:

$$X \sim \text{Gumbel}\,(\mu, \sigma)$$

or

$$X \sim \text{EV1}\,(\mu, \sigma)$$

where "EV1" stands for "Extreme Value (Type) 1." For reasons to be clarified in later lectures, the Gumbel distribution is typically used to model the *maximum* value among many different realizations, and it is the backbone of several "structural" econometric techniques used to model the behavior of socio-economic variables of interest that take values on a countable set.

Type II GEV distribution ($\xi > 0$): Fréchet

Restricting the wider class of GEV distributions to $\xi > 0$ defines the subfamily of so-called *Fréchet* or *Type II GEV* distributions. A Fréchet distribution has bounded support as per (2.110), and its defining equations have been given earlier for the standardized ($\mu = 0$ and $\sigma = 1$) case. Often, the distribution is reparametrized through the inverse of the shape parameter, $\alpha = \xi^{-1}$, and the following transformation is applied.

$$Y = \sigma + \mu\,(1 - \xi) + \xi X \tag{2.120}$$

[8] The Gumbel distribution also features a relatively simple expression for the moment generating function, that is $M_X\,(t; \mu, \sigma) = \exp\,(\mu t)\,\Gamma\,(1 - \sigma t)$.

In such a case, the support is $\mathbb{Y} = [\mu, \infty)$, the density function reads as:

$$f_Y(y; \alpha, \mu, \sigma) = \frac{\alpha}{\sigma}\left(\frac{y-\mu}{\sigma}\right)^{-\alpha-1} \exp\left(-\left(\frac{y-\mu}{\sigma}\right)^{-\alpha}\right) \quad \text{for} \quad y > \mu \tag{2.121}$$

the cumulative distribution function as:

$$F_Y(y; \alpha, \mu, \sigma) = \exp\left(-\left(\frac{y-\mu}{\sigma}\right)^{-\alpha}\right) \quad \text{for} \quad y > \mu \tag{2.122}$$

and the quantile function as follows.

$$Q_Y(p; \alpha, \mu, \sigma) = (-\log(p))^{\frac{1}{\alpha}} \quad \text{for} \quad p \in [0, 1) \tag{2.123}$$

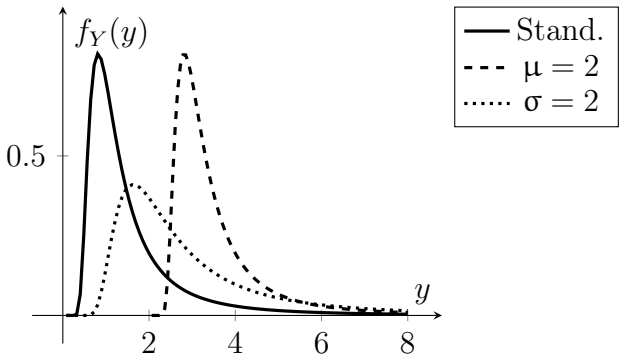


Figure 2.16: Different Fréchet probability densities with $\alpha = 2$

Three different Fréchet density for a fixed value of the shape parameter are shown above in Figure 2.16, again highlighting the role of the location and scale parameters; observe how the former (μ) also affects the bound of the distribution's support. The mean and the variance of the transformed random variable, if they are finite, are obtained by applying the standard properties of simple moments to (2.115) and (2.116) respectively. If it follows the Fréchet distribution, a random variable Y is indicated either as:

$$Y \sim \text{Frechet}(\alpha, \mu, \sigma)$$

or as:

$$Y \sim \text{EV2}(\alpha, \mu, \sigma)$$

where the latter notation now refers to "Type II." The most frequent use of the Fréchet distribution is similar to the Gumbel's, that is, for modeling the *maximum* value of many realizations. Furthermore, the Fréchet distribution features prominently in several economic and econometric models, especially in the field of international trade.

Type III GEV distribution ($\xi < 0$): (reverse) Weibull

The last subfamily of GEV distributions is defined for $\xi < 0$, and results in the *reverse Weibull* or *Type III GEV* distribution. Once more, the support of a reverse Weibull distribution is bounded as per (2.111) and its defining functions follow from the general GEV case. This distribution is typically reparametrized with $\alpha = -\xi^{-1}$; with the transformation (2.120) the analysis proceeds as in the Fréchet case (e.g. the support becomes $\mathbb{Y} = (-\infty, \mu]$). With the alternative transformation instead:

$$W = -\left[\sigma + \mu\left(1 - \xi\right) + \xi X\right] \tag{2.124}$$

one obtains the (traditional) *Weibull* distribution, which predates the theory of GEV distributions (thus explaining the name *reverse* Weibull for the Type III GEV case, since $W = -Y$). The traditional Weibull distribution has support $\mathbb{W} = [\mu, \infty)$; its density function is:

$$f_W\left(w; \alpha, \mu, \sigma\right) = \frac{\alpha}{\sigma}\left(\frac{w-\mu}{\sigma}\right)^{\alpha-1} \exp\left(-\left(\frac{w-\mu}{\sigma}\right)^{\alpha}\right) \quad \text{for} \quad w > \mu \tag{2.125}$$

its cumulative distribution function as:

$$F_W\left(w; \alpha, \mu, \sigma\right) = 1 - \exp\left(-\left(\frac{w-\mu}{\sigma}\right)^{\alpha}\right) \quad \text{for} \quad w > \mu \tag{2.126}$$

and its quantile function is as follows.

$$Q_W\left(p; \alpha, \mu, \sigma\right) = \left(-\log\left(1-p\right)\right)^{\frac{1}{\alpha}} \quad \text{for} \quad p \in [0, 1) \tag{2.127}$$

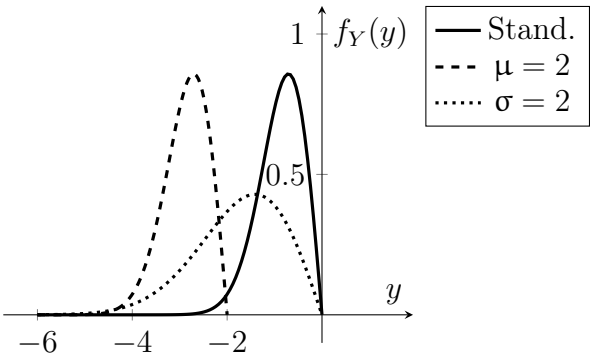


Figure 2.17: Different reverse Weibull probability densities with $\alpha = 2$

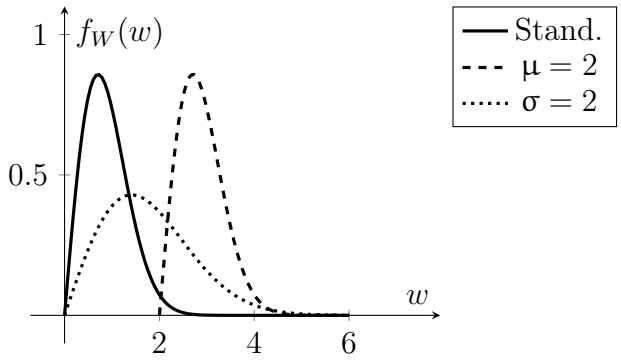


Figure 2.18: Different Weibull probability densities with $\alpha = 2$

Figures 2.17 and 2.18 display the reverse Weibull and the Weibull distribution respectively, for $\alpha = 2$ and the same perturbations of the location and scale parameters. The two figures clarify that each distribution is the perfect "mirror image" of the other, as their names and mathematical relationship suggest. The moments of both the inverse Weibull and the Weibull distribution can again be appropriately derived from the expressions in the GEV case. If a random variable Y follows the reverse Weibull distribution, this is best written with an explicit reference to the "Type III" GEV:

$$Y \sim \text{EV3}\,(\alpha, \mu, \sigma)$$

while one writes:

$$W \sim \text{Weibull}\,(\alpha, \mu, \sigma)$$

if a random variable W follows the (traditional) Weibull distribution.

This section is concluded by highlighting some relationships between the Weibull distribution, other types of GEV distributions, and the exponential distribution (and by extension, all distributions related to the exponential).

Observation 2.18. If $X \sim \text{Exp}\,(1)$, $Y = \mu - \sigma \log(X)$, and $W = \mu + \sigma X^{1/\alpha}$, it is $Y \sim \text{Gumbel}\,(\mu, \sigma)$ and $W \sim \text{Weibull}\,(\alpha, \mu, \sigma)$.

Observation 2.19. If $X \sim \text{Exp}\left(\sqrt{\alpha}\right)$ and $W \sim \text{Weibull}\left(\alpha, 0, \frac{1}{2}\right)$, the two random variables are symmetrically related: $X = \sqrt{W}$ and $W = X^2$.

Observation 2.20. If $Y \sim \text{Frechet}\,(\alpha, \mu_Y, \sigma)$, and $W = (Y - \mu_Y)^{-1} + \mu_W$, it is $W \sim \text{Weibull}\,(\alpha, \mu_W, \sigma^{-1})$.

The (traditional) Weibull distribution is most often used to model the *minimum* value among multiple realizations, contrary to the Gumbel and the Fréchet cases. A frequent application is in *survival analysis* (the statistical study of waiting times) along with the related exponential distribution.

Lecture 3

Random Vectors

This lecture introduces those conceptual and mathematical tools that are necessary to handle multiple random variables: these notions include those of random vector, joint versus marginal probability distribution, multivariate transformation of a random vector, independence, covariance and correlation, conditional distribution and conditional moments. While developing these concepts, this lecture introduces additional relationships between common univariate distributions, concluding with the treatment of two important multivariate distributions. The mathematical notation is chosen so to facilitate the later treatment of econometric theory.

3.1 Multivariate Distributions

In most practical settings of interest, a statistical analyst is often interested in describing the occurrence of multiple events, each best expressed through a single random variable, that are *possibly* related to one another through a probabilistic relationship of dependence. It is therefore important to extend the theory of probability distributions to a *multivariate* environment. The first step in this direction is the definition of a **random vector**.

Definition 3.1. Random Vector. A random vector $\boldsymbol{x}$ of length K is a collection of K random variables $X_1, \ldots, X_K$:

$$\boldsymbol{x} = \begin{pmatrix} X_1 \\ \vdots \\ X_K \end{pmatrix}$$

each with support $\mathbb{X}_k \subseteq \mathbb{R}$ for $k = 1, \ldots, K$.

In light of the definition, the name *random vector* is admittedly uninspiring.

It is worth to observe one specific choice about notation: random vectors are indicated with lower case, bold faced italic characters. In these lectures, a similar notation ($\mathbf{x}$) with roman – instead of italic – characters is instead reserved for the *realizations* of random vectors: that is, the collection of the specific realizations of each of the K random variables that define $\boldsymbol{x}$.

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_K \end{pmatrix}$$

It is also useful to specify what the support of a random vector is.

Definition 3.2. Support of a Random Vector. The support $\mathbb{X} \subseteq \mathbb{R}^K$ of a random vector $\boldsymbol{x}$ is the Cartesian product of all the supports of the random variables featured in $\boldsymbol{x}$.

$$\mathbb{X} = \mathbb{X}_1 \times \cdots \times \mathbb{X}_K$$

Note that the definition of random vector imposes no restriction on the original sample space on which the K random variables are based: it may well be that it is the same for all of them. In such a case, one should allow for the fact that the variables in question are probabilistically dependent, that is, the realization of one random variable provides information about the odds of another random variable's realizations (as expressed through the definition of conditional probability). A **joint probability distribution** is a mathematical function that easily handles such circumstances.

Definition 3.3. Joint Probability Cumulative Distribution. Given a random vector $\boldsymbol{x}$, its joint probability *cumulative* distribution is defined as the following function.

$$F_{\boldsymbol{x}}(\mathbf{x}) = \mathbb{P}(\boldsymbol{x} \leq \mathbf{x}) = \mathbb{P}(X_1 \leq x_1 \cap \cdots \cap X_K \leq x_K)$$

Obviously, a joint probability cumulative distribution takes values in the $[0, 1]$ interval. One can show that the properties of cumulative probability distributions for single random variables (Theorem 1.7) are extended to the joint, multivariate case; in particular, when the limit of *all* its arguments tends to minus infinity (plus infinity) the function tends to zero (one); the function is non-decreasing and right-continuous in all its arguments.

Even mass and density functions have their multivariate analogues.

Definition 3.4. Joint Probability Mass Function. Given any random vector $\boldsymbol{x}$ composed by *discrete* random variables *only*, its joint probability *mass* function $f_{\boldsymbol{x}}(\mathbf{x})$ is defined as follows, for all $\mathbf{x} = (x_1, \ldots, x_K) \in \mathbb{R}^K$.

$$f_{\boldsymbol{x}}(x_1, \ldots, x_K) = \mathbb{P}(X_1 = x_1 \cap \cdots \cap X_K = x_K)$$

A joint probability mass function is related to the cumulative function via the following relationship:

$$\mathbb{P}(\boldsymbol{x} \leq \mathbf{x}) = F_{\boldsymbol{x}}(\mathbf{x}) = \sum_{\mathbf{t} \in \mathbb{X}: \mathbf{t} \leq \mathbf{x}} f_{\boldsymbol{x}}(\mathbf{t})$$

where the summation is taken over the vectors $\mathbf{t}$ in $\mathbb{X}$ whose *all* elements are smaller or equal than *all* the elements of $\mathbf{x}$. In addition, the joint probability mass function must obviously satisfy the following condition.

$$\mathbb{P}(\boldsymbol{x} \in \mathbb{X}) = \sum_{\mathbf{x} \in \mathbb{X}} f_{\boldsymbol{x}}(\mathbf{x}) = 1$$

Definition 3.5. Joint Probability Density Function. Given any random vector $\boldsymbol{x}$ composed by *continuous* random variables *only*, its joint probability *density* function $f_{\boldsymbol{x}}(\mathbf{x})$ is defined as the function that satisfies the following relationship, for all $\mathbf{x} = (x_1, \ldots, x_K) \in \mathbb{R}^K$.

$$F_{\boldsymbol{x}}(x_1, \ldots, x_K) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_K} f_{\boldsymbol{x}}(x_1, \ldots, x_K)\, dx_1 \ldots dx_K$$

In analogy with its univariate counterpart, a joint probability density function only takes nonnegative values (possibly larger than one), and allows to express probabilities about events occurring within specified intervals – now, intervals in $\mathbb{R}^K$. Given two vectors $\mathbf{a} = (a_1, \ldots, a_K)$ and $\mathbf{b} = (b_1, \ldots, b_K)$ with $\mathbf{a}, \mathbf{b} \in \mathbb{R}^K$ and $b_k \geq a_k$ for $k = 1, \ldots, K$, it is as follows.

$$\begin{aligned}\mathbb{P}(a_1 \leq x_1 \leq b_1 \cap \cdots \cap a_K \leq x_K \leq b_K) &= F_{\boldsymbol{x}}(b_1, \ldots, b_K) - F_{\boldsymbol{x}}(a_1, \ldots, a_K) \\ &= \int_{a_1}^{b_1} \cdots \int_{a_K}^{b_K} f_{\boldsymbol{x}}(x_1, \ldots, x_K)\, dx_1 \ldots dx_K\end{aligned}$$

Clearly, the joint density integrates to one over the entire support of $\boldsymbol{x}$.

$$\mathbb{P}(\boldsymbol{x} \in \mathbb{X}) = \int_{\mathbb{X}_1} \ldots \int_{\mathbb{X}_K} f_{\boldsymbol{x}}(x_1, \ldots, x_K)\, dx_1 \ldots dx_K = 1$$

In a multivariate environment, the probability distributions of the individual random variables that compose a random vector are called **marginal** distributions, and can be derived from the joint mass and density functions.

Definition 3.6. Marginal Distribution (discrete case). For a given random vector $\boldsymbol{x}$ made of *discrete* random variables *only*, the probability mass function of X_k – the k-th random variable in $\boldsymbol{x}$ – is obtained as:

$$f_{X_k}(x_k) = \sum_{x_1 \in \mathbb{X}_1} \cdots \sum_{x_{k-1} \in \mathbb{X}_{k-1}} \sum_{x_{k+1} \in \mathbb{X}_{k+1}} \cdots \sum_{x_K \in \mathbb{X}_K} f_{\boldsymbol{x}}(x_1, \ldots, x_K)$$

and thus $F_{X_k}(x_k) = \sum_{t=\inf \mathbb{X}_k}^{x_k} f_{X_k}(t)$.

The above relationship expresses a summation over all the supports of all the other random variables in $\boldsymbol{x}$, *excluding* X_k. It has to be interpreted in a general sense, whatever the dimension of K and the actual index k are: if K is small and/or k is at either extreme of the list, it must be reformulated accordingly. This is best seen with small values of K.

Example 3.1. Joint Medical Outcomes. Recall Example 1.4 about the probability of getting sick following the take-up (or lack thereof) of some preemptive medical treatment. One could reformulate that example via a random vector (X, Y) where: $x = 1$ if an individual is a taker, $x = 0$ if he or she hesitates, $y = 1$ if an individual stays healthy, $y = 0$ if he or she gets sick. This is a *bivariate Bernoulli distribution* with:

$$f_{X,Y}(x = 1, y = 1) = 0.40, \quad f_{X,Y}(x = 1, y = 0) = 0.20,$$
$$f_{X,Y}(x = 0, y = 1) = 0.15, \quad f_{X,Y}(x = 0, y = 0) = 0.25,$$

and the marginal mass function of either Bernoulli-distributed random variable is obtained by appropriately summing over the support of the other.

$$f_X(x) = f_{X,Y}(x, y = 1) + f_{X,Y}(x, y = 0) \quad \text{for } x = 0, 1$$
$$f_Y(y) = f_{X,Y}(x = 1, y) + f_{X,Y}(x = 0, y) \quad \text{for } y = 0, 1$$

Bernoulli distributions are typically represented through *frequency tables*, where joint probabilities are displayed at the center, and marginal probabilities at the margins; a frequency table for this example is shown below.

	$Y = 0$	$Y = 1$	*Total*
$X = 0$	0.25	0.15	0.40
$X = 1$	0.20	0.40	0.60
Total	0.45	0.55	1

The denomination "marginal" clearly derives from this graphical device. ■

Analogously, the density functions of continuous marginal distributions can be obtained by integrating the joint density over the support of all the random variables in the random vector, except the one of interest.

Definition 3.7. Marginal Distribution (continuous case). For a given random vector $\boldsymbol{x}$ made of *continuous* random variables *only*, the probability density function of X_k – the k-th random variable in $\boldsymbol{x}$ – is obtained as:

$$f_{X_k}(x_k) = \int_{\times_{\ell \neq k} \mathbb{X}_\ell} f_{\boldsymbol{x}}(\mathbf{x}) \, \mathrm{d}\mathbf{x}_{-k}$$

and thus $F_{X_k}(x_k) = \int_{-\infty}^{x_k} f_{X_k}(t) \, dt$.

In this more compact definition, the notation $\times_{\ell \neq k} \mathbb{X}_\ell$ indicates the Cartesian product of all the supports of each random variable in $\boldsymbol{x}$ excluding X_k: e.g. $\times_{\ell \neq k} \mathbb{X}_\ell = \mathbb{X}_1 \times \cdots \times \mathbb{X}_{k-1} \times \mathbb{X}_{k+1} \times \cdots \times \mathbb{X}_K$; similarly the expression $\mathrm{d}\mathbf{x}_{-k}$ for the differential is to be interpreted as the product of all differentials except the one for x_k: $\mathrm{d}\mathbf{x}_{-k} = dx_1 \ldots dx_{k-1} dx_{k+1} \ldots dx_K$.

Example 3.2. Bivariate Normal Distribution. A two-dimensional random vector $\boldsymbol{x} = (X_1, X_2)$ follows a *bivariate* normal distribution if its joint density function $f_{X_1,X_2}(x_1, x_2)$ is, for some parameters $\mu_1 \in \mathbb{R}$, $\mu_2 \in \mathbb{R}$, $\sigma_1 \in \mathbb{R}_{++}$, $\sigma_2 \in \mathbb{R}_{++}$, and $\rho \in [-1, 1]$, expressed as follows.

$$f_{X_1,X_2}(x_1, x_2; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times \\ \times \exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2(1-\rho^2)} - \frac{(x_2-\mu_2)^2}{2\sigma_2^2(1-\rho^2)} + \frac{\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2(1-\rho^2)} \right) \quad (3.1)$$

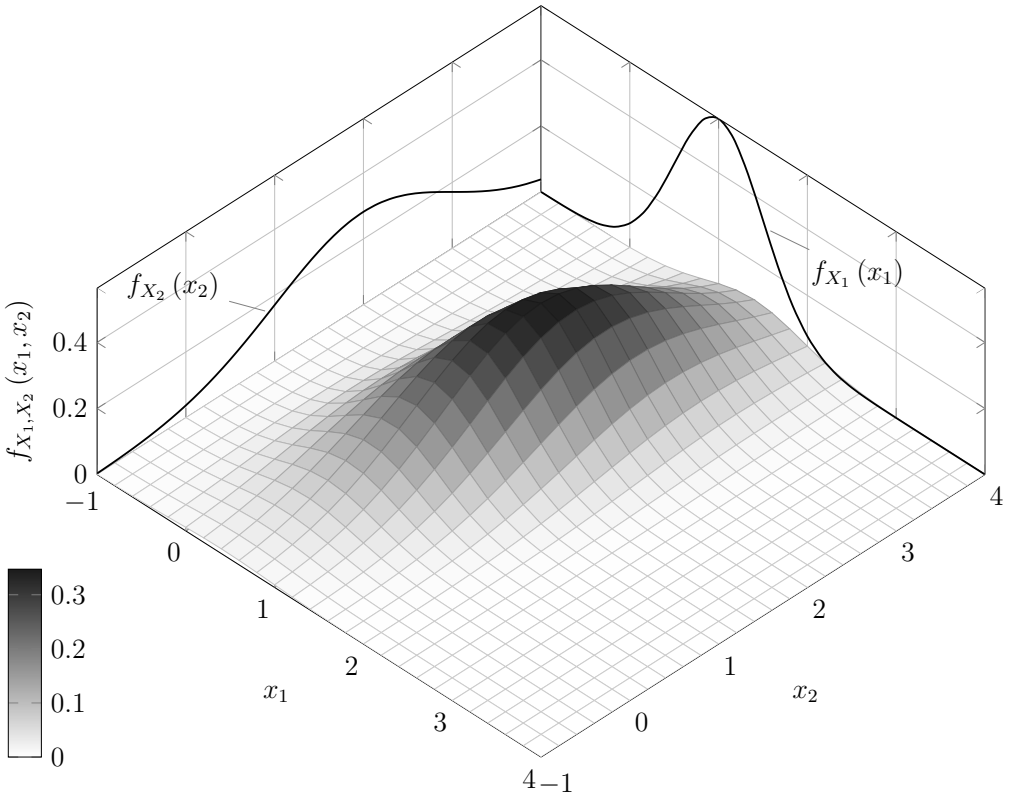


Figure 3.1: Bivariate normal for $\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1 = .5$, $\sigma_2 = 1$, $\rho = .4$.

Figure 3.1 represents a bivariate normal distribution (for selected values of the parameters) in a three-dimensional plot, with domain restricted to the $[-1, 4]^2$ area. Furthermore, it projects the density functions of the *marginal* distributions for X_1 and X_2 onto planes defined at the margins of the space under consideration. A somewhat tedious exercise in integration of the joint density would reveal that $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$. ■

In numerous cases of practical interest, a random vector features both continuous and discrete random variables. In such situations, obtaining the marginal probability mass or density function for a specific random variable is an exercise in combining summation and integration as appropriate.

Example 3.3. Mixed discrete-continuous random vector. Consider a simple example of a random vector $\boldsymbol{x} = (H, G)$ where H is a continuous random variable representing human height, while G is a Bernoulli random variable describing gender (e.g. $G = 1$ for females and $G = 0$ for males). It turns out that human height is normally distributed but with different parameters in the population. One can provide a full probability description of human height in the population as follows:

$$f_{H,G}(h, g = 1; \mu_F, \mu_M, \sigma_F, \sigma_M, p) = \frac{1}{\sigma_F}\phi\left(\frac{h - \mu_F}{\sigma_F}\right) \cdot p$$
$$f_{H,G}(h, g = 0; \mu_F, \mu_M, \sigma_F, \sigma_M, p) = \frac{1}{\sigma_M}\phi\left(\frac{h - \mu_M}{\sigma_M}\right) \cdot (1 - p)$$

where subscripts below parameters refer to genders (female or male) and $\phi(\cdot)$ is the standard normal density function. The marginal density function of H is obtained by simply summing the two expressions:

$$f_H(h; \mu_F, \mu_M, \sigma_F, \sigma_M, p) = \frac{1}{\sigma_F}\phi\left(\frac{h - \mu_F}{\sigma_F}\right) \cdot p + \frac{1}{\sigma_M}\phi\left(\frac{h - \mu_M}{\sigma_M}\right) \cdot (1 - p)$$

while the observation that both densities must integrate to one gives:

$$f_G(g = 1; p) = p$$
$$f_G(g = 0; p) = 1 - p$$

thus returning the marginal mass function for G. ■

The results about **transformations** of random variables can be generalized to random vectors. In this case, the interest falls on a transformed random vector $\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x})$ where $\boldsymbol{g}(\cdot)$ is a function taking K arguments and returning J values, with possibly $J \neq K$ (thus $\boldsymbol{y}$ would have dimension J).

Once again, this problem is tractable only so long as the transformation is invertible; in a multivariate setting this means that one can define a set of functions $g_1^{-1}(\cdot), \ldots, g_K^{-1}(\cdot)$ such that:

$$X_k = g_k^{-1}(Y_1, \ldots, Y_J)$$

for $k = 1, \ldots, K$. In such a case, a transformed discrete joint mass function can be obtained by generalizing (1.5)

$$f_{\boldsymbol{y}}(\mathbf{y}) = f_{\boldsymbol{x}}\left(g_1^{-1}(\mathbf{y}), \ldots, g_K^{-1}(\mathbf{y})\right) \tag{3.2}$$

while the cumulative distribution $F_{\boldsymbol{y}}(\mathbf{y})$ is derived consequently. For continuous random vectors, Theorem 1.11 is extended as follows, by imposing the additional restrictions that the transformation $\boldsymbol{g}(\cdot)$ is bijective (both injective and surjective, i.e. *"one-to-one and onto"*) and that $K = J$, that is, the transformation does not affect the length of the random vector.

Theorem 3.1. Joint Density of Transformed Random Vectors. *Let $\boldsymbol{x}$ and $\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x})$ be two random vectors of length K that are related by a bijective transformation $\boldsymbol{g}(\cdot)$ which preserves vector length, $\mathbb{X}$ and $\mathbb{Y}$ their respective supports, and $f_{\boldsymbol{x}}(\mathbf{x})$ the joint probability density function of $\boldsymbol{x}$, which is continuous on $\mathbb{X}$. If the inverse of the transformation function, $g_k^{-1}(\cdot)$, is continuously differentiable on $\mathbb{Y}$ for $k = 1, \ldots, K$, the joint probability density function of $\boldsymbol{y}$ can be calculated as:*

$$f_{\boldsymbol{y}}(\mathbf{y}) = \begin{cases} f_{\boldsymbol{x}}\left(g_1^{-1}(\mathbf{y}), \ldots, g_K^{-1}(\mathbf{y})\right) \cdot \left|\det\left(\frac{\partial}{\partial \boldsymbol{y}^{\mathrm{T}}} \boldsymbol{g}^{-1}(\mathbf{y})\right)\right| & \text{if } \mathbf{y} \in \mathbb{Y} \\ 0 & \text{if } \mathbf{y} \notin \mathbb{Y} \end{cases}$$

where $\boldsymbol{g}^{-1}(\mathbf{y}) = \left(g_1^{-1}(\mathbf{y}), \ldots, g_K^{-1}(\mathbf{y})\right)^{\mathrm{T}}$, with the following $K \times K$ Jacobian matrix:

$$\frac{\partial}{\partial \boldsymbol{y}^{\mathrm{T}}} \boldsymbol{g}^{-1}(\mathbf{y}) = \begin{bmatrix} \frac{\partial g_1^{-1}(\mathbf{y})}{\partial y_1} & \frac{\partial g_1^{-1}(\mathbf{y})}{\partial y_2} & \cdots & \frac{\partial g_1^{-1}(\mathbf{y})}{\partial y_K} \\ \frac{\partial g_2^{-1}(\mathbf{y})}{\partial y_1} & \frac{\partial g_2^{-1}(\mathbf{y})}{\partial y_2} & \cdots & \frac{\partial g_2^{-1}(\mathbf{y})}{\partial y_K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_K^{-1}(\mathbf{y})}{\partial y_1} & \frac{\partial g_K^{-1}(\mathbf{y})}{\partial y_2} & \cdots & \frac{\partial g_K^{-1}(\mathbf{y})}{\partial y_K} \end{bmatrix}$$

the absolute value of whose determinant is denoted as $\left|\det\left(\frac{\partial}{\partial \boldsymbol{y}^{\mathrm{T}}} \boldsymbol{g}^{-1}(\mathbf{y})\right)\right|$.

Proof. This result is a particular application of Jacobian transformations from multivariate calculus. □

This result can be additionally generalized to transformations that are not bijective, but are bijective on each element with positive probability of some partition of $\mathbb{X}$, similarly as in Theorem 1.12 for the univariate case.

Example 3.4. Bivariate lognormal distribution. Consider the bivariate normally distributed random vector from Example 3.2, and the random vector $\boldsymbol{y} = (Y_1, Y_2)$ which is obtained through the following transformation.

$$\boldsymbol{y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \boldsymbol{g} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} \exp(X_1) \\ \exp(X_2) \end{pmatrix}$$

This implies $X_1 = g_1^{-1}(Y_1, Y_2) = \log(Y_1)$ and $X_2 = g_2^{-1}(Y_1, Y_2) = \log(Y_2)$, hence $\det\left(\frac{\partial}{\partial \boldsymbol{y}^\top} \boldsymbol{g}^{-1}(y_1, y_2)\right) = (y_1 y_2)^{-1} > 0$ and:

$$f_{Y_1,Y_2}(y_1, y_2; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \frac{1}{y_1 y_2} \times$$
$$\times \exp\left(-\frac{(\log y_1 - \mu_1)^2}{2\sigma_1^2(1-\rho^2)} - \frac{(\log y_2 - \mu_2)^2}{2\sigma_2^2(1-\rho^2)} + \frac{\rho(\log y_1 - \mu_1)(\log y_2 - \mu_2)}{\sigma_1\sigma_2(1-\rho^2)} \right)$$

showing that $\boldsymbol{y} = (Y_1, Y_2)$ follows a *bivariate lognormal distribution* on $\mathbb{R}^2_{++}$ with parameters $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$, as displayed in Figure 3.2 below.

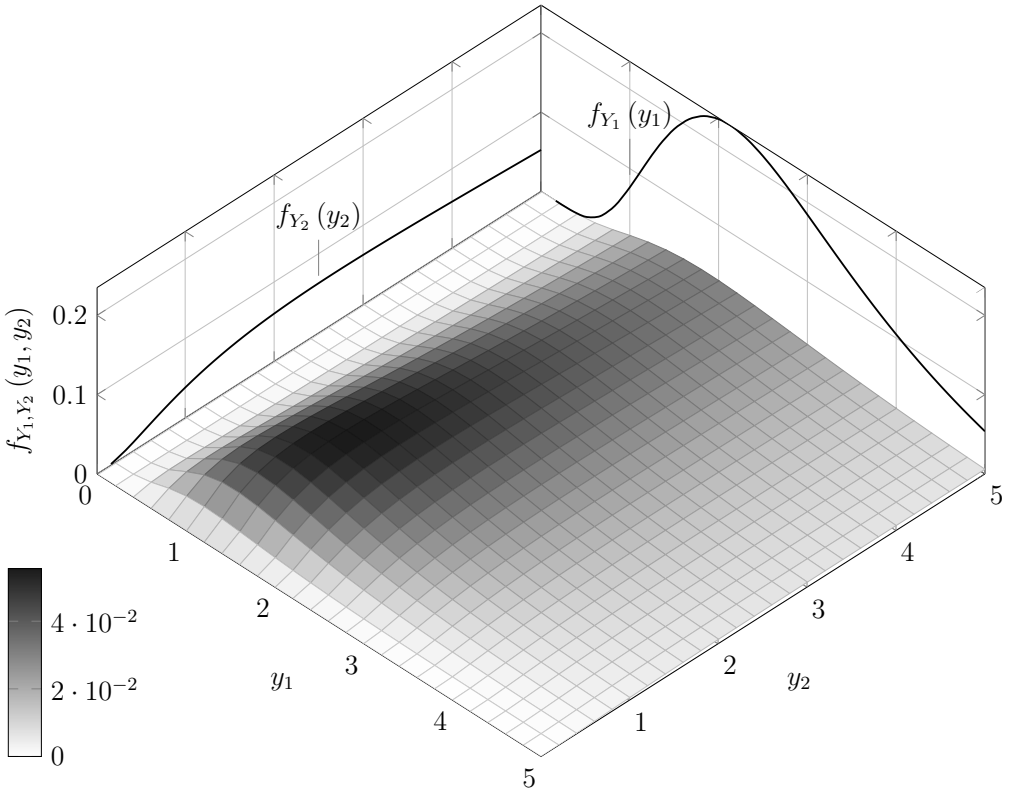


Figure 3.2: Bivariate lognormal, $\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1 = .5$, $\sigma_2 = 1$, $\rho = .4$.

This example is relatively simple, because the two original random variables X_1 and X_2 do not interact in the transformation (that is, the Jacobian is diagonal). Some more elaborate cases are discussed in the next section. However, this example is also useful in itself as an occasion to graphically visualize another bivariate distribution (in this case, the lognormal). ■

All the concepts and ideas discussed until this point in this lecture extend easily to **random matrices**, that is arrayed combinations of L random vectors, written e.g. $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_L \end{bmatrix}$. In analogy with random vectors, the realizations of random matrices adopt a romanized notation too, being written for example as $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_L \end{bmatrix}$. The necessity to deal with random matrices explains why uppercase letters are not used to denote random vectors. Random matrices do not involve particular conceptual hurdles, being nothing else but a different algebraic way of arraying random variables. However, they are necessary in multivariate statistical analysis and econometrics as a means to more elegantly, compactly and often clearly describe statistical estimators and their properties.

3.2 Independence and Random Ratios

When a random vector involves some underlying random variables that are probabilistically unrelated from one another, we say that these random variables are **independent**. Intuitively, this means that the realization of one specific random variable is uninformative with respect to the potential realization of the other(s). In other words (tracing back to a definition given in Lecture 1), any event described by either random variable is *statistically independent* to any event described by the other. Because the probability of events expressed by random variables are completely characterized by their mass or density functions, it is possible to establish the following definitions.

Definition 3.8. Independent Random Variables. Let $\boldsymbol{x} = (X, Y)$ be a random vector with joint probability mass or density function $f_{X,Y}(x, y)$, and marginal mass or density functions $f_X(x)$ and $f_Y(y)$. Let uppercase F denote corresponding cumulative distributions instead (joint or marginal). The two random variables X and Y are *independent* if the two equivalent conditions below hold.

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) \iff F_{X,Y}(x, y) = F_X(x) F_Y(y)$$

The definition of independence clearly extends to multiple random variables. However, one must be careful at distinguishing between independence

between random vectors versus independence *between multiple random variables* that might belong to the same random vectors. Thus, it is important to keep the following two definitions separate.

Definition 3.9. Mutually – or Pairwise – Independent Random Variables. Let $\boldsymbol{x} = (X_1, \ldots, X_K)$ be a random vector with joint probability mass or density function $f_{\boldsymbol{x}}(\mathbf{x})$, and marginal mass or density functions $f_{X_1}(x_1), \ldots, f_{X_K}(x_K)$. Let uppercase F denote corresponding cumulative distributions instead (joint or marginal). The random variables $X_1, \ldots, X_K$ are *pairwise independent* if every pair of random variables listed in $\boldsymbol{x}$ are independent, and they are *mutually independent* if the two equivalent conditions below hold.

$$f_{\boldsymbol{x}}(\mathbf{x}) = \prod_{k=1}^{K} f_{X_k}(x_k) \iff F_{\boldsymbol{x}}(\mathbf{x}) = \prod_{k=1}^{K} F_{X_k}(x_k)$$

Observe that while mutual independence implies pairwise independence, the converse is not true.

Definition 3.10. Independent Random Vectors. Let $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J)$ be a sequence of J random vectors whose joint probability mass or density function is written as $f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_J}(\mathbf{x}_1, \ldots, \mathbf{x}_J)$. Let the joint probability mass or density functions of an individual nested random vector be $f_{\boldsymbol{x}_i}(\mathbf{x}_i)$, where $i = 1, \ldots, J$, and the joint probability mass or density functions of any two random vectors indexed i, j (with $i \neq j$) as $f_{\boldsymbol{x}_i,\boldsymbol{x}_j}(\mathbf{x}_i, \mathbf{x}_j)$. Finally, let uppercase F denote corresponding cumulative distributions instead (joint or marginal). Any pair of random vectors indexed i and j are *independent* if the two equivalent conditions below hold.

$$f_{\boldsymbol{x}_i,\boldsymbol{x}_j}(\mathbf{x}_i, \mathbf{x}_j) = f_{\boldsymbol{x}_i}(\mathbf{x}_i) f_{\boldsymbol{x}_j}(\mathbf{x}_j) \iff F_{\boldsymbol{x}_i,\boldsymbol{x}_j}(\mathbf{x}_i, \mathbf{x}_j) = F_{\boldsymbol{x}_i}(\mathbf{x}_i) F_{\boldsymbol{x}_j}(\mathbf{x}_j)$$

If the above holds for any i, j distinct pair, the J random vectors are said to be *pairwise independent*. The J random vectors are *mutually independent* if the two equivalent conditions below hold.

$$f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_J}(\mathbf{x}_1, \ldots, \mathbf{x}_J) = \prod_{i=1}^{J} f_{\boldsymbol{x}_i}(\mathbf{x}_i) \iff F_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_J}(\mathbf{x}_1, \ldots, \mathbf{x}_J) = \prod_{i=1}^{J} F_{\boldsymbol{x}_i}(\mathbf{x}_i)$$

Note that within each random vector, the underlying random variables are not necessarily independent. Moreover, if all the random vectors in question have length one, these definitions reduce to those given above.

Two results about independent random variables are well worth of being discussed: the first helps the interpretation of independence, the second is of more practical use and instrumental to derive other properties and results.

Theorem 3.2. Independence of Events. *Any two events mapped by two independent random* variables X *and* Y *are statistically independent.*

Proof. (Outline.) This requires to show that, for any two events $\mathbb{A} \subset \mathbb{S}_X$ and $\mathbb{B} \subset \mathbb{S}_Y$ – where $\mathbb{S}_X$ and $\mathbb{S}_Y$ are the primitive sample spaces of X and Y respectively – it is:

$$\mathbb{P}\left(X \in X\left(\mathbb{A}\right) \cap Y \in Y\left(\mathbb{B}\right)\right) = \mathbb{P}\left(X \in X\left(\mathbb{A}\right)\right) \cdot \mathbb{P}\left(Y \in Y\left(\mathbb{B}\right)\right)$$

which follows from the definitions of (joint) cumulative distribution, mass and density functions, and that of independent events. □

Generalization: Mutual Independence between Events. *Any combination of events mapped by a sequence of* J *mutually independent random* vectors $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J)$ *are mutually independent.*

Proof. (Outline.) Extending the reasoning above, consider a collection of J events denoted by $\mathbb{A}_i \subset \mathbb{S}_{\boldsymbol{x}_i}$ for $i = 1, \ldots, J$, where $\mathbb{S}_{\boldsymbol{x}_i}$ is the primitive sample space of $\boldsymbol{x}_i$. It must be shown that:

$$\mathbb{P}\left(\bigcap_{i=1}^{J}\left(\boldsymbol{x}_i \in \boldsymbol{x}_i\left(\mathbb{A}_i\right)\right)\right) = \prod_{i=1}^{J}\mathbb{P}\left(\boldsymbol{x}_i \in \boldsymbol{x}_i\left(\mathbb{A}_i\right)\right)$$

which follows by analogous considerations. □

Theorem 3.3. Independence of Functions of Random Variables. *Consider two independent random variables* X *and* Y*, and let* $U = g_X\left(X\right)$ *be a transformation of* X *and* $V = g_Y\left(Y\right)$ *a transformation of* Y*. The two transformed random variables* U *and* V *are independent.*

Proof. (Outline.) This requires to show that,

$$f_{U,V}\left(u, v\right) = f_U\left(u\right) f_V\left(v\right) \iff F_{U,V}\left(u, v\right) = F_U\left(u\right) F_V\left(v\right)$$

which is achieved by manipulating the inverse mappings $g_X^{-1}\left([a, b]\right)$ and $g_Y^{-1}\left([a, b]\right)$ for any appropriate interval $[a, b] \subset \mathbb{R}$, with $a \leq b$. □

Generalization: Independence of Functions of Random Vectors. *Consider a sequence of mutually independent random vectors* $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J)$*, as well as a sequence of transformations* $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_J)$ *such that* $\boldsymbol{y}_i = \boldsymbol{g}_i\left(\boldsymbol{x}_i\right)$ *for* $i = 1, \ldots, J$*. The* J *transformed random vectors* $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_J)$ *are also themselves mutually independent.*

Proof. (Outline.) The proof extends the logic of the bivariate case to higher dimensions; it requires manipulating the J Jacobian transformations. □

The concept of independence is central in theoretical and applied statistics. In probability theory, it helps identify the distribution of functions of random variables that can be expressed as a **random ratio** ($Y = X_1/X_2$), or as a **random product** ($Y = X_1X_2$). Some results about random ratios are especially important in applied statistics, and they are developed next in the form of observations about distributions.

Observation 3.1. If $X_1 \sim \mathcal{N}(0,1)$ and $X_2 \sim \mathcal{N}(0,1)$, and the two random variables X_1 and X_2 are independent, the random variable Y obtained as $Y = X_1/X_2$ is such that $Y \sim \text{Cauchy}(0,1)$.

Proof. To demonstrate this assertion, consider first that if the two standard normal random variables X_1 and X_2 in question are independent, their joint density function is by definition a simplified version of (3.1).

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{2\pi}\exp\left(-\frac{x_1^2+x_2^2}{2}\right)$$

Consider the random vectors $\boldsymbol{x} = (X_1, X_2)$ and $\boldsymbol{y} = (Y, Z)$ where $Z = |X_2|$ and the support of $\boldsymbol{y}$ is $\mathbb{Y} = \mathbb{R} \times \mathbb{R}_+$. It is straightforward to verify that the transformation that relates $\boldsymbol{x}$ with $\boldsymbol{y}$ is not bijective, however the support of $\boldsymbol{x}$ can be partitioned in such a way that it is bijective on each component with positive probability, in the spirit of Theorem 1.12, as follows.

$$\begin{aligned}\mathbb{X}_0 &= \{(x_1,x_2) : x_2 = 0\}\\ \mathbb{X}_1 &= \{(x_1,x_2) : x_2 < 0\}\\ \mathbb{X}_2 &= \{(x_1,x_2) : x_2 > 0\}\end{aligned}$$

The intermediate objective is to derive the joint probability density of $\boldsymbol{y}$. To this end, one must analyze the Jacobian matrices defined on both sets $\mathbb{X}_1$ and $\mathbb{X}_2$, apply Theorem 3.1 on them, and sum the results. The inverse transformations on $\mathbb{X}_1$, where $Z = -X_2$, are:

$$\begin{aligned}X_1 &= g^{-1}_{1,\mathbb{X}_1}(Y,Z) = -YZ\\ X_2 &= g^{-1}_{2,\mathbb{X}_1}(Y,Z) = -Z\end{aligned}$$

therefore the determinant of the Jacobian matrix is as follows.

$$\det\left(\frac{\partial}{\partial \boldsymbol{y}^{\mathrm{T}}}\boldsymbol{g}^{-1}_{\mathbb{X}_1}(y,z)\right) = \det\begin{pmatrix}-z & -y\\ 0 & -1\end{pmatrix} = z > 0$$

In the case of $\mathbb{X}_2$ it is $Z = X_2$, hence the analysis proceeds as:

$$\begin{aligned}X_1 &= g^{-1}_{1,\mathbb{X}_2}(Y,Z) = YZ\\ X_2 &= g^{-1}_{2,\mathbb{X}_2}(Y,Z) = Z\end{aligned}$$

and:

$$\det\left(\frac{\partial}{\partial \boldsymbol{y}^{\mathrm{T}}}\boldsymbol{g}_{\mathbb{X}_2}^{-1}(y,z)\right)=\det\begin{pmatrix}z & y\\ 0 & 1\end{pmatrix}=z>0$$

and in both cases the result is equal to z which is positive by construction, leaving no need to take absolute values. In addition, since the two sets $\mathbb{X}_1$ and $\mathbb{X}_2$ in the partition are both symmetric around $x_2=0$ – just like the transformation that defines z is – the joint density of $\boldsymbol{y}$ can be obtained by applying 3.1 once on the joint density of $\boldsymbol{x}$, and multiplying the result by two.

$$f_{Y,Z}(y,z)=\frac{z}{\pi}\exp\left(-\frac{(y^2+1)\,z^2}{2}\right)$$

The final objective is to show that the marginal density of Y indeed follows the standard Cauchy distribution. To achieve this, the route is to integrate the joint density of $\boldsymbol{y}$ over the support of Z, which is $\mathbb{R}_+$:

$$\begin{aligned}f_Y(y)&=\int_0^{+\infty}f_{Y,Z}(y,z)\,dz\\&=\int_0^{+\infty}\frac{z}{\pi}\exp\left(-\frac{(y^2+1)\,z^2}{2}\right)dz\\&=\int_0^{+\infty}\frac{1}{2\pi}\exp\left(-\frac{(y^2+1)}{2}u\right)du\\&=\frac{1}{\pi(y^2+1)}\int_0^{+\infty}\frac{(y^2+1)}{2}\exp\left(-\frac{(y^2+1)}{2}u\right)du\\&=\frac{1}{\pi(y^2+1)}\end{aligned}$$

where in the third line the change of variable $u=z^2$ is applied, while the integral in the fourth line vanishes because it is the total probability of an exponential distribution with parameter $\lambda=(y^2+1)/2$. The final result is indeed the probability density function of a standard Cauchy distribution, as it was originally postulated. □

Observation 3.2. If $Z\sim\mathcal{N}(0,1)$ and $X\sim\chi^2(\nu)$, and the two random variables Z and X are independent, the random variable Y obtained as $Y=Z/\sqrt{X/\nu}$ is such that $Y\sim\mathcal{T}(\nu)$.

Proof. The first step is to show what the distribution of the transformed random variable $W=\sqrt{X/\nu}$ is. The distribution of the squared root of a random variable following the chi-squared distribution is called (unsurprisingly) the **chi distribution**, and W just follows one rescaled version of it. This transformation is monotone, it preserves the support $\mathbb{X}=\mathbb{W}=\mathbb{R}_+$, its

inverse is $X = g^{-1}(W) = \nu W^2$ and thus $\frac{dx}{dw} = 2\nu w > 0$, and consequently the density function of W is:

$$f_W(w;\nu) = \frac{\nu^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)\cdot 2^{\frac{\nu}{2}-1}} w^{\nu-1}\exp\left(-\frac{\nu w^2}{2}\right) \quad \text{for } w>0$$

and therefore the joint density function of the random vector $\boldsymbol{w} = (Z, W)$ is, given $\phi(z)$ the density function of the standard normal distribution:

$$\begin{aligned} f_{\boldsymbol{w}}(z,w;\nu) &= \phi(z) f_W(w;\nu) \\ &= \frac{1}{\sqrt{2\pi}}\frac{\nu^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)\cdot 2^{\frac{\nu}{2}-1}} w^{\nu-1}\exp\left(-\frac{z^2+\nu w^2}{2}\right) \end{aligned}$$

for $z \in \mathbb{R}$ and $w \in \mathbb{R}_{++}$; note that a product of the two densities is sufficient here because by Theorem 3.3, if Z is independent of X it is also independent of W. The next step is to obtain the joint density function of the random vector $\boldsymbol{y} = (Y, W)$, which is related to $\boldsymbol{w}$ by a support-preserving, bivariate, bijective transformation whose inverse has a Jacobian matrix with determinant $w > 0$, similarly to the analysis of Observation 3.1. Since $Z = YW$, the density function in question is, for $y \in \mathbb{R}$ and $w \in \mathbb{R}_{++}$:

$$\begin{aligned} f_{\boldsymbol{y}}(y,w;\nu) &= w\cdot f_{\boldsymbol{w}}(yw,w;\nu) \\ &= \frac{\nu^{\frac{\nu+1}{2}}}{\Gamma\left(\frac{\nu}{2}\right)\cdot 2^{\frac{\nu-1}{2}}}\frac{1}{\sqrt{\pi\nu}} w^{\nu}\exp\left(-\frac{(y^2+\nu)w^2}{2}\right) \end{aligned}$$

and the marginal density of Y is obtained by integrating W over $\mathbb{R}_{++}$. The result is the density function of a Student t-distribution with parameter ν.

$$\begin{aligned} f_Y(y;\nu) &= \int_0^{+\infty} f_{Y,W}(y,w)\,dw \\ &= \frac{1}{\Gamma\left(\frac{\nu}{2}\right)}\frac{1}{\sqrt{\pi\nu}}\int_0^{+\infty}\frac{\nu^{\frac{\nu+1}{2}}}{2^{\frac{\nu-1}{2}}} w^{\nu}\exp\left(-\frac{(\nu+y^2)w^2}{2}\right)dw \\ &= \frac{1}{\Gamma\left(\frac{\nu}{2}\right)}\frac{1}{\sqrt{\pi\nu}}\int_0^{+\infty}\left(\frac{\nu}{2}\right)^{\frac{\nu+1}{2}} u^{\frac{\nu-1}{2}}\exp\left(-\frac{\nu}{2}\left(1+\frac{y^2}{\nu}\right)u\right)du \\ &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\frac{1}{\sqrt{\pi\nu}}\left(1+\frac{y^2}{\nu}\right)^{-\frac{\nu+1}{2}}\times \\ &\qquad \times\int_0^{+\infty}\frac{1}{\Gamma\left(\frac{\nu+1}{2}\right)}\left(\frac{\nu+y^2}{2}\right)^{\frac{\nu+1}{2}} u^{\frac{\nu-1}{2}}\exp\left(-\left(\frac{\nu+y^2}{2}\right)u\right)du \\ &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\frac{1}{\sqrt{\pi\nu}}\left(1+\frac{y^2}{\nu}\right)^{-\frac{\nu+1}{2}} \end{aligned}$$

In the above analysis, the third line applies the change of variable $u = w^2$; the fourth line is obtained through some manipulation, whereas the integral therein is recognized as the density function of a Gamma distribution with parameters $\alpha = (\nu + 1)/2$ and $\beta = (\nu + y^2)/2$, thus vanishing. □

Observation 3.3. If $X_1 \sim \chi^2(\nu_1)$ and $X_2 \sim \chi^2(\nu_2)$, and the two random variables X_1 and X_2 are independent, the random variable Y obtained as $Y = (X_1/\nu_1)/(X_2/\nu_2)$ is such that $Y \sim \mathcal{F}(\nu_1, \nu_2)$.

Proof. (Outline.) This proceeds as in the previous two observations. First, define $W_1 = X_1/\nu_1$; the density function of this transformation is the same as X_1's but multiplied by ν_1, and similarly for $W_2 = X_2/\nu_2$. The next step is the transformation $Y = W_1/W_2$ and $Z = |W_2|$; the joint density function of Y and Z can be derived consequently from the one of W_1 and W_2. Some manipulation would then reveal that the marginal density of the ratio Y is that of an F-distribution with parameters ν_1 and ν_2. □

The last observation is presented completely without proof.

Observation 3.4. If $X_1 \sim \Gamma(\alpha, \gamma)$ and $X_2 \sim \Gamma(\beta, \gamma)$, and the two random variables X_1 and X_2 are independent, the random variable Y obtained as $Y = X_1/(X_1 + X_2)$ is such that $Y \sim \text{Beta}(\alpha, \beta)$, and is independent of the random variable W obtained as $W = X_1 + X_2$ such that $W \sim \Gamma(\alpha + \beta, \gamma)$.

This completes the picture about the best known results on random ratios. Among these Observations, 3.2 and 3.3 play an important role in statistical inference, as elaborated in the next lectures.

3.3 Multivariate Moments

To every random vector $\boldsymbol{x} = (X_1, \ldots, X_K)$ is associated some list of moments, both uncentered and centered, that pertain to all the random variables featured in $\boldsymbol{x}$ (insofar as these moments exist). There is not a distinct definition for these. However, it is interesting to analyze how they formally relate with the joint distribution of $\boldsymbol{x}$, through the link between joint and marginal mass or density functions. Beginning with the mean, the r-th *uncentered* moments of discrete and continuous random variables belonging to the random vector $\boldsymbol{x}$ can be written, respectively, as:

$$\mathbb{E}[X_k^r] = \sum_{x_1 \in \mathbb{X}_1} \cdots \sum_{x_K \in \mathbb{X}_K} x_k^r f_{\boldsymbol{x}}(\mathbf{x})$$

$$\mathbb{E}[X_k^r] = \int_{\mathbb{X}_1} \ldots \int_{\mathbb{X}_K} x_k^r f_{\boldsymbol{x}}(\mathbf{x}) \, \mathrm{d}\mathbf{x}$$

while the r-th *centered* moments, including the variance, as:

$$\mathbb{E}\left[(X_k - \mathbb{E}\left[X_k\right])^r\right] = \sum_{x_1 \in \mathbb{X}_1} \cdots \sum_{x_K \in \mathbb{X}_K} \left(x_k - \mathbb{E}\left[X_k\right]\right)^r f_{\boldsymbol{x}}\left(\mathbf{x}\right)$$

$$\mathbb{E}\left[(X_k - \mathbb{E}\left[X_k\right])^r\right] = \int_{\mathbb{X}_1} \dots \int_{\mathbb{X}_K} \left(x_k - \mathbb{E}\left[X_k\right]\right)^r f_{\boldsymbol{x}}\left(\mathbf{x}\right) \mathrm{d}\mathbf{x}$$

where in both cases, $\mathrm{d}\mathbf{x} = dx_1 \dots dx_K$ is the product of all the differentials.

In a multivariate context it is interesting to describe the degree by which two random variables tend to deviate from their mean in the same direction, in a probabilistic sense. This concept is expressed by the **covariance** (an absolute measure) and the **correlation** (a normalized one).

Definition 3.11. Covariance. For any two random variables X_k and X_ℓ belonging to a random vector $\boldsymbol{x}$, their specific covariance is defined as the expectation of a particular function of X_k and X_ℓ, that is, the product of both variables' deviations from their respective means.

$$\mathbb{C}\mathrm{ov}\left[X_k, X_\ell\right] = \mathbb{E}\left[\left(X_k - \mathbb{E}\left[X_k\right]\right)\left(X_\ell - \mathbb{E}\left[X_\ell\right]\right)\right]$$

The full expression is written as follows, for discrete and continuous random variables respectively.

$$\mathbb{C}\mathrm{ov}\left[X_k, X_\ell\right] = \sum_{x_1 \in \mathbb{X}_1} \cdots \sum_{x_K \in \mathbb{X}_K} \left(x_k - \mathbb{E}\left[X_k\right]\right)\left(x_\ell - \mathbb{E}\left[X_\ell\right]\right) f_{\boldsymbol{x}}\left(\mathbf{x}\right)$$

$$\mathbb{C}\mathrm{ov}\left[X_k, X_\ell\right] = \int_{\mathbb{X}_1} \dots \int_{\mathbb{X}_K} \left(x_k - \mathbb{E}\left[X_k\right]\right)\left(x_\ell - \mathbb{E}\left[X_\ell\right]\right) f_{\boldsymbol{x}}\left(\mathbf{x}\right) \mathrm{d}\mathbf{x}$$

The covariance takes positive values if the two variables (X_k and X_ℓ) tend to deviate from the mean in the same direction, and negative vice versa. It must be observed that, however, the covariance expresses a relationship of dependence which is essentially linear; if two random variables tend to move together in a very non-linear or irregular way, this may not be captured at all by the covariance. Similarly to the variance, the definition of covariance can be rewritten in a way that is often more convenient to handle.

$$\begin{aligned}
\mathbb{C}\mathrm{ov}\left[X_k, X_\ell\right] &= \mathbb{E}\left[\left(X_k - \mathbb{E}\left[X_k\right]\right)\left(X_\ell - \mathbb{E}\left[X_\ell\right]\right)\right] \\
&= \mathbb{E}\left[X_k X_\ell\right] - \mathbb{E}\left[X_k\, \mathbb{E}\left[X_\ell\right]\right] - \mathbb{E}\left[X_\ell\, \mathbb{E}\left[X_k\right]\right] + \mathbb{E}\left[X_k\right] \mathbb{E}\left[X_\ell\right] \\
&= \mathbb{E}\left[X_k X_\ell\right] - \mathbb{E}\left[X_k\right] \mathbb{E}\left[X_\ell\right]
\end{aligned}$$

Definition 3.12. Correlation. For any two random variables X_k and X_ℓ belonging to a random vector $\boldsymbol{x}$, their *population* correlation is defined as follows.

$$\mathbb{C}\mathrm{orr}\left[X_k, X_\ell\right] = \frac{\mathbb{C}\mathrm{ov}\left[X_k, X_\ell\right]}{\sqrt{\mathbb{V}\mathrm{ar}\left[X_k\right]}\sqrt{\mathbb{V}\mathrm{ar}\left[X_\ell\right]}}$$

The correlation is a "normalized covariance" which is comparable across different pairs of random variables, thanks to the following result.

Theorem 3.4. Properties of Correlation. *For any two random variables X and Y, it is:*

- **a.** $\mathbb{C}\mathrm{orr}[X, Y] \in [-1, 1]$, *and*
- **b.** $|\mathbb{C}\mathrm{orr}[X, Y]| = 1$ *if and only if there are some real numbers* $a \neq 0$ *and* b *such that* $\mathbb{P}(Y = aX + b) = 1$. *If* $\mathbb{C}\mathrm{orr}[X, Y] = 1$ *it is* $a > 0$, *if* $\mathbb{C}\mathrm{orr}[X, Y] = -1$ *it is* $a < 0$.

Proof. Define the following function:

$$\begin{aligned}\mathbb{C}(t) &= \mathbb{E}\left[[(X - \mathbb{E}[X]) \cdot t + (Y - \mathbb{E}[Y])]^2\right] \\ &= \mathbb{V}\mathrm{ar}[X] \cdot t^2 + 2\,\mathbb{C}\mathrm{ov}[X, Y] \cdot t + \mathbb{V}\mathrm{ar}[Y]\end{aligned}$$

and note that it is nonnegative, because it is defined as the expectation of the square of a random variable. Thus its solution must satisfy:

$$(2\,\mathbb{C}\mathrm{ov}[X, Y])^2 - 4\,\mathbb{V}\mathrm{ar}[X]\,\mathbb{V}\mathrm{ar}[Y] \leq 0 \tag{3.3}$$

or, equivalently:

$$-\sqrt{\mathbb{V}\mathrm{ar}[X]}\sqrt{\mathbb{V}\mathrm{ar}[Y]} \leq \mathbb{C}\mathrm{ov}[X, Y] \leq \sqrt{\mathbb{V}\mathrm{ar}[X]}\sqrt{\mathbb{V}\mathrm{ar}[Y]}$$

thus **a.** is proved. Next, consider that $|\mathbb{C}\mathrm{orr}[X, Y]| = 1$ only if (3.3) holds with equality, or equivalently, $\mathbb{C}(t) = 0$. For this to happen, it must be:

$$\mathbb{P}\left([(X - \mathbb{E}[X])\,t + (Y - \mathbb{E}[Y])]^2 = 0\right) = 1$$

or equivalently:

$$\mathbb{P}\left((X - \mathbb{E}[X])\,t + (Y - \mathbb{E}[Y]) = 0\right) = 1$$

which only occurs if, given $Y = aX + b$:

$$\begin{aligned}a &= -t \\ b &= \mathbb{E}[X] \cdot t + \mathbb{E}[Y] \\ t &= -\frac{\mathbb{C}\mathrm{ov}[X, Y]}{\mathbb{V}\mathrm{ar}[X]}\end{aligned}$$

and the proof of **b.** is completed by showing that a and $\mathbb{C}\mathrm{orr}[X, Y]$ must also share the same sign. □

Result **a.** in the above Theorem characterizes the normalized interpretation of correlation. Result **b.** instead specifies the linear nature of the relationship captured by measures of correlation, which equal either 1 or -1 if and only if the two random variables under consideration are connected through an exact linear dependence.

Additional insights can be acquired by looking at specific moments of *independent* random variables.

Theorem 3.5. Cross-expectation of independent random variables. *Given two* independent *random variables X and Y, it is*

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$$

if the above moments exist.

Proof. The left hand side of the above relationship is the expectation of a random variable which is defined as the product of X and Y:

$$\begin{aligned}\int_{\mathbb{X}}\int_{\mathbb{Y}} xyf_{X,Y}(x,y)\,dxdy &= \int_{\mathbb{X}}\int_{\mathbb{Y}} xyf_X(x)\,f_Y(y)\,dxdy \\ &= \int_{\mathbb{X}} xf_X(x)\,dx \cdot \int_{\mathbb{Y}} yf_Y(y)\,dy\end{aligned}$$

falling back to the product of two expressions corresponding to the definition of mean (for X and Y respectively); the first equality exploits the definition of independent random variables. □

Corollary 1. of Theorem 3.5. *Both the covariance and the correlation between two* independent *random variables X and Y equal zero.*

Corollary 2. of Theorem 3.5. *Given two transformations $U = g_X(X)$ and $V = g_Y(Y)$ of two* independent *random variables X and Y, it is:*

$$\mathbb{E}[UV] = \mathbb{E}[U]\,\mathbb{E}[V]$$

because U and V are also independent (so long as all the relevant moments exist); this also implies that U and V have zero covariance and correlation and that all higher moments of X and Y inherit this property, for example:

$$\begin{aligned}\mathbb{V}\mathrm{ar}[XY] &= \mathbb{E}\left[(X-\mathbb{E}[X])^2(Y-\mathbb{E}[Y])^2\right] \\ &= \mathbb{E}\left[(X-\mathbb{E}[X])^2\right]\mathbb{E}\left[(Y-\mathbb{E}[Y])^2\right] \\ &= \mathbb{V}\mathrm{ar}[X]\,\mathbb{V}\mathrm{ar}[Y]\end{aligned}$$

which is best seen by setting $U = (X-\mathbb{E}[X])^2$ and $V = (Y-\mathbb{E}[Y])^2$.

Example 3.5. Covariances and Correlations of Bivariate Normal Distributions. For every bivariate normal distribution such as the one in Example 3.2, the following holds.

$$\mathbb{E}[X_1X_2] = \rho\sigma_1\sigma_2 + \mu_1\mu_2$$

The demonstration of this result is simplified by a transformation of the random vector $\boldsymbol{x} = (X_1, X_2)$. Define the random vector $\boldsymbol{y} = (Y, Z)$ as:

$$\begin{aligned} Y &= \frac{X_1 - \mu_1}{\sigma_1} \frac{X_2 - \mu_2}{\sigma_2} \\ Z &= \frac{X_1 - \mu_1}{\sigma_1} \end{aligned}$$

the support of $\boldsymbol{y}$ is $\mathbb{Y} = \mathbb{R}^2$ and the transformation is clearly bijective; the inverse transformation is:

$$\begin{aligned} X_1 &= g_1^{-1}(Y, Z) = \sigma_1 Z + \mu_1 \\ X_2 &= g_2^{-1}(Y, Z) = \sigma_2 \frac{Y}{Z} + \mu_2 \end{aligned}$$

whose Jacobian has the following absolute value of the determinant.

$$\left| \det \left(\frac{\partial}{\partial \boldsymbol{y}^{\mathrm{T}}} \boldsymbol{g}^{-1}(y, z) \right) \right| = \left| \det \begin{pmatrix} 0 & \sigma_1 \\ \sigma_2 z^{-1} & -\sigma_2 y z^{-2} \end{pmatrix} \right| = \frac{\sigma_1 \sigma_2}{|z|} = \frac{\sigma_1 \sigma_2}{\sqrt{z^2}}$$

By Theorem 3.1, the joint density function of $\boldsymbol{y} = (Y, Z)$ is:

$$f_{Y,Z}(y, z; \rho) = \frac{1}{2\pi \sqrt{(1 - \rho^2) z^2}} \exp\left(-\frac{z^2 - 2\rho y + y^2 z^{-2}}{2(1 - \rho^2)} \right)$$

and, observing that $z^2 - 2\rho y + y^2 z^{-2} = (1 - \rho^2) z^2 + (y - \rho z^2)^2 z^{-2}$, it is:

$$\begin{aligned} \mathbb{E}[Y] &= \int_{-\infty}^{+\infty} \phi(z) \left[\int_{-\infty}^{+\infty} \frac{y}{\sqrt{2\pi (1 - \rho^2) z^2}} \exp\left(-\frac{(y - \rho z^2)^2}{2(1 - \rho^2) z^2} \right) dy \right] dz \\ &= \rho \int_{-\infty}^{+\infty} z^2 \phi(z)\, dz \\ &= \rho \end{aligned}$$

where $\phi(z)$ is the density function of the standard normal distribution, and where the second line follows from the observation that the inner integral in the first line is the mean of a normally distributed random variable with mean ρz^2 and variance $(1 - \rho^2) z^2$. Therefore, exploiting again the inverse transformation above:

$$\begin{aligned} \mathbb{E}[X_1 X_2] &= \sigma_1 \sigma_2 \mathbb{E}[Y] + \sigma_1 \mu_2 \mathbb{E}[Z] + \sigma_2 \mu_1 \mathbb{E}\left[\frac{Y}{Z} \right] + \mu_1 \mu_2 \\ &= \rho \sigma_1 \sigma_2 + \mu_1 \mu_2 \end{aligned}$$

as postulated, because $\mathbb{E}[Z] = \mathbb{E}\left[\frac{Y}{Z}\right] = 0$ are both expectations of random variables that follow the standard normal distribution.

In light of this result, it is:

$$\mathbb{C}\mathrm{ov}\left[X_1, X_2\right] = \rho\sigma_1\sigma_2$$
$$\mathbb{C}\mathrm{orr}\left[X_1, X_2\right] = \rho$$

hence, parameter ρ has an immediate interpretation as *correlation* (and in fact its range is confined in the $[-1, 1]$ interval).[1] For the sake of illustration, consider the bivariate distribution depicted in Figure 3.1, and suppose to invert the sign of ρ: the result would be as in Figure 3.3 below, which quite clearly manifests – in graphical form – an inversion of the linear relationship that connects X_1 and X_2 in Example 3.5.

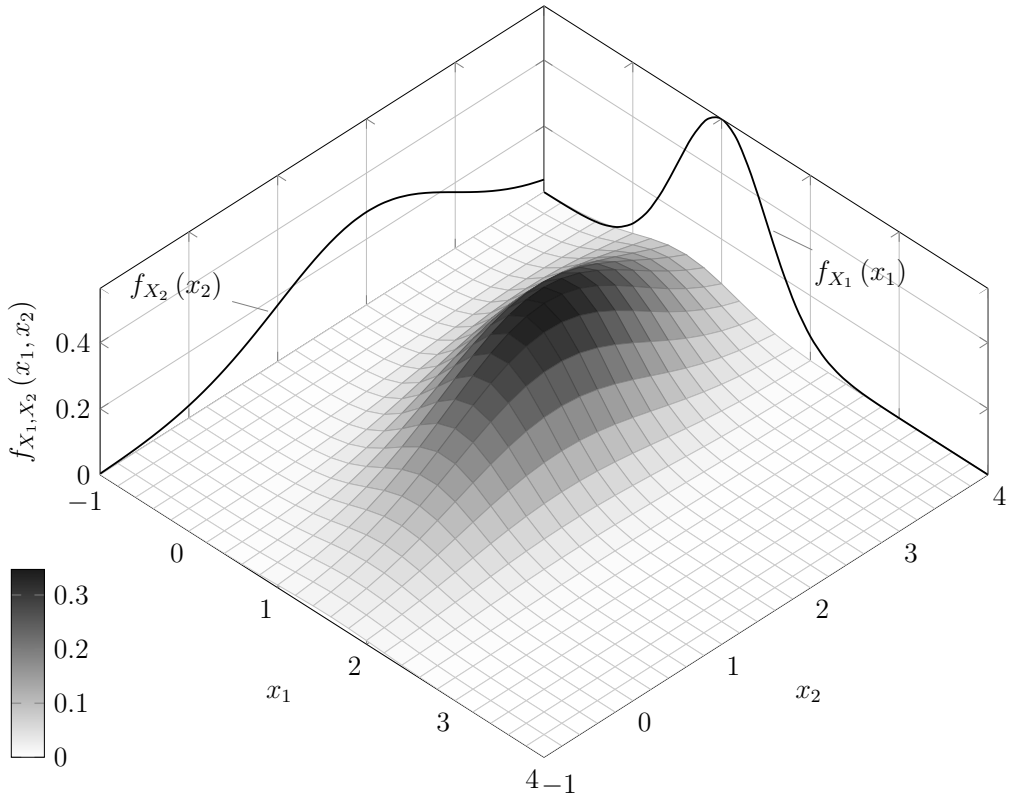


Figure 3.3: Bivariate normal as in Example 3.5, but with $\rho = -.4$.

If the two arguments of a bivariate normal distributions – the two "marginal" random variables – are independent, it must thus be that $\rho = 0$. This case is instead represented in Figure 3.4, which is obtained by again substituting only that parameter. There, no linear dependence between the two random variables can be visually detected, neither positive nor negative.

[1] Note that by construction of the previous transformation, $\mathbb{C}\mathrm{orr}\left[X_1, X_2\right] = \mathbb{E}\left[Y\right]$.

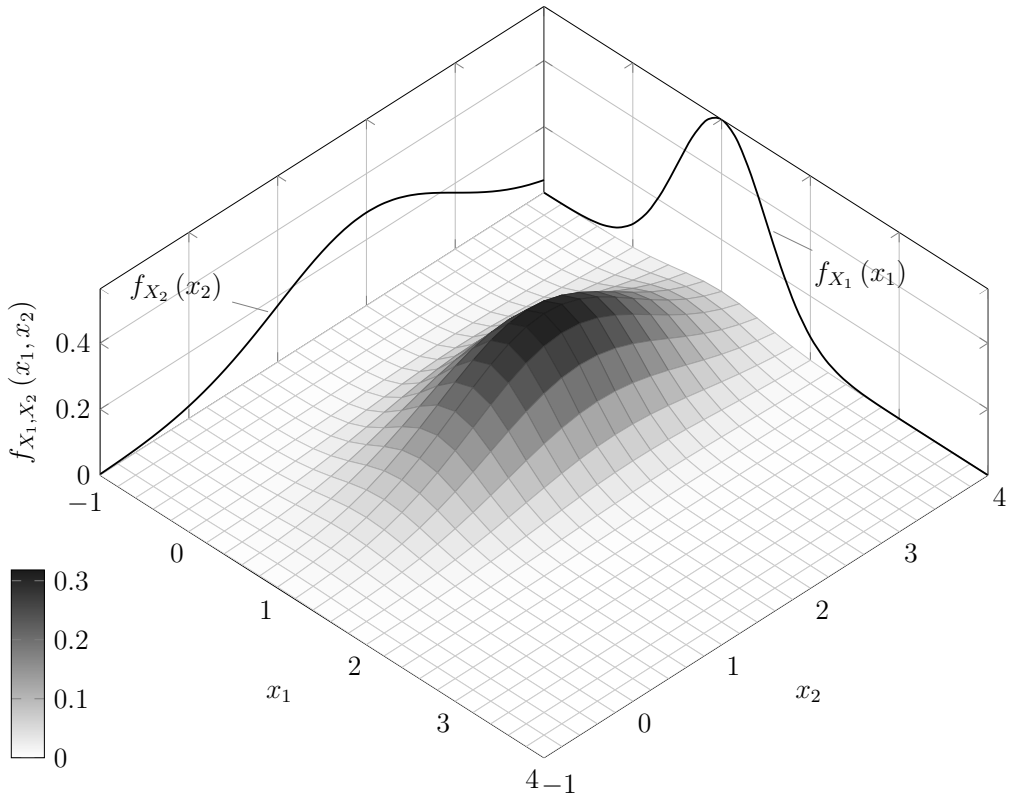


Figure 3.4: Bivariate normal as in Example 3.5, but with $\rho = 0$.

Note that in all these cases the marginal distributions stay the same, since they do not depend upon the parameter ρ (as already mentioned, a unique feature of the bivariate normal distribution is that marginal distributions do not depend on ρ, a unique feature of the bivariate normal). ■

Having characterized all the relevant moments of a given random vector $\boldsymbol{x} = (X_1, \ldots, X_K)$, it is useful to establish some notation for expressing *all* the relevant moments of a certain type of *all* the random variables involved. This is accomplished by means of tools typical of linear algebra, which turn out to be extremely useful to handle moments of multivariate distribution. In particular, the **mean vector** – usually denoted as $\mathbb{E}[\boldsymbol{x}]$ – is the collection of the means of all the random variables in random vector $\boldsymbol{x}$.

$$\mathbb{E}[\boldsymbol{x}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_K] \end{bmatrix}$$

The **variance-covariance matrix** instead, which is commonly denoted by $\mathbb{V}\mathrm{ar}[\boldsymbol{x}]$, collects the variances of each random variable *along* the diagonal, and the covariances between each pair of elements of $\boldsymbol{x}$ *outside* the diagonal.

$$\mathbb{V}\mathrm{ar}[\boldsymbol{x}] = \begin{bmatrix} \mathbb{V}\mathrm{ar}[X_1] & \mathbb{C}\mathrm{ov}[X_1, X_2] & \dots & \mathbb{C}\mathrm{ov}[X_1, X_K] \\ \mathbb{C}\mathrm{ov}[X_2, X_1] & \mathbb{V}\mathrm{ar}[X_2] & \dots & \mathbb{C}\mathrm{ov}[X_2, X_K] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\mathrm{ov}[X_K, X_1] & \mathbb{C}\mathrm{ov}[X_K, X_2] & \dots & \mathbb{V}\mathrm{ar}[X_K] \end{bmatrix}$$

Given how each element is calculated, it is often useful to express the variance of $\boldsymbol{x}$ as the expectation of a specific *random matrix*, as follows.

$$\mathbb{V}\mathrm{ar}[\boldsymbol{x}] = \mathbb{E}\left[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^{\mathrm{T}}\right]$$

It is worth making few observations about the variance-covariance matrix:

- it has dimension $K \times K$ and is symmetric;
- the elements along its diagonal, the variances, are always nonnegative; but those outside the diagonal, the covariances, can be negative;
- in analogy with the univariate case, one can establish the following.

$$\begin{aligned} \mathbb{V}\mathrm{ar}[\boldsymbol{x}] &= \mathbb{E}\left[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^{\mathrm{T}}\right] \\ &= \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right] - \mathbb{E}[\boldsymbol{x}]\,\mathbb{E}[\boldsymbol{x}]^{\mathrm{T}} - \mathbb{E}[\boldsymbol{x}]\,\mathbb{E}[\boldsymbol{x}]^{\mathrm{T}} + \mathbb{E}[\boldsymbol{x}]\,\mathbb{E}[\boldsymbol{x}]^{\mathrm{T}} \\ &= \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right] - \mathbb{E}[\boldsymbol{x}]\,\mathbb{E}[\boldsymbol{x}]^{\mathrm{T}} \end{aligned}$$

- Finally, $\mathbb{V}\mathrm{ar}[\boldsymbol{x}]$ is positive semi-definite, that is, for any non-zero vector $\mathbf{a}$ of length K, the quadratic form $\mathbf{a}^{\mathrm{T}}\,\mathbb{V}\mathrm{ar}[\boldsymbol{x}]\,\mathbf{a} \geq 0$ is nonnegative. This property is demonstrated later while analyzing the moments of linear transformations of random vectors.

Example 3.6. Summarizing the Moments of Bivariate Normal Distributions. All the moments of the bivariate normal distribution from the previous examples can be summarized using the following notation:

$$\boldsymbol{\mu} \equiv \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \mathbb{E}[\boldsymbol{x}]$$

and

$$\boldsymbol{\Sigma} \equiv \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} = \mathbb{V}\mathrm{ar}[\boldsymbol{x}]$$

and it is straightforward to verify that $\boldsymbol{\Sigma}$ complies with the properties of all variance-covariance matrices. If $\boldsymbol{x} = (X_1, X_2)$ follows the bivariate normal distribution, one can write $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. ■

With the aid of some linear algebra, the usual properties of means and variances are generalized to a multivariate environment. Consider a random vector $\boldsymbol{x}$ with mean $\mathbb{E}[\boldsymbol{x}]$ and variance $\mathbb{V}\text{ar}[\boldsymbol{x}]$ in the three following cases.

- **Linear Transformations returning Scalars.** Consider some vector $\mathbf{a} = (a_1, \ldots, a_K)^{\mathrm{T}}$ of length K which, multiplied to $\boldsymbol{x}$, returns the random variable $Y = \mathbf{a}^{\mathrm{T}}\boldsymbol{x}$ as a linear combination. Because expectations are linear operators, the mean of Y is:

$$\begin{aligned}\mathbb{E}[Y] &= \mathbb{E}\left[\mathbf{a}^{\mathrm{T}}\boldsymbol{x}\right] \\ &= \mathbb{E}[a_1 X_1 + \cdots + a_K X_K] \\ &= a_1\,\mathbb{E}[X_1] + \cdots + a_K\,\mathbb{E}[X_K] \\ &= \mathbf{a}^{\mathrm{T}}\,\mathbb{E}[\boldsymbol{x}]\end{aligned}$$

  as for the variance of Y instead:

$$\begin{aligned}\mathbb{V}\text{ar}[Y] &= \mathbb{V}\text{ar}\left[\mathbf{a}^{\mathrm{T}}\boldsymbol{x}\right] \\ &= \mathbb{E}\left[\left(\mathbf{a}^{\mathrm{T}}\boldsymbol{x} - \mathbb{E}\left[\mathbf{a}^{\mathrm{T}}\boldsymbol{x}\right]\right)\left(\mathbf{a}^{\mathrm{T}}\boldsymbol{x} - \mathbb{E}\left[\mathbf{a}^{\mathrm{T}}\boldsymbol{x}\right]\right)^{\mathrm{T}}\right] \\ &= \mathbb{E}\left[\mathbf{a}^{\mathrm{T}}\left(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]\right)\left(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]\right)^{\mathrm{T}}\mathbf{a}\right] \\ &= \mathbf{a}^{\mathrm{T}}\,\mathbb{E}\left[\left(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]\right)\left(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]\right)^{\mathrm{T}}\right]\mathbf{a} \\ &= \mathbf{a}^{\mathrm{T}}\,\mathbb{V}\text{ar}[\boldsymbol{x}]\,\mathbf{a}\end{aligned}$$

  where the last expression is a *quadratic form* that cannot be negative (showing that $\mathbb{V}\text{ar}[\boldsymbol{x}]$ is positive semi-definite); in particular:

$$\mathbb{V}\text{ar}[Y] = \sum_{k=1}^{K}\left[a_k^2\,\mathbb{V}\text{ar}[X_k] + 2\sum_{\ell=1}^{k-1} a_k a_\ell\,\mathbb{C}\text{ov}[X_k, X_\ell]\right]$$

  which exemplifies how it can be easier to work with matrices.

- **Linear Transformations returning Vectors.** Consider now a vector $\mathbf{a}$ of length $J > 1$, a matrix $\mathbf{B}$ of dimension $J \times K$, and the random vector $\boldsymbol{y} = \mathbf{a} + \mathbf{B}\boldsymbol{x} = (Y_1, \ldots, Y_J)^{\mathrm{T}}$ resulting from J different linear combinations of $\boldsymbol{x}$. Since expectations are linear operators, it is:

$$\mathbb{E}[\boldsymbol{y}] = \mathbb{E}[\mathbf{a} + \mathbf{B}\boldsymbol{x}] = \mathbf{a} + \mathbf{B}\,\mathbb{E}[\boldsymbol{x}]$$

  while the variance-covariance matrix of $\boldsymbol{y}$ is given by:

$$\mathbb{V}\text{ar}[\boldsymbol{y}] = \mathbb{V}\text{ar}[\mathbf{a} + \mathbf{B}\boldsymbol{x}] = \mathbf{B}\,\mathbb{V}\text{ar}[\boldsymbol{x}]\,\mathbf{B}^{\mathrm{T}}$$

  noting that, if $\mathbf{b}_i$ and $\mathbf{b}_j$ are the i-th and the j-th rows of $\mathbf{B}$, then the ij-th element of $\mathbb{V}\text{ar}[\boldsymbol{y}]$ equals $\mathbb{C}\text{ov}\left[\mathbf{b}_i^{\mathrm{T}}\boldsymbol{x}, \mathbf{b}_j^{\mathrm{T}}\boldsymbol{x}\right] = \mathbf{b}_i^{\mathrm{T}}\,\mathbb{V}\text{ar}[\boldsymbol{x}]\,\mathbf{b}_j$.

- **Non-linear Transformations of Random Vectors.** Finally look at the case of a J-dimensional non-linear vector-valued function $\boldsymbol{g}(\boldsymbol{x})$. A first-order Taylor expansion of $\boldsymbol{g}(\cdot)$ around $\mathbb{E}[\boldsymbol{x}]$ gives

$$\begin{aligned}\boldsymbol{g}(\boldsymbol{x}) &\approx \boldsymbol{g}(\mathbb{E}[\boldsymbol{x}]) + \frac{\partial}{\partial \boldsymbol{x}^{\mathrm{T}}}\boldsymbol{g}(\mathbb{E}[\boldsymbol{x}])[\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]] \\ &\approx \left[\boldsymbol{g}(\mathbb{E}[\boldsymbol{x}]) - \frac{\partial}{\partial \boldsymbol{x}^{\mathrm{T}}}\boldsymbol{g}(\mathbb{E}[\boldsymbol{x}])\mathbb{E}[\boldsymbol{x}]\right] + \frac{\partial}{\partial \boldsymbol{x}^{\mathrm{T}}}\boldsymbol{g}(\mathbb{E}[\boldsymbol{x}])\boldsymbol{x}\end{aligned}$$

  where $\frac{\partial}{\partial \boldsymbol{x}^{\mathrm{T}}}\boldsymbol{g}(\mathbb{E}[\boldsymbol{x}])$ is a $J \times K$ Jacobian matrix containing, in each jk-th element, the derivative of the j-th equation of $\boldsymbol{g}(\boldsymbol{x})$ with respect to the k-th element of $\boldsymbol{x}$. Hence, in analogy with the univariate case:

$$\mathbb{E}[\boldsymbol{g}(\boldsymbol{x})] \approx \boldsymbol{g}(\mathbb{E}[\boldsymbol{x}])$$

  is a generally poor approximation, but

$$\mathbb{V}\mathrm{ar}[\boldsymbol{g}(\boldsymbol{x})] \approx \left[\frac{\partial}{\partial \boldsymbol{x}^{\mathrm{T}}}\boldsymbol{g}(\mathbb{E}[\boldsymbol{x}])\right]\mathbb{V}\mathrm{ar}[\boldsymbol{x}]\left[\frac{\partial}{\partial \boldsymbol{x}^{\mathrm{T}}}\boldsymbol{g}(\mathbb{E}[\boldsymbol{x}])\right]^{\mathrm{T}}$$

  is a considerably better one.

In a multivariate environment, it is sometimes useful to summarize the covariances between the elements of two random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of length $K_{\boldsymbol{x}}$ and $K_{\boldsymbol{y}}$ respectively. This is best done via a **cross-covariance matrix** of dimension $K_{\boldsymbol{x}} \times K_{\boldsymbol{y}}$ (one should always be careful with terminology and not mistake it for a variance-covariance matrix). Such a matrix collects the covariances between every i-th element of $\boldsymbol{x}$ and every j-th element of $\boldsymbol{y}$ in its ij-th entries, and it is obviously symmetric.

$$\mathbb{C}\mathrm{ov}[\boldsymbol{x}, \boldsymbol{y}] = \begin{bmatrix} \mathbb{C}\mathrm{ov}[X_1, Y_1] & \mathbb{C}\mathrm{ov}[X_1, Y_2] & \dots & \mathbb{C}\mathrm{ov}\left[X_1, Y_{K_{\boldsymbol{y}}}\right] \\ \mathbb{C}\mathrm{ov}[X_2, Y_1] & \mathbb{C}\mathrm{ov}[X_2, Y_2] & \dots & \mathbb{C}\mathrm{ov}\left[X_2, Y_{K_{\boldsymbol{y}}}\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\mathrm{ov}\left[X_{K_{\boldsymbol{x}}}, Y_1\right] & \mathbb{C}\mathrm{ov}\left[X_{K_{\boldsymbol{x}}}, Y_2\right] & \dots & \mathbb{C}\mathrm{ov}\left[X_{K_{\boldsymbol{x}}}, Y_{K_{\boldsymbol{y}}}\right] \end{bmatrix}$$

Like a variance-covariance matrix, a cross-covariance matrix can be recast as the expectation of a random matrix, which can be simplified.

$$\begin{aligned}\mathbb{C}\mathrm{ov}[\boldsymbol{x}, \boldsymbol{y}] &= \mathbb{E}\left[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}])^{\mathrm{T}}\right] \\ &= \mathbb{E}\left[\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}\right] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{y}]^{\mathrm{T}} - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{y}]^{\mathrm{T}} + \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{y}]^{\mathrm{T}} \\ &= \mathbb{E}\left[\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}\right] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{y}]^{\mathrm{T}}\end{aligned}$$

Clearly, the cross-covariance matrix is a collection of zeros if $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent, as $\mathbb{E}\left[\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}\right] = \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{y}]^{\mathrm{T}}$.

Other properties are obtained by extending some previous observations to the case of two random vectors; here, they are summarized briefly.

- If two linearly transformed vectors $\boldsymbol{u} = \mathbf{a}_{\boldsymbol{x}} + \mathbf{B}_{\boldsymbol{x}}\boldsymbol{x}$ and $\boldsymbol{v} = \mathbf{a}_{\boldsymbol{y}} + \mathbf{B}_{\boldsymbol{y}}\boldsymbol{y}$ have length $J_{\boldsymbol{u}}$ and $J_{\boldsymbol{v}}$, their $J_{\boldsymbol{u}} \times J_{\boldsymbol{v}}$ cross-covariance matrix is:
$$\mathbb{C}\text{ov}\left[\boldsymbol{u}, \boldsymbol{v}\right] = \mathbb{C}\text{ov}\left[\mathbf{a}_{\boldsymbol{x}} + \mathbf{B}_{\boldsymbol{x}}\boldsymbol{x}, \mathbf{a}_{\boldsymbol{y}} + \mathbf{B}_{\boldsymbol{y}}\boldsymbol{y}\right] = \mathbf{B}_{\boldsymbol{x}}\,\mathbb{C}\text{ov}\left[\boldsymbol{x}, \boldsymbol{y}\right]\mathbf{B}_{\boldsymbol{y}}^{\mathrm{T}}$$
- however, if $\boldsymbol{u} = \boldsymbol{g}_{\boldsymbol{x}}\left(\boldsymbol{x}\right)$ and $\boldsymbol{v} = \boldsymbol{g}_{\boldsymbol{y}}\left(\boldsymbol{y}\right)$ are obtained via two non-linear transformations, the following approximation can be useful.
$$\mathbb{C}\text{ov}\left[\boldsymbol{u}, \boldsymbol{v}\right] \approx \left[\frac{\partial}{\partial \boldsymbol{x}^{\mathrm{T}}}\boldsymbol{g}_{\boldsymbol{x}}\left(\mathbb{E}\left[\boldsymbol{x}\right]\right)\right]\mathbb{C}\text{ov}\left[\boldsymbol{x}, \boldsymbol{y}\right]\left[\frac{\partial}{\partial \boldsymbol{y}^{\mathrm{T}}}\boldsymbol{g}_{\boldsymbol{y}}\left(\mathbb{E}\left[\boldsymbol{y}\right]\right)\right]^{\mathrm{T}}$$

Furthermore, the cross-covariance matrix of $\boldsymbol{x}$ and $\boldsymbol{y}$ relates with the respective variance-covariance matrices through the following relationship.
$$\mathbb{V}\text{ar}\left[\boldsymbol{x}\right] - \mathbb{C}\text{ov}\left[\boldsymbol{x}, \boldsymbol{y}\right]\left[\mathbb{V}\text{ar}\left[\boldsymbol{y}\right]\right]^{-1}\mathbb{C}\text{ov}\left[\boldsymbol{x}, \boldsymbol{y}\right]^{\mathrm{T}} \geq \mathbf{0}$$
The above inequality is to interpreted in the sense that the matrix on the left-hand side is positive semi-definite. This relationship is tedious to prove, but it essentially represents a bound on the cross-covariance matrix, extending the logic of the intermediate result (3.3) from Theorem 3.4.

3.4 Multivariate Moment Generation

Both the moment-generating and the characteristic functions are easily generalized to a multivariate environment, where they are especially useful for deriving the *distribution* of certain linear combinations of random variables.

Definition 3.13. Moment generating function (multivariate case). Given a random vector $\boldsymbol{x} = (X_1, \ldots, X_K)$ with support $\mathbb{X}$, the *moment generating* function $M_{\boldsymbol{x}}\left(\mathbf{t}\right)$ is defined, for $\mathbf{t} = (t_1, \ldots, t_K) \in \mathbb{R}^K$, as the expectation of the transformation $g\left(X\right) = \exp\left(\mathbf{t}^{\mathrm{T}}\boldsymbol{x}\right)$.
$$M_{\boldsymbol{x}}\left(\mathbf{t}\right) = \mathbb{E}\left[\exp\left(\mathbf{t}^{\mathrm{T}}\boldsymbol{x}\right)\right] = \mathbb{E}\left[\exp\left(\sum_{k=1}^{K} t_k X_k\right)\right]$$

Definition 3.14. Characteristic function (multivariate case). Given a random vector $\boldsymbol{x} = (X_1, \ldots, X_K)$ with support $\mathbb{X}$, the *characteristic* function $\varphi_{\boldsymbol{x}}\left(\mathbf{t}\right)$ is defined, for $\mathbf{t} = (t_1, \ldots, t_K) \in \mathbb{R}^K$, as the expectation of the transformation $g\left(X\right) = \exp\left(i\mathbf{t}^{\mathrm{T}}\boldsymbol{x}\right)$.
$$\varphi_{\boldsymbol{x}}\left(\mathbf{t}\right) = \mathbb{E}\left[\exp\left(i\mathbf{t}^{\mathrm{T}}\boldsymbol{x}\right)\right] = \mathbb{E}\left[\exp\left(i\sum_{k=1}^{K} t_k X_k\right)\right]$$

The r-th centered moments for each k-th element of the random vector $\boldsymbol{x}$ can be calculated in analogy with the univariate case.

$$\mathbb{E}\left[X_k^r\right] = \left.\frac{\partial^r M_{\boldsymbol{x}}\left(\mathbf{t}\right)}{\partial t_k^r}\right|_{\mathbf{t}=\mathbf{0}} = \frac{1}{i^r}\cdot\left.\frac{\partial^r \varphi_{\boldsymbol{x}}\left(\mathbf{t}\right)}{\partial t_k^r}\right|_{\mathbf{t}=\mathbf{0}}$$

Furthermore, the *cross-moments* are obtained, for two integers r and s, as:

$$\mathbb{E}\left[X_k^r X_\ell^s\right] = \left.\frac{\partial^{r+s} M_{\boldsymbol{x}}\left(\mathbf{t}\right)}{\partial t_k^r \partial t_\ell^s}\right|_{\mathbf{t}=\mathbf{0}} = \frac{1}{i^{r+s}}\cdot\left.\frac{\partial^{r+s} \varphi_{\boldsymbol{x}}\left(\mathbf{t}\right)}{\partial t_k^r \partial t_\ell^s}\right|_{\mathbf{t}=\mathbf{0}}$$

which follows since:

$$\frac{\partial^{r+s} M_{\boldsymbol{x}}\left(\mathbf{t}\right)}{\partial t_k^r \partial t_\ell^s} = \mathbb{E}\left[\frac{\partial^{r+s}}{\partial t_k^r \partial t_\ell^s}\,\exp\left(\sum_{k=1}^{K} t_k X_k\right)\right] = \mathbb{E}\left[X_k^r X_\ell^s \exp\left(\sum_{k=1}^{K} t_k X_k\right)\right]$$

and the case of the characteristic function is analogous. This fact allows to calculate covariances using these two important functions.

Example 3.7. Moment Generating Function and Covariance of the Bivariate Normal Distribution. The moment generating function of the bivariate normal distribution is the following.

$$\begin{aligned} M_{X_1,X_2}\left(t_1,t_2\right) &= \mathbb{E}\left[\exp\left(t_1X_1 + t_2X_2\right)\right] \\ &= \exp\left(t_1\mu_1 + t_2\mu_2 + \frac{1}{2}\left(t_1^2\sigma_1^2 + 2t_1t_2\rho\sigma_1\sigma_2 + t_2^2\sigma_2^2\right)\right) \end{aligned}$$

Obtaining this expression while keeping track of all these parameters is not as difficult as it is annoying, therefore a proper and more elegant derivation is postponed to the later, more general analysis of the multivariate normal distribution. Here the point is to show how the covariance between X_1 and X_2 can be derived via the moment generating function. It is not difficult to see that $\mathbb{E}\left[X_k\right] = \mu_k$ and $\mathbb{E}\left[X_k^2\right] = \sigma_k^2 + \mu_k^2$ for $k = 1, 2$, as in the univariate case. As per the first cross-moment, some calculations show that:

$$\begin{aligned} \frac{\partial^2}{\partial t_1 \partial t_2} M_{X_1,X_2}\left(t_1,t_2\right) = \big[&\left(\mu_1 + t_1\sigma_1^2 + t_2\rho\sigma_1\sigma_2\right)\left(\mu_2 + t_2\sigma_2^2 + t_1\rho\sigma_1\sigma_2\right) + \\ &+ \rho\sigma_1\sigma_2\big]\cdot\exp\left(t_1\mu_1 + t_2\mu_2 + \frac{1}{2}\left(t_1^2\sigma_1^2 + 2t_1t_2\rho\sigma_1\sigma_2 + t_2^2\sigma_2^2\right)\right) \end{aligned}$$

and evaluating the above expression for $t_1 = 0$ and $t_2 = 0$ gives:

$$\mathbb{E}\left[X_1X_2\right] = \left.\frac{\partial^2}{\partial t_1 \partial t_2} M_{X_1,X_2}\left(t_1,t_2\right)\right|_{t_1,t_2=0} = \rho\sigma_1\sigma_2 + \mu_1\mu_2$$

which is exactly as derived in Example 3.5. ■

Like in the univariate case, both the moment generating and the characteristic functions uniquely characterize a distribution, but only the more "complex" characteristic function is guaranteed to always exist. In addition, in the multivariate context it is possible to derive some results of extreme importance about independent random variables.

Theorem 3.6. Moment generating and characteristic functions of independent random variables. *If the random variables from a random vector $\boldsymbol{x} = (X_1, \ldots, X_K)$ are pairwise independent, the moment generating function (if it exists) and the characteristic function of $\boldsymbol{x}$ equal respectively the product of the K moment generating functions (if they exist) and the K characteristic functions of the random variables involved.*

$$M_{\boldsymbol{x}}(\mathbf{t}) = \prod_{k=1}^{K} M_{X_k}(t_k)$$
$$\varphi_{\boldsymbol{x}}(\mathbf{t}) = \prod_{k=1}^{K} \varphi_{X_k}(t_k)$$

Proof. This is an application Theorem 3.3 and Theorem 3.5 upon a sequence of K transformed random variables: $\exp(t_1 X_1), \ldots, \exp(t_K X_K)$ which are themselves mutually independent. For moment generating functions:

$$\begin{aligned} M_{\boldsymbol{x}}(\mathbf{t}) = \mathbb{E}\left[\exp\left(\mathbf{t}^{\mathrm{T}}\boldsymbol{x}\right)\right] &= \mathbb{E}\left[\exp\left(\sum_{k=1}^{K} t_k X_k\right)\right] \\ &= \mathbb{E}\left[\prod_{k=1}^{K} \exp(t_k X_k)\right] \\ &= \prod_{k=1}^{K} \mathbb{E}\left[\exp(t_k X_k)\right] \\ &= \prod_{k=1}^{K} M_{X_k}(t_k) \end{aligned}$$

and the case of characteristic functions is analogous. □

Theorem 3.7. Moment generating and characteristic functions of linear combinations of independent random variables. *Consider a random variable Y obtained as the sum of N linearly transformed, pairwise independent random variables $\boldsymbol{x} = (X_1, \ldots, X_N)$:*

$$Y = \sum_{i=1}^{N} (a_i + b_i X_i)$$

where $(a_i, b_i) \in \mathbb{R}^2$ *for* $i = 1, \ldots, N$*. The moment generating and characteristic functions of* Y *are obtained as follows.*

$$M_Y(t) = \exp\left(t \sum_{i=1}^{N} a_i\right) \prod_{i=1}^{N} M_{X_i}(b_i t)$$
$$\varphi_Y(t) = \exp\left(t \sum_{i=1}^{N} a_i\right) \prod_{i=1}^{N} \varphi_{X_i}(b_i t)$$

Proof. For moment generating functions this results is obtained as:

$$\begin{aligned} M_Y(t) = \mathbb{E}[\exp(tY)] &= \mathbb{E}\left[\exp\left(t \cdot \sum_{i=1}^{N} (a_i + b_i X_i)\right)\right] \\ &= \exp\left(t \sum_{i=1}^{N} a_i\right) \mathbb{E}\left[\exp\left(\sum_{i=1}^{N} t b_i X_i\right)\right] \\ &= \exp\left(t \sum_{i=1}^{N} a_i\right) \mathbb{E}\left[\prod_{i=1}^{N} \exp(t b_i X_i)\right] \\ &= \exp\left(t \sum_{i=1}^{N} a_i\right) \prod_{i=1}^{N} \mathbb{E}[\exp(t b_i X_i)] \\ &= \exp\left(t \sum_{i=1}^{N} a_i\right) \prod_{i=1}^{N} M_{X_i}(b_i t) \end{aligned}$$

where the second-to-last line follows from observing that $b_1 X_1, \ldots, b_N X_N$ are N mutually independent random variables (as per Theorem 3.3) as well. The case of characteristic functions is analogous. □

This powerful result often allows to easily obtain the moment distribution of some linear combination of random variables $\boldsymbol{x} = (X_1, \ldots, X_K)$, if their underlying distribution is known and its moment generating or characteristic function is manipulable in such a way that it returns the moment generating function of another known random variable. A list of important cases follows; for all results, the proof is either provided or outlined. Below, the notation X_i indicates one of N random variables (for $i = 1, \ldots, N$) that are all *mutually independent* and follow the indicated distribution.

Observation 3.5. If $X_i \sim \mathrm{Be}(p)$, it is $\sum_{i=1}^{N} X_i \sim \mathrm{BN}(p, N)$.

Proof. If $M_{X_i}(t) = p \exp(t) + (1 - p)$, it suffices to multiply the N identical moment generating functions: $M_{\sum_{i=1}^{N} X_i}(t) = [p \exp(t) + (1 - p)]^N$. □

The next five results are easily demonstrated through the same approach as in the previous observation: that is, by multiplying the moment generating functions of the N specified primitive, independent random variables X_i.

Observation 3.6. If $X_i \sim \mathrm{NB}(p, 1)$, it is $\sum_{i=1}^{N} X_i \sim \mathrm{NB}(p, N)$.

Observation 3.7. If $X_i \sim \mathrm{Pois}(\lambda)$, it is $\sum_{i=1}^{N} X_i \sim \mathrm{Pois}(N\lambda)$.

Observation 3.8. If $X_i \sim \mathrm{Exp}(\lambda)$, it is $\sum_{i=1}^{N} X_i \sim \Gamma(N, \lambda^{-1})$.

Observation 3.9. If $X_i \sim \chi^2(\kappa_i)$, it is $\sum_{i=1}^{N} X_i \sim \chi^2\left(\sum_{i=1}^{N} \kappa_i\right)$.

Observation 3.10. If $X_i \sim \Gamma(\alpha_i, \beta)$, it is $\sum_{i=1}^{N} X_i \sim \Gamma\left(\sum_{i=1}^{N} \alpha_i, \beta\right)$.

Observation 3.11. If $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, for all real a_i, b_i it is as follows.

$$Y = \sum_{i=1}^{N} (a_i + b_i X_i) \sim \mathcal{N}\left(\sum_{i=1}^{N} (a_i + b_i \mu_i), \sum_{i=1}^{N} b_i^2 \sigma_i^2\right)$$

Proof. This requires few steps:

$$\begin{aligned} M_Y(t) &= \exp\left(t \sum_{i=1}^{K} a_i\right) \prod_{i=1}^{K} M_{X_i}(b_i t) \\ &= \exp\left(t \sum_{i=1}^{K} a_i\right) \exp\left(t \sum_{i=1}^{K} b_i \mu_i + t^2 \sum_{i=1}^{K} \frac{b_i^2 \sigma_i^2}{2}\right) \\ &= \exp\left(t \sum_{i=1}^{K} (a_i + b_i \mu_i) + t^2 \sum_{i=1}^{K} \frac{b_i^2 \sigma_i^2}{2}\right) \end{aligned}$$

and the second line is recognized as the moment generating function of a normal distribution with the parameters indicated for Y. □

Observation 3.12. If $\log(X_i) \sim \mathcal{N}(\mu_i, \sigma_i^2)$, for all real a_i, b_i it is as follows.

$$\log\left(\prod_{i=1}^{N} \exp(a_i) X_i^{b_i}\right) \sim \mathcal{N}\left(\sum_{i=1}^{N} (a_i + b_i \mu_i), \sum_{i=1}^{N} b_i^2 \sigma_i^2\right)$$

Proof. Since $\log\left(\prod_{i=1}^{N} \exp(a_i) X_i^{b_i}\right) = \sum_{i=1}^{N} [a_i + b_i \log(X_i)]$, the previous observation extends easily. □

Observation 3.13. If $X_1 \sim \mathrm{Exp}(\lambda_1)$ and $X_2 \sim \mathrm{Exp}(\lambda_2)$, and the two random variables X_1 and X_2 are independent, the random variable Y obtained as $Y = X_1/\lambda_1 - X_2/\lambda_2$ is such that $Y \sim \mathrm{Laplace}(0, 1)$.

Proof. Define the two random variables $W_1 = X_1/\lambda_1$ and $W_2 = -X_2/\lambda_2$, which are obviously independent. By the properties of moment generating functions for linear transformations, the two transformed random variables have moment generating function:

$$M_{W_1}(t) = (1-t)^{-1}$$
$$M_{W_2}(t) = (1+t)^{-1}$$

and since $Y = W_1 + W_2$, the moment generating function of Y is:

$$M_Y(t) = M_{W_1}(t) M_{W_2}(t) = \left(1-t^2\right)^{-1}$$

that is, that of a standard Laplace distribution. □

Observation 3.14. If $X_1 \sim \text{Gumbel}(\mu_1, \sigma)$ and $X_2 \sim \text{Gumbel}(\mu_2, \sigma)$, and the two random variables X_1 and X_2 are independent, the random variable Y obtained as $Y = X_1 - X_2$ is such that $Y \sim \text{Logistic}(\mu_1 - \mu_2, \sigma)$.

Proof. The moment generating function of X_i – for $i = 1, 2$ – is given by $M_{X_i}(t) = \exp(\mu_i t)\,\Gamma(1-\sigma t)$. Similarly, the transformed random variables $W_i = -X_i$ – again for $i = 1, 2$ – have moment generating functions given by $M_{W_i}(t) = \exp(-\mu_i t)\,\Gamma(1+\sigma t)$. It is easy to see that X_1 is independent of W_2 and vice versa. Since $Y = X_1 + W_2$, the moment generating function of Y is therefore obtained as:

$$\begin{aligned} M_Y(t) &= M_{X_1}(t) M_{W_2}(t) \\ &= \exp(\mu_1 t)\,\Gamma(1-\sigma t) \cdot \exp(-\mu_2 t)\,\Gamma(1+\sigma t) \\ &= \exp(\mu_1 t - \mu_2 t)\,\frac{\Gamma(1-\sigma t)\,\Gamma(1+\sigma t)}{\Gamma(2)} \\ &= \exp((\mu_1 - \mu_2)\,t) \cdot \text{B}(1-\sigma t, 1+\sigma t) \end{aligned}$$

which is indeed the moment generating function of the logistic distribution with specified parameters (note that $\Gamma(2) = 1! = 1$). □

3.5 Conditional Distributions

Many conceptual and practical exercises involving multivariate distributions involve "fixing" the value of specific random variables, or "restricting" them to a subset of their support, and analyzing the resulting distribution of the remaining random variables. Such exercises are also called *conditioning* and result in **conditional distributions**; these, together with the **conditional moments** that are derived from them, are of extreme practical importance in statistics and econometrics. For simplicity, only conditioning on specific values (as opposed to subsets of the support) is discussed here.

The ensuing discussion considers two generic random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of dimension $K_{\boldsymbol{x}} \geq 1$ and $K_{\boldsymbol{y}} \geq 1$ (with possibly $K_{\boldsymbol{x}} \neq K_{\boldsymbol{y}}$) and supports $\mathbb{X}$ and $\mathbb{Y}$ respectively, which are expressed as follows.

$$\begin{aligned}\boldsymbol{x} &= \begin{pmatrix} X_1 & \dots & X_{K_{\boldsymbol{x}}} \end{pmatrix}^{\mathrm{T}} \\ \boldsymbol{y} &= \begin{pmatrix} Y_1 & \dots & Y_{K_{\boldsymbol{y}}} \end{pmatrix}^{\mathrm{T}}\end{aligned}$$

In what follows, it is presumed for simplicity's sake that both $\boldsymbol{x}$ and $\boldsymbol{y}$ are either composed by discrete random variables only or by continuous random variables only, but the two types should not coincide between vectors (that is, $\boldsymbol{x}$ might include only discrete random variables and $\boldsymbol{y}$ only continuous ones, or vice versa). The definition of conditional mass or density function is the point of departure of the discussion, as it allows to subsequently define the cumulative conditional distribution.

Definition 3.15. Conditional mass or density function. Consider the combined random vector $(\boldsymbol{x}, \boldsymbol{y})$ with joint mass/density function $f_{\boldsymbol{x},\boldsymbol{y}}(\mathbf{x}, \mathbf{y})$. Suppose that the random vectors $\boldsymbol{x}$ has a probability mass or density function $f_{\boldsymbol{x}}(\mathbf{x})$. The *conditional* mass or density function of $\boldsymbol{y}$, *given* $\boldsymbol{x} = \mathbf{x}$, is defined as follows for all $\mathbf{x} \in \mathbb{X}$:

$$f_{\boldsymbol{y}|\boldsymbol{x}}(\mathbf{y}|\,\boldsymbol{x} = \mathbf{x}) = \frac{f_{\boldsymbol{x},\boldsymbol{y}}(\mathbf{x}, \mathbf{y})}{f_{\boldsymbol{x}}(\mathbf{x})}$$

It is a conditional mass function if all the random variables in $\boldsymbol{y}$ are discrete, and a conditional density function if they are all continuous.

Definition 3.16. Conditional cumulative distribution. Consider the combined random vector $(\boldsymbol{x}, \boldsymbol{y})$ with joint mass/density function $f_{\boldsymbol{x},\boldsymbol{y}}(\mathbf{x}, \mathbf{y})$. The *conditional* cumulative distribution of $\boldsymbol{y}$, *given* $\boldsymbol{x} = \mathbf{x}$ is defined as:

$$F_{\boldsymbol{y}|\boldsymbol{x}}(\mathbf{y}|\,\boldsymbol{x} = \mathbf{x}) = \sum_{\mathbf{t} \in \mathbb{Y}: \mathbf{t} \leq \mathbf{y}} f_{\boldsymbol{y}|\boldsymbol{x}}(\mathbf{t}|\,\boldsymbol{x} = \mathbf{x})$$

if all the random variables in $\boldsymbol{y}$ are discrete, and

$$F_{\boldsymbol{y}|\boldsymbol{x}}(\mathbf{y}|\,\boldsymbol{x} = \mathbf{x}) = \int_{-\infty}^{y_1} \dots \int_{-\infty}^{y_{K_{\boldsymbol{y}}}} f_{\boldsymbol{y}|\boldsymbol{x}}(\mathbf{t}|\,\boldsymbol{x} = \mathbf{x})\, \mathrm{d}\mathbf{t}$$

if all the random variables in $\boldsymbol{y}$ are continuous.

When $\mathbf{x}$ is some generic (undetermined) realization of $\boldsymbol{x}$ – as in the virtual entirety of the conditional probabilities and moments analyzed in these lectures – the conditional density and the conditional cumulative distribution are often written more simply as $f_{\boldsymbol{y}|\boldsymbol{x}}(\mathbf{y}|\,\mathbf{x})$ and $F_{\boldsymbol{y}|\boldsymbol{x}}(\mathbf{y}|\,\mathbf{x})$, respectively.

If both $\boldsymbol{x}$ and $\boldsymbol{y}$ are all-discrete random vectors, the interpretation of a conditional mass function in terms of conditional probability of $\boldsymbol{y}$ *given* $\boldsymbol{x} = \mathbf{x}$ is obvious from the definition. If the two random vectors are instead both all-continuous, the definition restricts the analysis to an hyperplane of the original space considered by the joint distribution, and allows the make conditional probability statements of the following sort.

$$\mathbb{P}\left(a_1 \leq Y_1 \leq b_1 \cap \cdots \cap a_{K_{\boldsymbol{y}}} \leq Y_{K_{\boldsymbol{y}}} \leq b_{K_{\boldsymbol{y}}} \middle| X_1 = x_1 \cap \cdots \cap X_{K_{\boldsymbol{x}}} = x_{K_{\boldsymbol{x}}}\right) = \\ = \int_{a_1}^{b_1} \cdots \int_{a_{K_{\boldsymbol{y}}}}^{b_{K_{\boldsymbol{y}}}} f_{\boldsymbol{y}|\boldsymbol{x}}\left(\mathbf{y} \middle| \boldsymbol{x} = \mathbf{x}\right) \mathrm{d}\mathbf{y}$$

Example 3.8. Conditional Normal Distribution. Consider the bivariate normal distribution from Example 3.2. Some tedious algebraic calculations would reveal that the conditional distribution of one variable, say X_1, conditional on the other variable, say X_2, is obtained as:

$$f_{X_1|X_2}\left(x_1 \middle| x_2\right) = \frac{1}{\sqrt{2\pi\sigma_1^2\left(1-\rho^2\right)}} \exp\left(-\frac{\left[x_1 - \mu_1 - \rho\frac{\sigma_1}{\sigma_2}\left(x_2 - \mu_2\right)\right]^2}{2\sigma_1^2\left(1-\rho^2\right)}\right)$$

where the parameters are dropped in the expression on the left hand side for the sake of brevity. One can observe that the resulting density is that of another univariate normal distribution with different parameters, which can be expressed in compact form as follows for any $X_2 = x_2$.

$$X_1 \middle| X_2 = x_2 \sim \mathcal{N}\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}\left(x_2 - \mu_2\right), \sigma_1^2\left(1-\rho^2\right)\right)$$

Clearly, the expression for the distribution of X_2 conditional on $X_1 = x_2$, for any $x_2 \in \mathbb{R}$, is symmetrical. ■

If $\boldsymbol{y}$ is an all-continuous random vector and $\boldsymbol{x}$ is an all-discrete one, the definition of conditional density function may not be directly applicable – short of resorting to a more general mathematical definition of joint density that allows for discrete mass points. However, the concept is still valid as much as it is useful, and it is best illustrated with an example.

Example 3.9. Conditional height distribution. Remember Example 3.3 about the height distribution with mixed genders. If one aims to describe the density function of height for females only, the appropriate concept is that of a conditional distribution:

$$f_{H|G=1}\left(h \middle| g = 1\right) = \frac{1}{\sigma_F}\phi\left(\frac{h - \mu_F}{\sigma_F}\right)$$

and symmetrically for males. ■

It is important to observe that the concept of conditional distribution is in a sense moot for independent random variables. Suppose, indeed, that every single random variable in $\boldsymbol{y}$ is independent from every single random variable in $\boldsymbol{x}$. This implies that the joint mass/density function of the two random vectors can be described as the product of the two "marginal" joint densities for each separate random vector:

$$f_{\boldsymbol{x},\boldsymbol{y}}(\mathbf{x},\mathbf{y}) = f_{\boldsymbol{x}}(\mathbf{x})\, f_{\boldsymbol{y}}(\mathbf{y})$$

by the definition of conditional mass/density function, this implies that:

$$\begin{aligned} f_{\boldsymbol{y}|\boldsymbol{x}}(\mathbf{y}|\,\mathbf{x}) &= f_{\boldsymbol{y}}(\mathbf{y}) \\ f_{\boldsymbol{x}|\boldsymbol{y}}(\mathbf{x}|\,\mathbf{y}) &= f_{\boldsymbol{x}}(\mathbf{x}) \end{aligned}$$

in other words, the conditional distribution of $\boldsymbol{y}$ *given* $\mathbf{x}$ is equal to the *unconditional* distribution of $\boldsymbol{y}$, and vice versa.

The mean, the variance and other moments are appropriately defined for conditional distributions as well. The **conditional expectation**, which is also called **regression**, is defined for discrete random variables as:

$$\mathbb{E}[\boldsymbol{y}|\,\mathbf{x}] = \sum_{y_1 \in \mathbb{Y}_1} \cdots \sum_{y_K \in \mathbb{Y}_{K_{\boldsymbol{y}}}} \mathbf{y} f_{\boldsymbol{y}|\boldsymbol{x}}(\mathbf{y}|\,\mathbf{x})$$

and for continuous random variables as:

$$\mathbb{E}[\boldsymbol{y}|\,\mathbf{x}] = \int_{\mathbb{Y}_1} \cdots \int_{\mathbb{Y}_{K_{\boldsymbol{y}}}} \mathbf{y} f_{\boldsymbol{y}|\boldsymbol{x}}(\mathbf{y}|\,\mathbf{x})\, \mathrm{d}\mathbf{y}$$

and it is a vector if $K_{\boldsymbol{y}} > 1$. Similarly, the **conditional variance** is, in the discrete case:

$$\mathbb{V}\mathrm{ar}[\boldsymbol{y}|\,\mathbf{x}] = \sum_{y_1 \in \mathbb{Y}_1} \cdots \sum_{y_K \in \mathbb{Y}_{K_{\boldsymbol{y}}}} (\mathbf{y} - \mathbb{E}[\boldsymbol{y}|\,\mathbf{x}])(\mathbf{y} - \mathbb{E}[\boldsymbol{y}|\,\mathbf{x}])^{\mathrm{T}} f_{\boldsymbol{y}|\boldsymbol{x}}(\mathbf{y}|\,\mathbf{x})$$

and in the continuous case:

$$\mathbb{V}\mathrm{ar}[\boldsymbol{y}|\,\mathbf{x}] = \int_{\mathbb{Y}_1} \cdots \int_{\mathbb{Y}_{K_{\boldsymbol{y}}}} (\mathbf{y} - \mathbb{E}[\boldsymbol{y}|\,\mathbf{x}])(\mathbf{y} - \mathbb{E}[\boldsymbol{y}|\,\mathbf{x}])^{\mathrm{T}} f_{\boldsymbol{y}|\boldsymbol{x}}(\mathbf{y}|\,\mathbf{x})\, \mathrm{d}\mathbf{y}$$

and it is a (conditional) variance-covariance matrix if $K_{\boldsymbol{y}} > 1$. As usual with variances, even its conditional version can be expressed in a more compact form and decomposed into simpler uncentered moments.

$$\begin{aligned} \mathbb{V}\mathrm{ar}[\boldsymbol{y}|\,\mathbf{x}] &= \mathbb{E}\left[(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}|\,\mathbf{x}])(\mathbf{y} - \mathbb{E}[\mathbf{y}|\,\mathbf{x}])^{\mathrm{T}} \middle| \mathbf{x} \right] \\ &= \mathbb{E}\left[\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}} \middle| \mathbf{x} \right] - \mathbb{E}[\boldsymbol{y}|\,\mathbf{x}]\, \mathbb{E}[\boldsymbol{y}|\,\mathbf{x}]^{\mathrm{T}} \end{aligned}$$

The results analyzed above relative to the moments of functions of random vectors naturally extend to conditional moments as well.

Example 3.10. Conditional Moments of the Bivariate Normal Distribution. The mean and variance for the conditional distribution $X_1 | X_2$ examined in Example 3.8 are clearly the following.

$$\mathbb{E}[X_1 | X_2 = x_2] = \mu_1 + \rho \frac{\sigma_1}{\sigma_2}(x_2 - \mu_2) \tag{3.4}$$

$$\mathbb{V}\text{ar}[X_1 | X_2 = x_2] = \sigma_1^2 \left(1 - \rho^2\right) \tag{3.5}$$

Symmetric expressions apply to the moments of $X_2 | X_1$. ■

It is quite common to express conditional moments as *functions* of the conditioned variables. In such a case they take names such as **conditional expectation function** (CEF) or **conditional variance function** (CVF), and the conditioned objects are best indicated with the notation for random variables/vectors (e.g. X or $\boldsymbol{y}$) as opposed to their realizations (e.g. x or $\mathbf{y}$). Two results about conditional moment functions turn out to be extremely useful for the sake (among others) of analyzing econometric estimators.

Theorem 3.8. Law of Iterated Expectations. *Given any two random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, it is:*

$$\mathbb{E}[\boldsymbol{y}] = \mathbb{E}_{\boldsymbol{x}}[\mathbb{E}[\boldsymbol{y} | \boldsymbol{x}]]$$

where $\mathbb{E}_{\boldsymbol{x}}[\cdot]$ denotes an expectation taken over the support of $\boldsymbol{x}$.

Proof. In the continuous case, apply the following decomposition:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{y}] &= \int_{\mathbb{X}} \int_{\mathbb{Y}} \mathbf{y} f_{\boldsymbol{x},\boldsymbol{y}}(\mathbf{x}, \mathbf{y}) \, \mathrm{d}\mathbf{y} \mathrm{d}\mathbf{x} \\
&= \int_{\mathbb{X}} \int_{\mathbb{Y}} \mathbf{y} f_{\boldsymbol{y}|\boldsymbol{x}}(\mathbf{y} | \mathbf{x}) f_{\boldsymbol{x}}(\mathbf{x}) \, \mathrm{d}\mathbf{y} \mathrm{d}\mathbf{x} \\
&= \int_{\mathbb{X}} f_{\boldsymbol{x}}(\mathbf{x}) \left[\int_{\mathbb{Y}} \mathbf{y} f_{\boldsymbol{y}|\boldsymbol{x}}(\mathbf{y} | \mathbf{x}) \, \mathrm{d}\mathbf{y} \right] \mathrm{d}\mathbf{x} \\
&= \mathbb{E}_{\boldsymbol{x}}[\mathbb{E}[\boldsymbol{y} | \boldsymbol{x}]]
\end{aligned}$$

and the discrete case is analogous (summations substitute integrals). □

It must be observed that the Law of Iterated Expectations is often applied in situations where $\boldsymbol{y}$ is a function of $\boldsymbol{x}$ itself. In these cases, the evaluation of the inner expectation is often simplified since if $\boldsymbol{x}$ is conditioned upon it is treated as constant, hence it can be taken outside the expectation operator.

Example 3.11. The Bivariate Linear Regression Model. Econometrics revolves around the analysis of statistical models, which are based upon economic theory, that specify the response of certain *endogenous* variables

Y_i to some other *endogenous* variables X_i (or Z_i), where the subscript i denotes the unit of observation of a sample – see lecture 5. These relationships are best framed via conditional distributions and conditional moments. The point of departure for much of econometrics are **linear regression models**, that is linear specifications of the conditional expectation function, like:

$$\mathbb{E}\left[Y_i \middle| X_{1i}, \ldots, X_{Ki}\right] = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{ki}$$

where $(\beta_0, \beta_1, \ldots, \beta_K)$ are the **parameters of interest**.

The simplest linear regression model is the bivariate one, which involves two random variables Y_i and X_i characterized by the following relationship:

$$\mathbb{E}\left[Y_i \middle| X_i\right] = \beta_0 + \beta_1 X_i \tag{3.6}$$

which is equivalently written as follows.

$$\mathbb{E}\left[Y_i - \beta_0 - \beta_1 X_i \middle| X_i\right] = 0 \tag{3.7}$$

The parameter β_0 is typically loaded with the interpretation as the conditional mean of Y_i given $X_i = 0$, or equivalently, as the "constant" coefficient that satisfies the following relationship.

$$\mathbb{E}\left[Y_i - \beta_0 - \beta_1 X_i\right] = 0$$

Consider the following application of the Law of Iterated Expectations: the objective is to analyze the mean of the random variable $X_i\left(Y_i - \beta_0 - \beta_1 X_i\right)$.

$$\begin{aligned}
\mathbb{E}\left[X_i\left(Y_i - \beta_0 - \beta_1 X_i\right)\right] &= \mathbb{E}_X\left[\mathbb{E}\left[X_i\left(Y_i - \beta_0 - \beta_1 X_i\right) \middle| X_i\right]\right] \\
&= \mathbb{E}_X\left[X_i \cdot \mathbb{E}\left[\left(Y_i - \beta_0 - \beta_1 X_i\right) \middle| X_i\right]\right] \\
&= 0
\end{aligned}$$

In the second line X_i can be taken outside the inner expectation operator since there it is treated as a constant; the result is ultimately zero because of (3.7). Consequently, one can establish the following system featuring two equations and two unknown parameters, β_0 and β_1.

$$\mathbb{E}\left[Y_i - \beta_0 - \beta_1 X_i\right] = 0 \tag{3.8}$$

$$\mathbb{E}\left[X_i\left(Y_i - \beta_0 - \beta_1 X_i\right)\right] = 0 \tag{3.9}$$

After some manipulation, the solution for β_0 and β_1 obtains as:

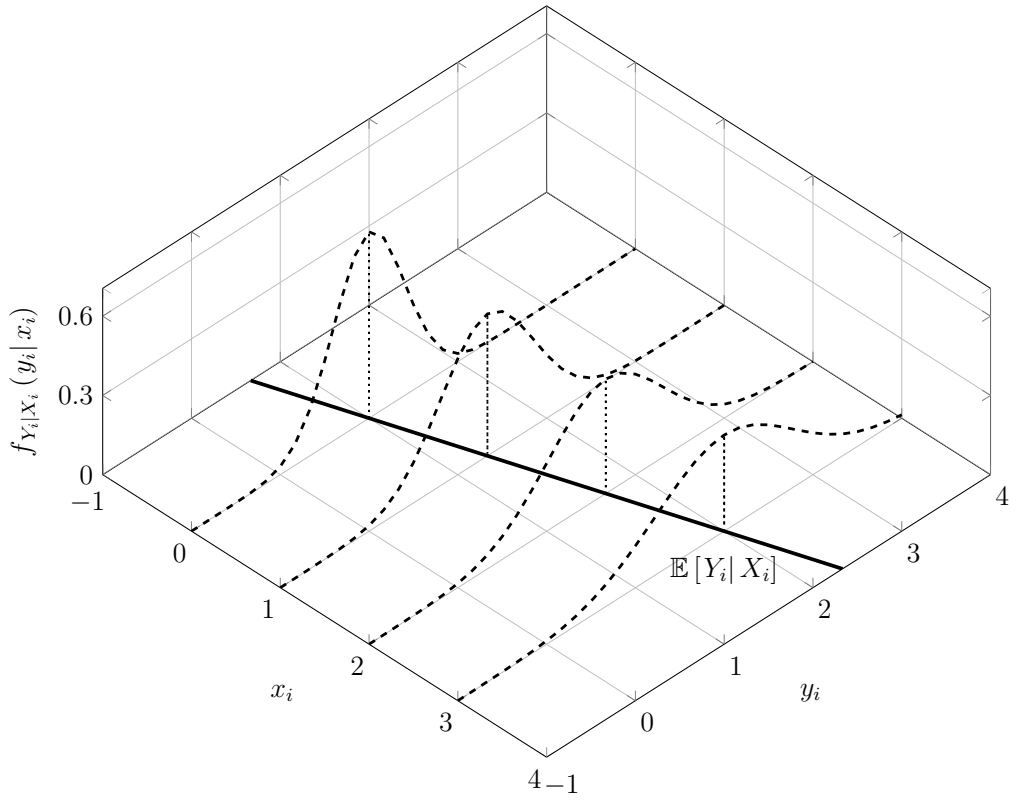
$$\beta_0 = \mathbb{E}\left[Y_i\right] - \frac{\mathbb{C}\mathrm{ov}\left[X_i, Y_i\right]}{\mathbb{V}\mathrm{ar}\left[X_i\right]} \cdot \mathbb{E}\left[X_i\right] \tag{3.10}$$

$$\beta_1 = \frac{\mathbb{C}\mathrm{ov}\left[X_i, Y_i\right]}{\mathbb{V}\mathrm{ar}\left[X_i\right]} \tag{3.11}$$

although (3.10) is more commonly written as $\beta_0 = \mathbb{E}[Y_i] - \beta_1 \mathbb{E}[X_i]$. Note that these are relationships that hold *in the population* as a consequence of the initial assumption (3.6) about linearity of the CEF. Parameter β_1, in particular, is called the **regression slope** and it bears an interpretation as the average response of Y_i to marginal variations in X_i. It is related to the correlation between X_i and Y_i through the following relationship.

$$\beta_1 = \mathbb{C}\text{orr}[X_i, Y_i] \cdot \frac{\sqrt{\mathbb{V}\text{ar}[Y_i]}}{\sqrt{\mathbb{V}\text{ar}[X_i]}}$$

In later lectures, these results are generalized to the multivariate case.



Note: the conditional distribution $Y_i|X_i$ is normal, but with parameters that vary as a function of X_i. Selected density functions of $Y_i|X_i$ are displayed for $x_i = \{0, 1, 2, 3\}$.

Figure 3.5: A bivariate linear regression model for $\beta_0 = 1$ and $\beta_1 = 1/3$

Figure 3.5 depicts an example of a bivariate linear regression model where the conditional distribution of $Y_i|X_i$ is normal; in this case, the parameters of the normal distribution vary visibly along the support of X_i. Note that the conditional distribution of $Y_i|X_i$ can be left unrestricted, so long as the conditional moment $\mathbb{E}[Y_i|X_i]$ exists and complies with (3.6). ■

The other important result about conditional moment functions follows.

Theorem 3.9. Law of Total Variance. *This result, otherwise known as the **variance decomposition**, states that given any two random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$:*

$$\mathbb{V}\mathrm{ar}\left[\boldsymbol{y}\right] = \mathbb{V}\mathrm{ar}_{\boldsymbol{x}}\left[\mathbb{E}\left[\boldsymbol{y}\middle|\boldsymbol{x}\right]\right] + \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{V}\mathrm{ar}\left[\boldsymbol{y}\middle|\boldsymbol{x}\right]\right]$$

where $\mathbb{V}\mathrm{ar}_{\boldsymbol{x}}\left[\cdot\right]$ denotes that the variance is obtained by summing/integrating over the support of $\boldsymbol{x}$; while $\mathbb{E}_{\boldsymbol{x}}\left[\cdot\right]$ represents a summation or an integral applied to a matrix and returning, element-by-element, yet another matrix.

Proof. The proof repeatedly applies the Law of Iterated Expectations.

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left[\boldsymbol{y}\right] &= \mathbb{E}\left[\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}\right] - \mathbb{E}\left[\boldsymbol{y}\right]\mathbb{E}\left[\boldsymbol{y}\right]^{\mathrm{T}} \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}\left[\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}\middle|\boldsymbol{x}\right]\right] - \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}\left[\boldsymbol{y}\middle|\boldsymbol{x}\right]\right]\left[\mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}\left[\boldsymbol{y}\middle|\boldsymbol{x}\right]\right]\right]^{\mathrm{T}} \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}\left[\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}\middle|\boldsymbol{x}\right] - \mathbb{E}\left[\boldsymbol{y}\middle|\boldsymbol{x}\right]\mathbb{E}\left[\boldsymbol{y}\middle|\boldsymbol{x}\right]^{\mathrm{T}}\middle|\boldsymbol{x}\right] \\
&\quad + \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}\left[\boldsymbol{y}\middle|\boldsymbol{x}\right]\mathbb{E}\left[\boldsymbol{y}\middle|\boldsymbol{x}\right]^{\mathrm{T}}\middle|\boldsymbol{x}\right] - \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}\left[\boldsymbol{y}\middle|\boldsymbol{x}\right]\right]\left[\mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}\left[\boldsymbol{y}\middle|\boldsymbol{x}\right]\right]\right]^{\mathrm{T}} \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{V}\mathrm{ar}\left[\boldsymbol{y}\middle|\boldsymbol{x}\right]\right] + \mathbb{V}\mathrm{ar}_{\boldsymbol{x}}\left[\mathbb{E}\left[\boldsymbol{y}\middle|\boldsymbol{x}\right]\right]
\end{aligned}$$

The result follows by adding and subtracting $\mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}\left[\boldsymbol{y}\middle|\boldsymbol{x}\right]\mathbb{E}\left[\boldsymbol{y}\middle|\boldsymbol{x}\right]^{\mathrm{T}}\middle|\boldsymbol{x}\right]$. □

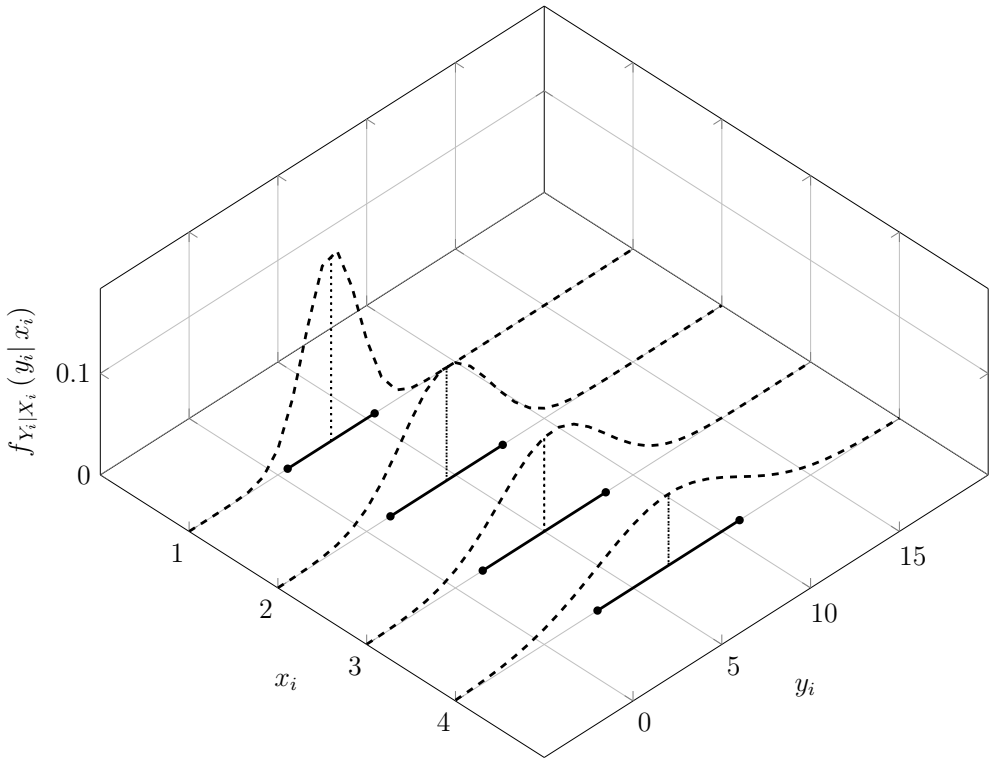
Example 3.12. Variance decomposition. Suppose that a researcher is examining how a certain continuous random variable Y (say, the logarithm of income) differs across four specific groups in the population. These groups are coded as $X = 1, 2, 3, 4$ – X is clearly a discrete random variable – and may correspond to, say, different ethnicities or combinations of two binary characteristics such as age and education. The researcher finds out that the log-income distribution, conditional on each demographic group, is normal:

$$\begin{aligned}
Y\middle|X = 1 &\sim \mathcal{N}\left(3, 1.5\right) \\
Y\middle|X = 2 &\sim \mathcal{N}\left(4.5, 2.5\right) \\
Y\middle|X = 3 &\sim \mathcal{N}\left(5.5, 3\right) \\
Y\middle|X = 4 &\sim \mathcal{N}\left(7, 4\right)
\end{aligned}$$

therefore, it appears that to a higher conditional mean corresponds a higher conditional variance. Furthermore, each group is found with equal probability in the population, meaning that $\mathbb{P}\left(X = x\right) = 0.25$ for $x = 1, 2, 3, 4$. Hence, by the Law of Iterated Expectations, the grand mean of Y in the population can be calculated as follows.

$$\mathbb{E}\left[Y\right] = \mathbb{E}_X\left[\mathbb{E}\left[Y\middle|X\right]\right] = \frac{1}{4}\sum_{x=1}^{4}\mathbb{E}\left[Y\middle|X = x\right] = 5$$

The researcher, however, seems more intent into analyzing how the variation of Y differs across groups with respect to the variation of Y in the population as a whole. The interest of the researcher can be, for example, to gauge to what extent the resentment against inequality interacts with issues about ethnicity. The concern of the researcher is all but enhanced after having visualized a plot about the four conditional distributions, which is reported in Figure 3.6 below. The figure shows how not only the means of Y markedly differ across the four groups, but also the variation of log-income is quite heterogeneous. Problems like this are the domain of the so-called *analysis of the variance*, a set of statistical techniques for assessing differences about certain characteristics between groups of a population.



Note: the figure displays the four conditionally normal distributions of $Y_i|X_i$ for every $X_i = 1, 2, 3, 4$. The four straight lines delimited by circles and placed beneath each density function denote the range of all values of $Y_i|X_i$ within two standard deviations below or above each group's conditional mean.

Figure 3.6: Hypothetical log-income distribution by four groups

To the rescue of the researcher comes the Law of Total Variance, which in this bivariate case reads as follows.

$$\mathbb{V}\text{ar}[Y] = \mathbb{V}\text{ar}_X[\mathbb{E}[Y|X]] + \mathbb{E}_X[\mathbb{V}\text{ar}[Y|X]]$$

The first component on the right hand side, $\mathbb{V}\text{ar}_X\left[\mathbb{E}\left[Y\middle|X\right]\right]$ is the so-called **between group variation** and is interpreted as the "variance of the conditional means" – that is, how much do the four groups differ on average (a more direct measure of cross-group inequality). Here, this is:

$$\mathbb{V}\text{ar}_X\left[\mathbb{E}\left[Y\middle|X\right]\right] = \frac{1}{4}\sum_{x=1}^{4}\left(\mathbb{E}\left[Y\middle|X=x\right] - \mathbb{E}_X\left[\mathbb{E}\left[Y\middle|X\right]\right]\right)^2 = 2.125$$

where $\mathbb{E}_X\left[\mathbb{E}\left[Y\middle|X\right]\right] = 5$ – as previously calculated by averaging the conditional means over X. The second component on the right hand side of the variance decomposition, $\mathbb{E}_X\left[\mathbb{V}\text{ar}\left[Y\middle|X\right]\right]$ is called **within group variation** and is interpreted as the "mean of the conditional variances," identifying that part of the variance which depends on the relative position all individuals against their own group average (that is, their conditional mean) over *all* the groups. In this case this is:

$$\mathbb{E}_X\left[\mathbb{V}\text{ar}\left[Y\middle|X\right]\right] = \frac{1}{4}\sum_{x=1}^{4}\mathbb{V}\text{ar}\left[Y\middle|X=x\right] = 2.75$$

with about the same magnitude as the between group variation. Therefore, the researcher concludes that the overall inequality has a very strong group component, which is likely to bear social and political consequences. ■

3.6 Two Multivariate Distributions

This lecture is concluded with the discussion of two important multivariate distributions, a discrete and a continuous one, which play a special role in both statistics and econometrics. The discrete one is the *multinomial* distribution, a generalization of the binomial distribution. The continuous one cannot be anything but the generalized *multivariate normal* distribution, whose relationship with the univariate case is obvious from its name.

Multinomial Distribution

The multinomial distribution describes a variation of the binomial experiment with many, mutually exclusive, alternatives. Specifically, suppose one is making n draws, and each of these can end up with the realization of one and only one event between $K \geq 2$ that are possible. All these events have probability $p_k \in [0,1]$ to happen (for $k = 1,\ldots,K$), with $\sum_{k=1}^{K} p_k = 1$. After the n draws, the result is a list of *success counts* for each alternative,

a list that one could write as $\boldsymbol{x} = (X_1, \ldots, X_K)$ with $X_k \in \{0, 1, \ldots, n\}$ for $k = 1, \ldots, K$. The support of this random vector is thus the following set.

$$\mathbb{X} = \left\{ \mathbf{x} = (x_1, \ldots, x_K) \in \{0, 1, \ldots, n\}^n : \sum_{k=1}^{K} x_k = n \right\}$$

The probability mass function of this particular random vector is given by:

$$f_{\boldsymbol{x}}(x_1, \ldots, x_K; n, p_1, \ldots, p_K) = \frac{n!}{\prod_{k=1}^{K} x_k!} \prod_{k=1}^{K} p_k^{x_k} \tag{3.12}$$

where the *multinomial coefficient* $n! \cdot \prod_{k=1}^{K} (x_k!)^{-1}$ counts, in what appears to be an extension of the binomial coefficient, the number of realizations containing exactly $(x_1, \ldots, x_K)$ successes for each alternative out of n draws. The cumulative distribution clearly sums the mass function over points in the support as follows, for $\mathbf{t} = (t_1, \ldots, t_K)$.

$$F_{\boldsymbol{x}}(x_1, \ldots, x_K; n, p_1, \ldots, p_K) = \sum_{\mathbf{t} \in \mathbb{X}: \mathbf{t} \leq \mathbf{x}} \frac{n!}{\prod_{k=1}^{K} t_k!} \prod_{k=1}^{K} p_k^{t_k} \tag{3.13}$$

This distribution draws its name from the multinomial theorem, which is useful to analyze it. It helps show that the total probability mass equals 1:

$$\begin{aligned}
\mathbb{P}(\boldsymbol{x} \in \mathbb{X}) &= \sum_{\mathbf{x} \in \mathbb{X}} f_{\boldsymbol{x}}(x_1, \ldots, x_K; n, p_1, \ldots, p_K) \\
&= \sum_{\mathbf{x} \in \mathbb{X}} \frac{n!}{\prod_{k=1}^{K} x_k!} \prod_{k=1}^{K} p_k^{x_k} \\
&= \left(\sum_{k=1}^{K} p_k \right)^n \\
&= 1
\end{aligned}$$

as well as the calculation of the moment generating function:

$$\begin{aligned}
M_{\boldsymbol{x}}(t_1, \ldots, t_K) &= \sum_{\mathbf{x} \in \mathbb{X}} \exp\left(\sum_{k=1}^{K} t_k x_k \right) \cdot f_{\boldsymbol{x}}(x_1, \ldots, x_K; n, p_1, \ldots, p_K) \\
&= \sum_{\mathbf{x} \in \mathbb{X}} \frac{n!}{\prod_{k=1}^{K} x_k!} \prod_{k=1}^{K} (p_k \cdot \exp(t_k))^{x_k} \\
&= \left(\sum_{k=1}^{K} p_k \cdot \exp(t_k) \right)^n
\end{aligned} \tag{3.14}$$

in both cases, the multinomial theorem is applied at the third line.

Through appropriate differentiation of the moment generating function and additional calculations, one can see that for all $k = 1, \ldots, K$:

$$\mathbb{E}[X_k] = np_k \tag{3.15}$$
$$\mathbb{V}\text{ar}[X_k] = np_k(1 - p_k) \tag{3.16}$$

and for all $k, \ell = 1, \ldots, K$:

$$\mathbb{C}\text{ov}[X_k, X_\ell] = -np_k p_\ell \tag{3.17}$$

and the covariance is always negative because an increasing number of successes for one alternative implies a decreasing number for another alternative. By writing the vector of probabilities as:

$$\mathbf{p} \equiv \begin{pmatrix} p_1 \\ \vdots \\ p_K \end{pmatrix}$$

hence one can write the mean vector and the variance-covariance matrix for this distribution in more compact and elegant form as follows.

$$\mathbb{E}[\boldsymbol{x}] = n\mathbf{p} \tag{3.18}$$
$$\mathbb{V}\text{ar}[\boldsymbol{x}] = n\left(\text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^{\mathrm{T}}\right) \tag{3.19}$$

Multivariate Normal Distribution

A generic random vector $\boldsymbol{x} = (X_1, \ldots, X_K)$ of length K whose support is $\mathbb{X} = \mathbb{R}^K$ is said to follow the *multivariate* normal distribution if its joint probability density function is given by:

$$f_{\boldsymbol{x}}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^K |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \tag{3.20}$$

with two collections of parameters: a vector $\boldsymbol{\mu}$ of length K and a symmetric, positive semi-definite matrix $\boldsymbol{\Sigma}$ of dimension $K \times K$ and full rank:

$$\boldsymbol{\mu} \equiv \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_K \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} \equiv \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1K} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{K1} & \sigma_{K2} & \ldots & \sigma_{KK} \end{pmatrix}$$

with $\sigma_{ij} = \sigma_{ji}$ for $i, j = 1, \ldots, K$ and where $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$.

The expression

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

typically indicates that the random vector $\boldsymbol{x}$ follows the multivariate normal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. A particular case of the multivariate normal distribution is the *standardized* one, with $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}$. If a random vector $\boldsymbol{z}$ follows the standard multivariate normal distribution, this is written as follows.

$$\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Note that since $\boldsymbol{\Sigma}$ is symmetric and positive semi-definite, a Cholesky decomposition can always be applied to it so to find some matrix $\boldsymbol{\Sigma}^{\frac{1}{2}}$ such that $\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-\frac{1}{2}} = \mathbf{I}$. Therefore, a random vector that follows a generic normal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is always related to a random vector $\boldsymbol{z}$ that follows the standard multivariate normal via the transformations:

$$\begin{aligned} \boldsymbol{z} &= \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{x} - \boldsymbol{\mu}) \\ \boldsymbol{x} &= \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{z} + \boldsymbol{\mu} \end{aligned}$$

which is analogous to the univariate case; also observe that the two transformations together comply with Theorem 3.1 about the transformation of continuous random vectors.

As usual, the cumulative distribution of the normal distribution lacks a closed form, hence it must be expressed as a multiple integral.

$$\begin{aligned} &F_{\boldsymbol{x}}(x_1, \ldots, x_K; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \\ &\quad = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_K} \frac{1}{\sqrt{(2\pi)^K |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{t} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\mathbf{t} - \boldsymbol{\mu})\right) \mathrm{d}\mathbf{t} \end{aligned} \quad (3.21)$$

Like in the univariate case, it is not immediate to show that the joint density function integrates to one; the demonstration is a tedious extension of the one from Lecture 2 for $K = 1$. However, obtaining the moment generating function is again a relatively simpler task if one starts from the standardized case, $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$\begin{aligned} M_{\boldsymbol{z}}(\mathbf{t}) &= \int_{\mathbb{R}^K} \exp\left(\mathbf{t}^{\mathrm{T}}\mathbf{z}\right) \frac{1}{\sqrt{(2\pi)^K}} \exp\left(-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\mathbf{z}\right) \mathrm{d}\mathbf{z} \\ &= \exp\left(\frac{\mathbf{t}^{\mathrm{T}}\mathbf{t}}{2}\right) \int_{\mathbb{R}^K} \frac{1}{\sqrt{(2\pi)^K}} \exp\left(-\frac{(\mathbf{z} - \mathbf{t})^{\mathrm{T}}(\mathbf{z} - \mathbf{t})}{2}\right) \mathrm{d}\mathbf{z} \\ &= \exp\left(\frac{\mathbf{t}^{\mathrm{T}}\mathbf{t}}{2}\right) \end{aligned}$$

where the integral in the second line equates that of an another multivariate normal distribution with $\boldsymbol{\mu} = \mathbf{t}$ and $\boldsymbol{\Sigma} = \mathbf{I}$, hence it integrates to one. To obtain the general expression of the moment generating function, note that:

$$\begin{aligned} M_{\boldsymbol{x}}(\mathbf{t}) &= \mathbb{E}\left[\exp\left(\mathbf{t}^{\mathrm{T}}\boldsymbol{x}\right)\right] \\ &= \mathbb{E}\left[\exp\left(\mathbf{t}^{\mathrm{T}}\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{z} + \boldsymbol{\mu}\right)\right)\right] \\ &= \exp\left(\mathbf{t}^{\mathrm{T}}\boldsymbol{\mu}\right) \cdot \mathbb{E}\left[\exp\left(\mathbf{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{z}\right)\right] \\ &= \exp\left(\mathbf{t}^{\mathrm{T}}\boldsymbol{\mu} + \frac{\mathbf{t}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{t}}{2}\right) \end{aligned}$$

since the expectation in the third line corresponds to the definition of moment generating function for the standardized normal distribution but with a rescaled argument: $\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{t}$ instead of $\mathbf{t}$ (recall that $\boldsymbol{\Sigma}^{\frac{1}{2}}$ is symmetric).

By analyzing the above moment generating function, one can conclude the following about the moments of a multivariate normal distribution.

$$\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu} \tag{3.22}$$

$$\mathbb{V}\mathrm{ar}[\boldsymbol{x}] = \boldsymbol{\Sigma} \tag{3.23}$$

Note that this is a different parametrization to that of the particular bivariate case (see e.g. Example 3.6); there, σ_1 and σ_2 correspond for convenience to *standard deviations* (square roots of variances) instead of just variances. Here instead, the variances are denoted as $\mathbb{V}\mathrm{ar}[X_k] = \sigma_{kk}$ for $k = 1, \ldots, K$ and the covariances as $\mathbb{C}\mathrm{ov}[X_k, X_\ell] = \sigma_{k\ell}$ for $k, \ell = 1, \ldots, K$ and $k \neq \ell$. An additional observation about the moment generating function is that for any J-dimensional linear combination of $\boldsymbol{x}$ – write it $\boldsymbol{y} = \mathbf{a} + \mathbf{B}\boldsymbol{x}$ where $\mathbf{a}$ is a vector of length J and $\mathbf{B}$ is a $J \times K$ matrix – one can obtain the moment generating function for $\boldsymbol{y}$: $M_{\boldsymbol{y}}(\mathbf{t})$ (where now $\mathbf{t}$ has length J) following the same procedure as above:

$$\begin{aligned} M_{\boldsymbol{y}}(\mathbf{t}) &= \mathbb{E}\left[\exp\left(\mathbf{t}^{\mathrm{T}}\boldsymbol{y}\right)\right] \\ &= \mathbb{E}\left[\exp\left(\mathbf{t}^{\mathrm{T}}\left(\mathbf{a} + \mathbf{B}\boldsymbol{x}\right)\right)\right] \\ &= \exp\left(\mathbf{t}^{\mathrm{T}}\mathbf{a}\right) \cdot \mathbb{E}\left[\exp\left(\mathbf{t}^{\mathrm{T}}\mathbf{B}\boldsymbol{x}\right)\right] \\ &= \exp\left(\mathbf{t}^{\mathrm{T}}\left(\mathbf{B}\boldsymbol{\mu} + \mathbf{a}\right) + \frac{\mathbf{t}^{\mathrm{T}}\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^{\mathrm{T}}\mathbf{t}}{2}\right) \end{aligned}$$

which is the moment generating function of another multivariate normal distribution, that is, $\boldsymbol{y} \sim \mathcal{N}\left(\mathbf{B}\boldsymbol{\mu} + \mathbf{a}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^{\mathrm{T}}\right)$. In plain words, *any collection of linear combinations of some possibly dependent normal distributions itself follows a multivariate normal distribution.* This result is frequently applied to derive the distribution of just one single linear combination ($J = 1$).

A final observation concerns the marginal and conditional distributions obtained from multivariate normal distributions. Suppose that $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$ can be split into two subvectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ of length K_1 and K_2 respectively, with $K_1 + K_2 = K$. Partition the original collection of parameters as follows:

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

where $\boldsymbol{\mu}_1$ is a vector of length K_1, $\boldsymbol{\mu}_2$ one of length K_2, $\boldsymbol{\Sigma}_{11}$ a symmetric $K_1 \times K_1$ matrix, $\boldsymbol{\Sigma}_{22}$ a symmetric $K_2 \times K_2$ matrix, while $\boldsymbol{\Sigma}_{12}$ and $\boldsymbol{\Sigma}_{21}$ are two matrices, one being the transpose of the other, of dimension $K_1 \times K_2$ and $K_2 \times K_1$ respectively. By the properties of partioned inverse matrices:

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \overline{\boldsymbol{\Sigma}}_1^{-1} & -\overline{\boldsymbol{\Sigma}}_1^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ -\overline{\boldsymbol{\Sigma}}_2^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} & \overline{\boldsymbol{\Sigma}}_2^{-1} \end{pmatrix}$$

where:

$$\overline{\boldsymbol{\Sigma}}_1 \equiv \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$
$$\overline{\boldsymbol{\Sigma}}_2 \equiv \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$$

and:

$$|\boldsymbol{\Sigma}| = \left|\overline{\boldsymbol{\Sigma}}_1\right| \cdot |\boldsymbol{\Sigma}_{22}| = \left|\overline{\boldsymbol{\Sigma}}_2\right| \cdot |\boldsymbol{\Sigma}_{11}|$$

relating the determinant of $\boldsymbol{\Sigma}$ to those of the matrices expressing its partitioned inverse. With all this in mind, (3.20) can be rewritten as:

$$f_{\boldsymbol{x}}\left(\mathbf{x}_1, \mathbf{x}_2; \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{12}, \boldsymbol{\Sigma}_{21}, \boldsymbol{\Sigma}_{22}\right) = \frac{1}{\sqrt{(2\pi)^K \left|\overline{\boldsymbol{\Sigma}}_1\right| \cdot |\boldsymbol{\Sigma}_{22}|}} \times$$
$$\times \exp\left(\frac{1}{2}\left(\mathbf{x}_1 - \boldsymbol{\mu}_1\right)^{\mathrm{T}} \overline{\boldsymbol{\Sigma}}_1^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right) - \frac{1}{2}\left(\mathbf{x}_1 - \boldsymbol{\mu}_1\right)^{\mathrm{T}} \overline{\boldsymbol{\Sigma}}_1^{-1}\left(\mathbf{x}_1 - \boldsymbol{\mu}_1\right) + \right.$$
$$\left. + \frac{1}{2}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right)^{\mathrm{T}} \overline{\boldsymbol{\Sigma}}_2^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\left(\mathbf{x}_1 - \boldsymbol{\mu}_1\right) - \frac{1}{2}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right)^{\mathrm{T}} \overline{\boldsymbol{\Sigma}}_2^{-1}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right)\right)$$

or alternatively, by inverting the '1' and '2' subscripts. The above expression of the joint density and its symmetric version can be exploited to show, after more calculations, that the "marginalized" distributions for $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are:

$$\boldsymbol{x}_1 \sim \mathcal{N}\left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}\right)$$
$$\boldsymbol{x}_2 \sim \mathcal{N}\left(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}\right)$$

while the conditional ones, for one subvector given the other, are:

$$\boldsymbol{x}_1 | \, \boldsymbol{x}_2 \sim \mathcal{N}\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\boldsymbol{x}_2 - \boldsymbol{\mu}_2\right), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)$$
$$\boldsymbol{x}_2 | \, \boldsymbol{x}_1 \sim \mathcal{N}\left(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\left(\boldsymbol{x}_1 - \boldsymbol{\mu}_1\right), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\right)$$

generalizing the observations already made for the simpler bivariate case.

Lecture 4

Samples and Statistics

This lecture introduces some core concepts associated with the practice of statistical analysis: samples and statistics that are calculated from samples. A short introduction to the general notion of statistical sample is followed by a focused analysis of an important special case: that of random sample, with special emphasis on random samples drawn from the normal distribution. In developing these concepts, the first half of the lecture introduces properties of common sample statistics, while the second half is devoted to the analysis of sample statistics with specific properties: order and sufficient statistics.

4.1 Random Samples

The practice of all statistical analysis is based on the analysis of some *data* that are extracted from some population of interest. All data are organized by *samples*, that is collections of information associated with different units of analysis of the population (individuals, firms, polities, etc.). The practice of statistical analysis requires to make probabilistic evaluations about data samples, which must thus be connected to probability theory. The first step in this direction is a probabilistic definition of a sample.

Definition 4.1. Sample. A *sample* is a collection $\{\mathbf{x}_i\}_{i=1}^N$ of *realizations* of some N random vectors $\{\boldsymbol{x}_i\}_{i=1}^N$ associated with some population of interest. Each unit of such a population is typically named a *unit of observation* and its associated realization $\mathbf{x}_i$ is identified by a unique subscript i.

This definition is already general enough that it allows for the observation of multiple variables for every unit of analysis; in the simpler cases when each of these is associated with the realization of just one random variable, a sample is written as $\{x_i\}_{i=1}^N$. Conversely, the definition can be extended to

realizations drawn from random matrices, in which case a sample is written as $\{\mathbf{X}_i\}_{i=1}^N$. In all these cases, the following terminology applies.

Definition 4.2. Sample size. The dimension N of a sample is called *size*.

A sample can be called in various ways, depending on the commonalities between the random variables from which the observations are drawn.

Definition 4.3. Random sample. A sample is said to be *random* if all the realizations that compose it are drawn from **independent and identically distributed** (**i.i.d.**) random vectors $\{\boldsymbol{x}_i\}_{i=1}^N$ (or variables, or matrices).

The hypothesis motivating random samples is that all the realizations are obtained from a given population characterized by some joint probability distribution expressed by some random vector $\boldsymbol{x}$. A sample complies with this framework if, for example, each realization is obtained by extracting every unit of observation from some specific population through a protocol that assigns to all such units the same probability of being drawn into the sample, a process known as *sampling with replacement* (this name derives from sampling protocols applied to finite populations where a unit is allowed to produce multiple realizations x_i). Conversely, other protocols like such as *sampling without replacement*, where every realization is drawn sequentially from a population and is not allowed to be extracted again, do not comply with the random sample framework. It is important to realize that not all samples are random.

Definition 4.4. Non-random sample. A sample is *non-random* if the realizations that compose it are not drawn from i.i.d. random variables, or vectors, or matrices. Instead, these may be:

- **independent and not identically distributed** (**i.n.i.d.**);
- **not independent and identically distributed** (**n.i.i.d.**);
- **not independent and not identically distributed** (**n.i.n.i.d.**).

Intuitively, the more a sample departs from the i.i.d. benchmark the more statistical inference is complicated. Yet in social sciences non-random samples are common. For example, the data may be composed by observations obtained from recognizably different distributions (i.n.i.d.), or whose characteristics are characterized by statistical dependence due to some underlying socio-economic phenomenon, like group behavior or the response to economic events (n.i.i.d.). While statistical inference is still possible on non-random samples, the asymptotic framework is better suited to these settings. The present lecture only focuses on random samples. Yet it must

be mentioned that the modern theory of econometrics extends to the general n.i.n.i.d. case because of the necessity to deal with real data that hardly fit the i.n.i.d. case, and almost never (except special cases) the i.i.d. one.

The rest of this section heretofore focuses on the random, i.i.d. case. A first observation about random samples is that since their realizations are independent, it is easy to express their associated joint probability mass or density function, as the product of every unit of observation's joint density:

$$f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}\left(\mathbf{x}_1,\ldots,\mathbf{x}_N;\boldsymbol{\theta}\right)=f_{\boldsymbol{x}}\left(\mathbf{x}_1;\boldsymbol{\theta}\right)\times\cdots\times f_{\boldsymbol{x}}\left(\mathbf{x}_N;\boldsymbol{\theta}\right)=\prod_{i=1}^{N}f_{\boldsymbol{x}}\left(\mathbf{x}_i;\boldsymbol{\theta}\right)$$

where $\boldsymbol{\theta}$ is the collection of parameters that are associated with the probability distribution of $\boldsymbol{x}$. This is an extremely useful fact aiding the analysis of selected **statistics** of a random sample.

Definition 4.5. Statistic. A function of the N random variables, vectors or matrices that are specific to each i-th unit of observation and that generate a sample is called a *statistic*. Any statistic is itself a random variable, vector or matrix.

Definition 4.6. Sampling distribution. The probability distribution of a statistic is called its *sampling* distribution.

The two most common and better known statistics are the following.

Definition 4.7. Sample mean. In samples derived from random vectors, the *sample mean* is a vector-valued statistic which is usually denoted as $\bar{\boldsymbol{x}}$ and defined as follows.

$$\bar{\boldsymbol{x}}=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i$$

This definition can be reduced to samples that are drawn from univariate random variables, in which case the usual notation is $\bar{X}$:

$$\bar{X}=\frac{1}{N}\sum_{i=1}^{N}X_i$$

or extended to samples drawn from random matrices, where one can write $\bar{\boldsymbol{X}}$ and the definition is again analogous.

Definition 4.8. Sample variance-covariance. In samples collected from random vectors, the *sample variance-covariance* is a matrix-valued statistic which is usually denoted by $\boldsymbol{S}$ and defined as follows.

$$\boldsymbol{S}=\frac{1}{N-1}\sum_{i=1}^{N}\left(\boldsymbol{x}_i-\bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_i-\bar{\boldsymbol{x}}\right)^{\mathrm{T}}$$

In samples from univariate random variables, this statistic is simply called *sample variance*, its associated notation is S^2, and it is a scalar.

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2$$

In this case, the square root of the sample variance is written $S = \sqrt{S^2}$ and called the *standard deviation*. In order to extend this definition to sampling from random matrices it is necessary to develop three-dimensional arrays.

These statistics have some important properties, which are proved here in the vector-valued case.

Theorem 4.1. Properties of simple sample statistics (1). *Consider a sample* $\{\boldsymbol{x}_i\}_{i=1}^N$*, its sample mean* $\bar{\boldsymbol{x}}$*, and its sample variance-covariance* $\boldsymbol{S}$*. The following two properties are true:*

a. $\bar{\boldsymbol{x}} = \arg\min_{\mathbf{a}\in\mathbb{R}^K} \sum_{i=1}^N (\boldsymbol{x}_i - \mathbf{a})(\boldsymbol{x}_i - \mathbf{a})^{\mathrm{T}}$*;*

b. $(N-1)\,\boldsymbol{S} = \sum_{i=1}^N \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} - N\cdot\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}$.

Proof. To show point **a.** note that:

$$\begin{aligned}
\sum_{i=1}^{N}(\boldsymbol{x}_i-\mathbf{a})(\boldsymbol{x}_i-\mathbf{a})^{\mathrm{T}} &= \sum_{i=1}^{N}(\boldsymbol{x}_i-\bar{\boldsymbol{x}}+\bar{\boldsymbol{x}}-\mathbf{a})(\boldsymbol{x}_i-\bar{\boldsymbol{x}}+\bar{\boldsymbol{x}}-\mathbf{a})^{\mathrm{T}} \\
&= \sum_{i=1}^{N}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\mathrm{T}} + \sum_{i=1}^{N}(\bar{\boldsymbol{x}}-\mathbf{a})(\bar{\boldsymbol{x}}-\mathbf{a})^{\mathrm{T}} \\
&\quad + \sum_{i=1}^{N}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}-\mathbf{a})^{\mathrm{T}} + \sum_{i=1}^{N}(\bar{\boldsymbol{x}}-\mathbf{a})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\mathrm{T}} \\
&= \sum_{i=1}^{N}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\mathrm{T}} + \sum_{i=1}^{N}(\bar{\boldsymbol{x}}-\mathbf{a})(\bar{\boldsymbol{x}}-\mathbf{a})^{\mathrm{T}}
\end{aligned}$$

where two terms in the second line are both equal to zero by definition of sample mean; in the last line, the first term does not depend on $\mathbf{a}$ while the second is minimized at $\mathbf{a} = \bar{\boldsymbol{x}}$. To show **b.** simply note that:

$$\begin{aligned}
\sum_{i=1}^{N}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\mathrm{T}} &= \sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} - \sum_{i=1}^{N}\boldsymbol{x}_i\bar{\boldsymbol{x}}^{\mathrm{T}} - \sum_{i=1}^{N}\bar{\boldsymbol{x}}\boldsymbol{x}_i^{\mathrm{T}} + N\cdot\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}} \\
&= \sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} - N\cdot\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}
\end{aligned}$$

and the result again follows from the definition of a sample mean. □

Theorem 4.2. Properties of simple sample statistics (2). *Consider a* random *sample* $\{\boldsymbol{x}_i\}_{i=1}^N$ *drawn from a random vector* $\boldsymbol{x}$*, a transformation of this vector* $\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x})$*, and suppose that all the moments expressed in the mean vector* $\mathbb{E}[\boldsymbol{y}]$ *and in the variance-covariance matrix* $\mathbb{V}\text{ar}[\boldsymbol{y}]$ *are defined. The following two properties are true:*

a. $\mathbb{E}\left[\sum_{i=1}^N \boldsymbol{y}_i\right] = N \cdot \mathbb{E}[\boldsymbol{y}_i]$*;*

b. $\mathbb{V}\text{ar}\left[\sum_{i=1}^N \boldsymbol{y}_i\right] = N \cdot \mathbb{V}\text{ar}[\boldsymbol{y}_i]$.

Proof. To show **a.** simply observe that:

$$\mathbb{E}\left[\sum_{i=1}^N \boldsymbol{y}_i\right] = \sum_{i=1}^N \mathbb{E}[\boldsymbol{y}_i] = N \cdot \mathbb{E}[\boldsymbol{y}_i]$$

where the first equality follows from the linear properties of the expectation operator and the second equality follows from the fact that the distributions of $\boldsymbol{y}_i$ for $i = 1, \ldots, N$ are identical (this particular result does not require independence and is also valid for n.i.i.d. samples). Regarding **b.** it is:

$$\begin{aligned}
\mathbb{V}\text{ar}\left[\sum_{i=1}^N \boldsymbol{y}_i\right] &= \mathbb{E}\left[\left(\sum_{i=1}^N \boldsymbol{y}_i - \mathbb{E}\left[\sum_{i=1}^N \boldsymbol{y}_i\right]\right)\left(\sum_{i=1}^N \boldsymbol{y}_i - \mathbb{E}\left[\sum_{i=1}^N \boldsymbol{y}_i\right]\right)^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^N (\boldsymbol{y}_i - \mathbb{E}[\boldsymbol{y}_i])\right)\left(\sum_{i=1}^N (\boldsymbol{y}_i - \mathbb{E}[\boldsymbol{y}_i])\right)^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^N (\boldsymbol{y}_i - \mathbb{E}[\boldsymbol{y}_i])(\boldsymbol{y}_i - \mathbb{E}[\boldsymbol{y}_i])^{\mathrm{T}}\right] \\
&= \sum_{i=1}^N \mathbb{E}\left[(\boldsymbol{y}_i - \mathbb{E}[\boldsymbol{y}_i])(\boldsymbol{y}_i - \mathbb{E}[\boldsymbol{y}_i])^{\mathrm{T}}\right] \\
&= N \cdot \mathbb{V}\text{ar}[\boldsymbol{y}_i]
\end{aligned}$$

where the first line is just the definition of variance for $\sum_{i=1}^N \boldsymbol{y}_i$, the second line applies the linear properties of expectations while also rearranging terms, the third line rearranges terms again after observing that, for $i \neq j$:

$$\mathbb{E}\left[(\boldsymbol{y}_i - \mathbb{E}[\boldsymbol{y}_i])(\boldsymbol{y}_j - \mathbb{E}[\boldsymbol{y}_j])^{\mathrm{T}}\right] = \mathbf{0}$$

which follows from the independence of the realizations in the random sample, the fourth line is another application of the linear properties of expectations, while the fifth line again exploits the fact that all the realizations follow from identically distributed random variables. □

Theorem 4.3. Properties of simple sample statistics (3). *Consider a* random *sample* $\{\boldsymbol{x}_i\}_{i=1}^N$ *drawn from a random vector* $\boldsymbol{x}$ *whose mean vector is* $\mathbb{E}[\boldsymbol{x}]$ *and whose variance-covariance matrix is* $\mathbb{V}\text{ar}[\boldsymbol{x}] < \infty$ *and is finite. The following three properties are true:*

a. $\mathbb{E}[\bar{\boldsymbol{x}}] = \mathbb{E}[\boldsymbol{x}]$*;*

b. $\mathbb{V}\text{ar}[\bar{\boldsymbol{x}}] = \mathbb{V}\text{ar}[\boldsymbol{x}]/N$*;*

c. $\mathbb{E}[\boldsymbol{S}] = \mathbb{V}\text{ar}[\boldsymbol{x}]$.

Proof. To show **a.** it is sufficient to apply Theorem 4.2, point **a.** for $\boldsymbol{y} = \boldsymbol{x}$:

$$\mathbb{E}[\bar{\boldsymbol{x}}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\boldsymbol{x}_i] = \frac{1}{N}\cdot N\cdot\mathbb{E}[\boldsymbol{x}_i] = \mathbb{E}[\boldsymbol{x}]$$

and point **b.** proceeds similarly.

$$\mathbb{V}\text{ar}[\bar{\boldsymbol{x}}] = \mathbb{V}\text{ar}\left[\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i\right] = \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{V}\text{ar}[\boldsymbol{x}_i] = \frac{1}{N^2}\cdot N\cdot\mathbb{V}\text{ar}[\boldsymbol{x}_i] = \frac{\mathbb{V}\text{ar}[\boldsymbol{x}]}{N}$$

The proof of point **c.** is as follows:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{S}] &= \mathbb{E}\left[\frac{1}{N-1}\left(\sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} - N\cdot\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}\right)\right] \\
&= \frac{1}{N-1}\left(\sum_{i=1}^{N}\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right] - N\cdot\mathbb{E}\left[\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}\right]\right) \\
&= \frac{1}{N-1}\left(N\cdot\mathbb{V}\text{ar}[\boldsymbol{x}_i] - N\cdot\mathbb{V}\text{ar}[\bar{\boldsymbol{x}}]\right) \\
&= \frac{N}{N-1}\left(1-\frac{1}{N}\right)\mathbb{V}\text{ar}[\boldsymbol{x}] \\
&= \mathbb{V}\text{ar}[\boldsymbol{x}]
\end{aligned}$$

where the third line follows after adding and subtracting $N\cdot\mathbb{E}[\boldsymbol{x}]\,\mathbb{E}[\boldsymbol{x}]^{\mathrm{T}}$. □

The theorem examined last is the culmination of the results analyzed previously, and specifies how to obtain quantities that can be used to evaluate – more precisely, *estimate* – the moments of the underlying distribution. By selecting the sample mean and the sample variance-covariance for the sake of estimating the corresponding moments, one can rely on the property that the expectation of those quantities is indeed the moment sought after, in both cases. Later, this property is defined *unbiasedness*.

4.2 Normal Sampling

Once the expectations of both the sample mean and the sample variance-covariance are known, the next step is to identify the sampling distribution of the sampling mean. This is necessary for the sake of characterizing the probability about the occurrence of different samples. In general, there is no unique answer to this problem, as it ultimately depends on the underlying distribution of the random vector that generates a random sample. In some specific cases, however, it is possible to leverage upon Theorem 3.7 in order to derive the distribution of univariate sample means obtained from selected distributions. The most important of these cases is the one about **sampling from the normal distribution**. Specifically, if a random sample $\{x_i\}_{i=1}^{N}$ is drawn from a random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, it is easy to see that:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right) \tag{4.1}$$

which is an extension of Observation 3.11 from Lecture 3. The above result can be alternatively expressed in a *standardized* fashion, which can be more convenient for calculating probabilities about the sample mean.

$$\sqrt{N}\,\frac{\bar{X} - \mu}{\sigma} \sim \mathcal{N}(0, 1) \tag{4.2}$$

In practical settings, those analysts who are interested about evaluating – better, *estimating* – a specific value for the mean of a normal distribution find results (4.1) and (4.2) of limited use. The reason is that, to manipulate the density function of the normal distribution in order to make statements about the probability that the parameter μ falls within specified ranges – that is to perform *hypothesis tests*, see Lecture 5 – it is necessary to know the parameter σ beforehand. In actual practice this information is generally inaccessible to researchers. An intuitive workaround is to substitute σ with its associated sample statistic, that is the standard deviation S. Doing so is tantamount to working with a well-known statistic.

Definition 4.9. The *t*-statistic. Given some univariate sample $\{x_i\}_{i=1}^{N}$ of size N drawn from a sequence of random variables $X_1, \ldots, X_N$, a *t-statistic* is defined as the following quantity:

$$\mathrm{t} = \sqrt{N}\,\frac{\bar{X} - \mu}{S}$$

where $\bar{X}$ is the sample mean whose expectation is $\mu = \mathbb{E}\left[\bar{X}\right]$, and S is the sample standard deviation.

The next result is central in statistical inference and allows to derive the sampling distribution of the t-statistic when the sample is drawn from the normal distribution.

Theorem 4.4. Sampling from the Normal Distribution. *Consider a random sample* $\{x_i\}_{i=1}^N$ *which is drawn from a random variable following the normal distribution* $X \sim \mathcal{N}(\mu, \sigma^2)$*, and the random variables corresponding to the two sample statistics* $\bar{X} = \frac{1}{N}\sum_{i=1}^N X_i$ *and* $S^2 = \frac{1}{N-1}\sum_{i=1}^N \left(X_i - \bar{X}\right)^2$*. The following three properties are true:*

a. $\bar{X}$ *and* S^2 *are independent;*

b. $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/N)$*;*

c. $(N-1)\, S^2/\sigma^2 \sim \chi^2_{N-1}$*.*

Proof. Point **a.** is the most crucial to show. To this end, it is useful to start from the observation that the sample variance can be expressed in terms of only $N-1$ of the original random variables, say $X_2, \ldots, X_N$:

$$\begin{aligned} S^2 &= \frac{1}{N-1}\sum_{i=1}^N \left(X_i - \bar{X}\right)^2 \\ &= \frac{1}{N-1}\left[\left(X_1 - \bar{X}\right)^2 + \sum_{i=2}^N \left(X_i - \bar{X}\right)^2\right] \\ &= \frac{1}{N-1}\left[\left(\sum_{i=2}^N \left(X_i - \bar{X}\right)\right)^2 + \sum_{i=2}^N \left(X_i - \bar{X}\right)^2\right] \end{aligned}$$

where the last line follows from $\sum_{i=1}^N \left(X_i - \bar{X}\right) = 0$. Consequently, proving that the sample mean is independent of the sample variance requires to show that $\bar{X}$ is independent of $N-1$ out of the N *demeaned* normally distributed random variables, say $X_2 - \bar{X}, \ldots, X_N - \bar{X}$. To do so, a convenient approach is to define the following random vector $\widetilde{\boldsymbol{z}}$ of length N, which is a function of the *standardized* random variables $Z_i = (X_i - \mu)/\sigma$ for $i = 1, \ldots, N$.

$$\widetilde{\boldsymbol{z}} = \begin{pmatrix} \bar{Z} \\ \widetilde{Z}_2 \\ \vdots \\ \widetilde{Z}_N \end{pmatrix} = \begin{pmatrix} \bar{Z} \\ Z_2 - \bar{Z} \\ \vdots \\ Z_N - \bar{Z} \end{pmatrix} = \begin{bmatrix} N^{-1} & N^{-1} & \ldots & N^{-1} \\ -N^{-1} & 1 - N^{-1} & \ldots & -N^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ -N^{-1} & -N^{-1} & \ldots & 1 - N^{-1} \end{bmatrix} \boldsymbol{z}$$

Here, $\boldsymbol{z} = (Z_i, \ldots, Z_N)$. One can prove that this linear transformation has Jacobian determinant equaling $1/N$, and therefore it is invertible; it follows

that the Jacobian determinant *of the inverse transformation* is equal to N. Thus, the joint distribution of $\widetilde{\boldsymbol{z}}$ can be obtained from that of $\boldsymbol{z}$ by applying Theorem 3.1 quite straightforwardly:

$$\begin{aligned} f_{\widetilde{\boldsymbol{z}}}\left(\bar{z}, \widetilde{z}_2, \ldots, \widetilde{z}_N\right) &= \frac{N}{\sqrt{(2\pi)^N}} \exp\left(-\frac{1}{2}\left(\bar{z} - \sum_{i=2}^{N} \widetilde{z}_i\right)^2 - \frac{1}{2}\sum_{i=2}^{N}\left(\bar{z} + \widetilde{z}_i\right)^2\right) \\ &= \sqrt{\frac{N}{2\pi}} \exp\left(-\frac{N\bar{z}^2}{2}\right) \times \\ &\quad \times \sqrt{\frac{N}{(2\pi)^{N-1}}} \exp\left(-\frac{1}{2}\left(\sum_{i=2}^{N} \widetilde{z}_i\right)^2 - \frac{1}{2}\sum_{i=2}^{N} \widetilde{z}_i^2\right) \\ &= f_{\bar{Z}}\left(\bar{z}\right) \cdot f_{\widetilde{\boldsymbol{z}}_{-1}}\left(\widetilde{z}_2, \ldots, \widetilde{z}_N\right) \end{aligned}$$

and it can be clearly decomposed into the product of two components: the density function of $\bar{Z}$ and that of all the other elements of $\widetilde{\boldsymbol{z}}$, implying that $\bar{X}$ is independent of $X_2 - \bar{X}, \ldots, X_N - \bar{X}$, and consequently of S^2.

To continue the proof about the other points in the statement, note that point **b.** is, as said, a consequence of Theorem 3.7. In order to demonstrate point **c.** instead, it is easiest to proceed as follows:

$$\begin{aligned} \left(N-1\right)\frac{S^2}{\sigma^2} &= \sum_{i=1}^{N} \frac{\left(X_i - \bar{X}\right)^2}{\sigma^2} \\ &= \sum_{i=1}^{N} \frac{\left(X_i - \mu + \mu - \bar{X}\right)^2}{\sigma^2} \\ &= \sum_{i=1}^{N} \frac{\left(X_i - \mu\right)^2}{\sigma^2} - \frac{N\left(\bar{X} - \mu\right)^2}{\sigma^2} - 2\left(\bar{X} - \mu\right)\sum_{i=1}^{N}\frac{X_i - \mu}{\sigma^2} \\ &= \sum_{i=1}^{N}\left(\frac{X_i - \mu}{\sigma}\right)^2 - \left(\sqrt{N}\frac{\bar{X} - \mu}{\sigma}\right)^2 \end{aligned}$$

that is, the statistic $\left(N-1\right)S^2/\sigma^2$ is shown to be the sum of the *squares* of N independent random variables all of which follow the standard normal distribution (the standardized versions of $X_i, \ldots, X_N$) *minus* the square of another random variable that follows the standard normal distribution (the standardized version of the sample mean $\bar{X}$). By the demonstration of point **a.** the latter is independent of the former. Consequently, the distribution of

the statistic of interest can be obtained by a variation of Observation 3.9:

$$M_{\bar{\bar{Z}}^2}(t)\, M_{(N-1)\frac{S^2}{\sigma^2}}(t) = \prod_{i=1}^{N} M_{Z_i^2}(t)$$

where $\bar{\bar{Z}} \equiv \sqrt{N}\left(\bar{X} - \mu\right)/\sigma$ and $Z_i \equiv (X_i - \mu)/\sigma$, or equivalently:

$$\begin{aligned} M_{(N-1)\frac{S^2}{\sigma^2}}(t) &= \frac{1}{M_{\bar{\bar{Z}}^2}(t)} \prod_{i=1}^{N} M_{Z_i^2}(t) \\ &= (1-2t)^{-\frac{1}{2}(N-1)} \end{aligned}$$

which follows since all the $N+1$ moment generating functions involved on the right-hand side are those of a chi-squared distribution with one degree of freedom; consequently the end result is the moment-generating function of a chi-squared distribution with $N-1$ degrees of freedom. □

This result took some effort to show, but it allows to derive the sampling distribution of the t-statistic. In fact, the ratio:

$$\mathrm{t} = \sqrt{N}\,\frac{\bar{X} - \mu}{S} = \frac{\sqrt{N}\,\dfrac{\bar{X} - \mu}{\sigma}}{\sqrt{(N-1)\,\dfrac{S^2}{\sigma^2}\,\dfrac{1}{\sqrt{N-1}}}} \sim \mathcal{T}_{N-1}$$

is easily seen as the ratio between two independent random variables; the one in the numerator follows the standard normal distribution, while the one in the denominator equals the square root of a random variable that follows the chi-squared distribution with $N-1$ degrees of freedom, *divided* by the square root of $N-1$. Hence, by Observation 3.2, a t-statistic follows the Student's t-distribution with $N-1$ degrees of freedom. Theorem 4.4 is also useful to obtain the sampling distribution of another important statistic.

Definition 4.10. Normal variance ratio. Consider two univariate random samples $\{x_i\}_{i=1}^{N_X}$ and $\{y_i\}_{i=1}^{N_Y}$ of sizes N_X and N_Y respectively, each drawn from two *independent* sequences of random variables $(X_1, \ldots, X_{N_X})$ and $(Y_1, \ldots, Y_{N_Y})$ whose distributions are as follows.

$$\begin{aligned} X_i &\sim \mathcal{N}\left(\mu_X, \sigma_X^2\right) \text{ for } i = 1, \ldots, N_X \\ Y_j &\sim \mathcal{N}\left(\mu_Y, \sigma_Y^2\right) \text{ for } j = 1, \ldots, N_Y \end{aligned}$$

The normal variance ratio is defined as the following F-statistic:

$$\mathrm{F} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$$

where S_X and S_Y are the sample variances of the two random samples.

The variance ratio is the statistic used by analysts to evaluate whether two population of interest, call them X and Y by the denomination of the respective random variables, have the same variance – or more generally, whether their variance is equal to some given quantity σ_X^2/σ_Y^2. Once again, these evaluations – these exercises in statistical inference – require knowledge about the distribution of the statistic in question. Theorem 4.4 comes again to the rescue, ensuring that both the numerator and the denominator of F, if multiplied by $N_X - 1$ and $N_Y - 1$ respectively, follow a chi-squared distribution with those numbers as degrees of freedom; so, by Observation 3.3:

$$\mathrm{F} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim \mathcal{F}_{N_X-1,N_Y-1}$$

that is, the variance ratio F follows the Snedecor F-distribution with paired degrees of freedom given by $N_X - 1$ and $N_Y - 1$. This fact is exploited in statistical tests about the variances of different populations drawn from the uniform distribution, as discussed in Lecture 5.

Most of these ideas can be extended to **multivariate** normal sampling, where analysts have access to a vector-valued random sample $\{\mathbf{x}_i\}_{i=1}^N$ drawn from some multivariate normal distribution $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Even in this environment, statistical inference requires the development of sample statistics with a clearly recognizable distribution. Clearly, the standardized multivariate sample mean in this case also follows a multivariate normal distribution, thanks to the linear properties of the latter.

$$\bar{\boldsymbol{x}} \sim \mathcal{N}\left(\boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{N}\right) \tag{4.3}$$

However, the above statistic is a random vector, which makes it unsuitable in some specific stages of statistical inference, like tests of hypotheses. This observation led to the development of the following statistic.

Definition 4.11. *u*-statistic. Given some multivariate sample $\{\mathbf{x}_i\}_{i=1}^N$ of size N drawn from a sequence of random vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, a *u-statistic*[1] is defined as the following quantity:

$$\begin{aligned} \mathrm{u} &= N\left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}\right) \\ &= N \sum_{k=1}^{K} \sum_{\ell=1}^{K} \sigma_{k\ell}^{*-1} \left(\bar{X}_k - \mu_k\right) \left(\bar{X}_\ell - \mu_\ell\right) \end{aligned}$$

where $\bar{\boldsymbol{x}}$ is the sample mean whose expectation and variance are $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{x}_i]$, and $\boldsymbol{\Sigma}/N = \mathbb{V}\mathrm{ar}[\boldsymbol{x}_i]$ respectively, and where $\sigma_{k\ell}^{*-1}$ is $k\ell$-th element of $\boldsymbol{\Sigma}^{-1}$.

[1]Note that this nomenclature is not standard, but is applied throughout the lectures.

To better interpret the u-statistic, it is useful to analyze its development as a quadratic form in the second line of the above definition: the statistic is a second degree polynomial of the K deviations of all univariate sample means from their respective mean parameters, *normalized* by the population variance-covariance. In this respect, the u-statistic intuitively appears to be a multivariate generalization of the *squared* standardized univariate sample mean. The following result should also be intuitive enough upon recalling the relationship between the normal and chi-squared distributions.

Theorem 4.5. Sampling from the Multivariate Normal Distribution. *Consider a random sample* $\{\mathbf{x}_i\}_{i=1}^{N}$ *drawn from a* K*-dimensional random vector following the multivariate normal distribution,* $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*. In this environment, the* u*-statistic follows the chi-squared distribution with* K *degrees of freedom.*

$$\mathrm{u} = N(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \sim \chi^2_K$$

Proof. Derive the moment-generating function of the u-statistic, exploiting $\bar{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}/N)$. With the aid of some linear algebra:

$$\begin{aligned}
M_{\mathrm{u}}(t) &= \int_{\mathbb{R}^K} \exp\left(N(\bar{\mathbf{x}} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu})\, t\right) f_{\bar{\boldsymbol{x}}}\left(\bar{\mathbf{x}}; \boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{N}\right) \mathrm{d}\bar{\mathbf{x}} \\
&= \int_{\mathbb{R}^K} \sqrt{\frac{1}{(2\pi)^K} \frac{N^K}{|\boldsymbol{\Sigma}|}} \exp\left(-\frac{N}{2} (\bar{\mathbf{x}} - \boldsymbol{\mu})^{\mathrm{T}} \left[(1-2t)\boldsymbol{\Sigma}^{-1}\right] (\bar{\mathbf{x}} - \boldsymbol{\mu})\, t\right) \mathrm{d}\bar{\mathbf{x}} \\
&= \sqrt{\frac{1}{(1-2t)^K}} \times \int_{\mathbb{R}^K} \sqrt{\frac{1}{(2\pi)^K} \frac{[(1-2t)N]^K}{|\boldsymbol{\Sigma}|}} \times \\
&\qquad \times \exp\left(-\frac{N}{2} (\bar{\mathbf{x}} - \boldsymbol{\mu})^{\mathrm{T}} \left[(1-2t)\boldsymbol{\Sigma}^{-1}\right] (\bar{\mathbf{x}} - \boldsymbol{\mu})\, t\right) \mathrm{d}\bar{\mathbf{x}} \\
&= (1-2t)^{-\frac{K}{2}}
\end{aligned}$$

where the integral in the third line disappears because it is the expression of the probability density function of a multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance $\boldsymbol{\Sigma}/(1-2t)N$. The result is the moment-generating function of a chi-squared distribution with parameter K. □

While this is a nice theoretical result, its direct applications are rather limited, since in applied statistical analysis the variance-covariance matrix $\boldsymbol{\Sigma}$ is seldom known *a-priori*. Just like in the univariate case, the immediate temptation is to substitute $\boldsymbol{\Sigma}$ with its corresponding sample statistic, the sample variance-covariance $\boldsymbol{S}$. This gives rise to yet another statistic, which intuitively is the multivariate generalization of the t-statistic.

Definition 4.12. Hotelling's "t-squared" statistic. Given some multivariate sample $\{\mathbf{x}_i\}_{i=1}^N$ of size N drawn from a sequence of random vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, Hotelling's *t-squared statistic* is defined as the random variable:

$$\begin{aligned} \mathrm{t}^2 &= N\left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{S}^{-1}\left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}\right) \\ &= N \sum_{k=1}^{K} \sum_{\ell=1}^{K} S_{k\ell}^{*-1}\left(\bar{X}_k - \mu_k\right)\left(\bar{X}_\ell - \mu_\ell\right) \end{aligned}$$

where $\bar{\boldsymbol{x}}$ is the sample mean whose expectation is $\boldsymbol{\mu} = \mathbb{E}\left[\boldsymbol{x}_i\right]$, $\boldsymbol{S}$ is the sample variance-covariance, and $S_{k\ell}^{*-1}$ is $k\ell$-th element of $\boldsymbol{S}^{-1}$.

Another result, which is not proved here, states that the following *rescaled* version of Hotelling's t-squared statistic:

$$\frac{N-K}{K\left(N-1\right)}\,\mathrm{t}^2 = \frac{N\left(N-K\right)}{K\left(N-1\right)}\left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{S}^{-1}\left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}\right) \sim \mathcal{F}_{K,N-K}$$

follows the F-distribution with paired degrees of freedom K and $N-K$.[2] This finding allows to conduct statistical inference on multivariate normal samples through a well-known univariate distribution (see Lecture 5).

4.3 Order Statistics

In statistical analysis it is often useful to study the values that the realization of a random variable take at a given position of the *order* of realizations (e.g. the smallest value, the largest value, *et cetera*). These are themselves realizations of random variables: more precisely, of the following statistics.

Definition 4.13. Order statistics. Consider some sample $\{x_i\}_{i=1}^N$ of realizations obtained from univariate random variables, $\{X_i\}_{i=1}^N$. Suppose that these values are placed in *ascending order*, where subscripts surrounded by parentheses denote one observation's position in the order:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(N)}$$

thus $x_{(1)} = \min\{x_i\}_{i=1}^N$ and $x_{(N)} = \max\{x_i\}_{i=1}^N$. The j-th *order statistic* is the random variable, denoted as $X_{(j)}$, that generates the j-th realization in the above sequence, that is $x_{(j)}$. Any univariate sample has N associated order statistics that must satisfy the following property.

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(N)}$$

[2] To prove this result one must develop the distribution of the *random matrix* $\boldsymbol{S}$: the so-called *Wishart* distribution, which is outside the scope of this analysis.

Definition 4.14. Sample Minimum. The sample *minimum* is the first order statistic, $X_{(1)}$.

Definition 4.15. Sample Maximum. The sample *maximum* is the N-th order statistic, $X_{(N)}$.

Definition 4.16. Sample Range. The sample *range* R is the difference between the sample maximum and the sample minimum: $R = X_{(N)} - X_{(1)}$.

Definition 4.17. Sample Median. The sample *median* M is a function of a sample's most central order statistics.

$$M = \begin{cases} X_{\left(\frac{N+1}{2}\right)} & \text{if } N \text{ is odd} \\ \frac{1}{2}\left(X_{\left(\frac{N}{2}\right)} + X_{\left(\frac{N}{2}+1\right)}\right) & \text{if } N \text{ is even} \end{cases}$$

It is occasionally useful to analyze the sampling distribution of selected order statistics. Fortunately, this task is simplified greatly if the sample is random. In fact, the cumulative distribution of the j-th order statistic can be expressed in terms of the following joint probability:

$$F_{X_{(j)}}(x) = \mathbb{P}\left(X_{(1)} \leq x \cap \cdots \cap X_{(j)} \leq x\right)$$

that is, for the j-th order statistic to be less or equal than some x, all the inferior order statistic must also be less or equal than x (while the superior ones can be larger, equal or lower than x). In a random sample, where all the realizations obtain from independent and identically distributed random variables, the above expression is considerably easier to evaluate.

Theorem 4.6. Sampling distribution of order statistics in a random sample. *In a univariate random sample, the cumulative distribution of the j-th order statistic is based on the binomial distribution:*

$$F_{X_{(j)}}(x) = \sum_{k=j}^{N} \binom{N}{k} \left[F_X(x)\right]^k \left[1 - F_X(x)\right]^{N-k}$$

where $F_X(x)$ is the cumulative distribution of the random variable X that generates the sample. As two particular cases, the cumulative distributions of the minimum and the maximum are as follows.

$$F_{X_{(1)}}(x) = 1 - \left[1 - F_X(x)\right]^N$$
$$F_{X_{(N)}}(x) = \left[F_X(x)\right]^N$$

Proof. For *at least* j realizations to be less or equal than x, the event defined as $X_i \leq x$ must occur an integer number of $j \leq k \leq N$ times, whereas the complementary event $X_i > x$ must instead occur $N-k$ times. If the sample is random (i.i.d.), these two fundamental events occur with probabilities that are constant across all realizations:

$$\begin{aligned}\mathbb{P}\left(X_i \leq x\right) &= F_X\left(x\right)\\ \mathbb{P}\left(X_i > x\right) &= 1 - F_X\left(x\right)\end{aligned}$$

and since realizations are independent, any joint combination of said events can be expressed as the appropriate product of those probabilities. Clearly, for a given k any joint events can be expressed through a binomial distribution, with the binomial coefficient counting all potential combinations with k "successes" $(X_i \leq x)$ and $N-k$ "failures" $(X_i > x)$. Summing over the eligible values of k delivers the result sought after, of which the distributions for the minimum and the maximum are special cases. □

Corollary. *If X is a continuous distribution with density function $f_X\left(x\right)$, the density function of the j-th order statistic is the following.*

$$f_{X_{(j)}}\left(x\right) = \frac{N!}{(j-1)!\,(N-j)!}\, f_X\left(x\right)\left[F_X\left(x\right)\right]^{j-1}\left[1 - F_X\left(x\right)\right]^{N-j}$$

Proof. This follows from manipulating the derivative of the cumulative distribution $F_{X_{(j)}}\left(x\right)$. By the chain rule one obtains the density function:

$$\begin{aligned}f_{X_{(j)}}\left(x\right) &= \frac{dF_{X_{(j)}}\left(x\right)}{dx}\\ &= \sum_{k=j}^{N}\binom{N}{k}\Big(k\left[F_X\left(x\right)\right]^{k-1}\left[1 - F_X\left(x\right)\right]^{N-k} f_X\left(x\right) -\\ &\qquad - \left(N-k\right)\left[F_X\left(x\right)\right]^{k}\left[1 - F_X\left(x\right)\right]^{N-k-1} f_X\left(x\right)\Big)\\ &= \frac{N!}{(j-1)!\,(N-j)!}\, f_X\left(x\right)\left[F_X\left(x\right)\right]^{j-1}\left[1 - F_X\left(x\right)\right]^{N-j} +\\ &\qquad + \sum_{k=j+1}^{N}\binom{N}{k} k\left[F_X\left(x\right)\right]^{k-1}\left[1 - F_X\left(x\right)\right]^{N-k} f_X\left(x\right) -\\ &\qquad - \sum_{k=j}^{N}\binom{N}{k}\left(N-k\right)\left[F_X\left(x\right)\right]^{k}\left[1 - F_X\left(x\right)\right]^{N-k-1} f_X\left(x\right)\end{aligned}$$

where the third line obtains by isolating the term for $k = j$ in the summation that results from taking the derivative. All that is left to do is to show that

the two "residual" summations in the third line cancel out. To this end, some additional manipulation is necessary. In particular, a re-indexing of the *first* of the two concerned residual summations, as well as the observation that the term for $k = N$ in the *second* residual summation equals zero, gives:

$$\begin{aligned} f_{X_{(j)}}(x) = \frac{N!}{(j-1)!\,(N-j)!} f_X(x) \left[F_X(x)\right]^{j-1} \left[1 - F_X(x)\right]^{N-j} + \\ + \sum_{k=j}^{N-1} \binom{N}{k+1} (k+1) \left[F_X(x)\right]^{k} \left[1 - F_X(x)\right]^{N-k-1} f_X(x) - \\ - \sum_{k=j}^{N-1} \binom{N}{k} (N-k) \left[F_X(x)\right]^{k} \left[1 - F_X(x)\right]^{N-k-1} f_X(x) \end{aligned}$$

and since, by simple manipulation of factorials, it is:

$$\binom{N}{k+1}(k+1) = \frac{N!}{k!\,(N-k-1)!} = \binom{N}{k}(N-k)$$

it clearly follows that the two terms do indeed cancel out. □

This theorem is important, but its practical use is circumscribed as the application of the above formulae seldom returns expressions that relate to known distributions. One particular result, however, stands out.

Observation 4.1. Consider a random sample drawn from the standard continuous uniform distribution, $X \sim \mathcal{U}(0,1)$. The j-th order statistic is such that $X_{(j)} \sim \text{Beta}(j, N-j+1)$.

Proof. Since $F_X(x) = x$ and $f_X(x) = 1$ for $x \in (0,1)$, while $F_X(x) = x$ and $f_X(x) = 0$ otherwise, the density function of $X_{(j)}$ is, for $x \in (0,1)$:

$$\begin{aligned} f_{X_{(j)}}(x) &= \frac{N!}{(j-1)!\,(N-j)!} x^{j-1} (1-x)^{N-j} \\ &= \frac{\Gamma(N+1)}{\Gamma(j)\,\Gamma(N-j+1)} x^{j-1} (1-x)^{(N-j+1)-1} \\ &= \text{B}(j, N-j+1) \cdot x^{j-1} (1-x)^{(N-j+1)-1} \end{aligned}$$

where the second line follows from the properties of the Gamma function; the result is the density function of the postulated Beta distribution. □

This result, in conjunction with Theorem 1.13, allows to derive the sampling distribution of the order statistic of a sample of *percentiles* p drawn from some known distribution, as $p \sim \mathcal{U}(0,1)$.

Other relevant results are, instead, restricted to the two extreme order statistics: the minimum and the maximum. In particular, certain distribution have the useful feature to return – in random samples – minima and maxima that follows a distribution in the same sub-family.

Definition 4.18. Extreme order statistics (min-max) stability. Consider a random sample drawn from some known distribution. If the sample minimum (maximum) follows another distribution of the same family, that distribution is said to be *min-stable* (*max-stable*).

For example, the exponential distribution is notoriously min-stable.

Observation 4.2. Consider a random sample drawn from the exponential distribution with parameter λ, $X \sim \mathrm{Exp}(\lambda)$. The first order statistic – the minimum – is such that $X_{(1)} \sim \mathrm{Exp}(N^{-1}\lambda)$.

Proof. By applying the formula for the distribution of the minimum:

$$\begin{aligned} F_{X_{(1)}}(x;\lambda,N) &= 1-\left[1-F_X(x;\lambda)\right]^N \\ &= 1-\left[\exp\left(-\frac{1}{\lambda}x\right)\right]^N \\ &= 1-\exp\left(-\frac{N}{\lambda}x\right) \end{aligned}$$

the postulated (cumulative) distribution is obtained straightforwardly. □

All three types of distributions in the GEV family, instead, are max-stable: this is another motivation for their collective name.

Observation 4.3. Consider a random sample drawn from the Type I GEV (Gumbel) distribution with parameters μ and σ, $X \sim \mathrm{EV1}(\mu,\sigma)$. The top order statistic – the maximum – is such that $X_{(N)} \sim \mathrm{EV1}(\mu-\sigma\log(N),\sigma)$.

Proof. By applying the formula for the distribution of the maximum:

$$\begin{aligned} F_{X_{(N)}}(x;\mu,\sigma,N) &= \left[F_X(x;\mu,\sigma)\right]^N \\ &= \exp\left(-\exp\left(\frac{x-\mu}{\sigma}\right)\right)^N \\ &= \exp\left(-N\exp\left(\frac{x-\mu}{\sigma}\right)\right) \\ &= \exp\left(-\exp\left(\frac{x-\mu+\sigma\log(N)}{\sigma}\right)\right) \end{aligned}$$

one obtains the Gumbel cumulative distribution that was argued. □

Observation 4.4. Consider a random sample drawn from the Type II GEV (Fréchet) distribution with parameters α, μ, and σ, $Y \sim \text{EV2}(\alpha, \mu, \sigma)$. The top order statistic – the maximum – is such that $Y_{(N)} \sim \text{EV2}\left(\alpha, \mu, \sigma N^{1/\alpha}\right)$. The result is identical in random sampling from the Type III GEV (reverse Weibull) case: with $Y \sim \text{EV3}(\alpha, \mu, \sigma)$, it is $Y_{(N)} \sim \text{EV3}\left(\alpha, \mu, \sigma N^{1/\alpha}\right)$.

Proof. Here, applying the formula for the distribution of the maximum:

$$\begin{aligned} F_{Y_{(N)}}(y; \alpha, \mu, \sigma, N) &= \left[F_Y(y; \alpha, \mu, \sigma)\right]^N \\ &= \exp\left(-\left(\frac{y-\mu}{\sigma}\right)^{-\alpha}\right)^N \\ &= \exp\left(-\left[N^{-\frac{1}{\alpha}}\left(\frac{y-\mu}{\sigma}\right)\right]^{-\alpha}\right) \end{aligned}$$

is equally valid to show both the Fréchet and the reverse Weibull results. □

A last observation about the traditional Weibull distribution – which should be intuitive in light of the relationships between the traditional Weibull, the exponential, and the GEV distributions – is presented next.

Observation 4.5. Consider a random sample drawn from the traditional Weibull distribution with parameters α, μ, and σ, $W \sim \text{Weibull}(\alpha, \mu, \sigma)$. The sample minimum is such that $W_{(1)} \sim \text{Weibull}\left(\alpha, \mu, \sigma N^{1/\alpha}\right)$.

Proof. Things proceed similarly to the Fréchet and reverse Weibull cases.

$$\begin{aligned} F_{W_{(1)}}(w; \alpha, \mu, \sigma, N) &= 1 - \left[1 - F_W(w; \alpha, \mu, \sigma)\right]^N \\ &= 1 - \exp\left(-\left(\frac{w-\mu}{\sigma}\right)^{-\alpha}\right)^N \\ &= 1 - \exp\left(-\left[N^{-\frac{1}{\alpha}}\left(\frac{w-\mu}{\sigma}\right)\right]^{-\alpha}\right) \end{aligned}$$

The difference is that here, the formula for the minimum is applied. □

It is difficult to identify other situations where the *exact* distribution of an order statistic of interest can be computed and related to some known common distribution. In an asymptotic environment things are different, as the so-called Extreme Value Theorem (see Lecture 6) allows to circumscribe the set of sampling distribution of order statistics to the three different types of the GEV family – so long as the sample size is large enough. Along with the above "stability" results, the Theorem in question motivates the use of the GEV distributions for modeling extreme order statistics.

4.4 Sufficient Statistics

After introducing the concept of a sample and the statistics that summarize its characteristics – the sample mean, the sample variance-covariance, and the order statistics – the next logical step is to use the sample in order to learn important facts about the population from which the sample is drawn. The final objective is to perform evaluations and tests about certain features of the probability distribution that generates the sample, such as selected parameters or moments; these exercises are known as *statistical estimation* and *inference*. Before reaching that point, however, it is useful to recognize that in selected situations – quite frequent ones actually – certain statistics help simplify statistical evaluations. These are called **sufficient statistics** and this long section is specifically devoted to them. A definition follows.

Definition 4.19. Sufficient statistics. Consider a sample generated by a list of random vectors $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$. Suppose that the joint distribution of the sample depends, among the others, on some parameter θ; write the associated probability mass or density function as $f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}(\mathbf{x}_1, \ldots, \mathbf{x}_N; \theta)$. A statistic $T = T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ is said to be *sufficient* if the joint distribution of the sample, conditional on it, does not depend on θ:

$$f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N|T}(\mathbf{x}_1, \ldots, \mathbf{x}_N | T(\mathbf{x}_1, \ldots, \mathbf{x}_N)) = \frac{f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}(\mathbf{x}_1, \ldots, \mathbf{x}_N; \theta)}{q_T(T(\mathbf{x}_1, \ldots, \mathbf{x}_N); \theta)}$$

where $q_T(T(\mathbf{x}_1, \ldots, \mathbf{x}_N); \theta)$ is the probability mass or density function of the sufficient statistic in question.

Note that the parameter θ disappears from the expression of the conditional density on the left-hand side above. Equivalently, this can be expressed by saying that the joint conditional density is constant as a function of θ.

The usefulness of sufficient statistics is that they intuitively "exhaust" all the information about θ that is contained in a sample. This aids estimation and inference in various ways which are best appreciated in more advanced treatments (some related results are mentioned in the subsequent Lecture). The role of sufficient statistics in inference is summarized by the following **statistical principle**, that is, a postulate (axiom) of statistical analysis.

Statistical Principle 1. Sufficiency. If $T = T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ is a sufficient statistic for a parameter θ, any evaluation about the latter should depend solely on the sufficient statistic or a function thereof. That is, if $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ and $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ are two – possibly different – sample realizations such that $T(\mathbf{x}_1, \ldots, \mathbf{x}_N) = T(\mathbf{y}_1, \ldots, \mathbf{y}_N)$, all evaluations about θ should be identical regardless of the exact observed values in either realization.

Examples are useful to build intuition; it is best start from simpler ones.

Example 4.1. A sufficient statistic for the Bernoulli parameter p. Suppose that a random (i.i.d.) sample is obtained from a random variable X following the Bernoulli distribution with parameter p, or $X \sim \mathrm{Be}(p)$. It turns out that the statistic counting the number of "successes," define it as:

$$T = T(X_1, \ldots, X_N) = \sum_{i=1}^{N} X_i$$

is a sufficient statistic for p. This is shown by applying the definition, after observing that $T \sim \mathrm{BN}(p, N)$. Call t the *realization* of T, that is the value of the sufficient statistic evaluated in terms of the N actual realizations of X that are observed *in the sample*:

$$t = T(x_1, \ldots, x_N) = \sum_{i=1}^{N} x_i$$

thus:

$$\begin{aligned} \frac{f_{X_1,\ldots,X_N}(x_1, \ldots, x_N; p)}{q_T(t; p, N)} &= \frac{\prod_{i=1}^{N} p^{x_i} (1-p)^{1-x_i}}{\binom{N}{t} p^t (1-p)^{N-t}} \\ &= \frac{p^t (1-p)^{N-t}}{\binom{N}{t} p^t (1-p)^{N-t}} \\ &= \frac{t!\,(N-t)!}{N!} \end{aligned}$$

where the first equality follows since the sample is random and the second from the definition of t. Thus it is proved that the distribution of the sample, conditional on the sufficient statistic, does not depend on p – as postulated. The intuition for this is that upon knowing the number of "successes" t over N attempts, there is no other information in the sample that helps "learn" (perform inference) about the parameter p. ■

Example 4.2. A sufficient statistic for μ in the normal distribution. Suppose that a random (i.i.d.) sample is obtained from a random variable X following the normal distribution with location parameter μ and scale parameter σ^2, or $X \sim \mathcal{N}(\mu, \sigma^2)$. The sample mean $\bar{X}$ is a sufficient statistic for the mean parameter μ. The demonstration proceeds as in the previous case, recalling now that $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/N)$. Similarly as above, it is useful to define the actual *realization* of the sample mean in the data.

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Proving the claim thus proceeds as above:

$$\begin{aligned}
\frac{f_{X_1,\ldots,X_N}(x_1,\ldots,x_N;\mu,\sigma^2)}{q_{\bar{X}}(\bar{x};\mu,\sigma^2/N)} &= \frac{\prod_{i=1}^{N}\sqrt{(2\pi\sigma^2)^{-1}}\cdot\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)}{\sqrt{(2\pi\sigma^2)^{-1}N}\cdot\exp\left(-\frac{N(\bar{x}-\mu)^2}{2\sigma^2}\right)} \\
&= \frac{\sqrt{(2\pi\sigma^2)^{-N}}\cdot\exp\left(-\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}\right)}{\sqrt{(2\pi\sigma^2)^{-1}N}\cdot\exp\left(-\frac{N(\bar{x}-\mu)^2}{2\sigma^2}\right)} \\
&= \frac{\exp\left(-\sum_{i=1}^{N}\frac{(x_i-\bar{x})^2}{2\sigma^2}-\frac{N(\bar{x}-\mu)^2}{2\sigma^2}\right)}{\sqrt{(2\pi\sigma^2)^{N-1}N}\cdot\exp\left(-\frac{N(\bar{x}-\mu)^2}{2\sigma^2}\right)} \\
&= \frac{\exp\left(-\sum_{i=1}^{N}\frac{(x_i-\bar{x})^2}{2\sigma^2}\right)}{\sqrt{(2\pi\sigma^2)^{N-1}N}}
\end{aligned}$$

where the third line obtains since:

$$\sum_{i=1}^{N}(x_i-\mu)^2=\sum_{i=1}^{N}(x_i-\bar{x})^2-2(\bar{x}-\mu)\underbrace{\sum_{i=1}^{N}(x_i-\bar{x})}_{=0}+N(\bar{x}-\mu)^2$$

which is a decomposition similar to that used to prove part **c.** of Theorem 4.4. Again, the final expression of the ratio does not depend on the location parameter μ, however it does depend on σ^2. The intuition is that the sample mean provides all the information that the sample can deliver about the *average* of the population being sampled, but it does not provide enough knowledge about its overall spread or *variation*. ■

Example 4.3. A sufficient statistic for the uniform distribution. An order statistics can be sufficient as well. Suppose that a random (i.i.d.) sample is obtained from a random variable X following the uniform distribution with the infimum of the support is zero, and the supremum is θ, an unknown parameter $X\sim\mathcal{U}(0,\theta)$. The maximum $X_{(N)}$ is indeed a sufficient statistic

for θ! Showing this result is quite simple: calling $x_{(N)} = \max\{x_1, \ldots, x_N\}$ the realization of $X_{(N)}$, and writing that the latter's density function as

$$\begin{aligned} q_{X_{(N)}}\left(x_{(N)};\theta\right) &= \frac{d}{dx_{(N)}}\left[\left(\frac{x_{(N)}}{\theta}\right)^N \cdot \mathbb{1}\left[x_{(N)} \in (0,\theta)\right]\right] \\ &= \frac{N x_{(N)}^{N-1}}{\theta^N} \cdot \mathbb{1}\left[x_{(N)} \in (0,\theta)\right] \end{aligned}$$

it is straightforward to see that:

$$\frac{f_{X_1,\ldots,X_N}(x_1,\ldots,x_N;\theta)}{q_{X_{(N)}}\left(x_{(N)};\theta\right)} = \frac{1}{N x_{(N)}^{N-1}} \cdot \mathbb{1}\left[x_{(N)} \in (0,\theta)\right]$$

as $f_{X_1,\ldots,X_N}(x_1,\ldots,x_N;\theta) = [f_X(x;\theta)]^N = \theta^{-N}$ because the sample is random. Again, the result is intuitive: since the support of the uniform distribution is bounded above, the highest value found in the sample is the most informative about the limit of the the support. ■

Sometimes, it is difficult to verify that a statistic is effectively sufficient for a certain parameter of interest. Fortunately, the following theorem often helps simplify the analysis.

Theorem 4.7. Fisher-Neyman's Factorization Theorem. *Consider a sample generated by a list of random vectors $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$, whose joint distribution has mass or density function $f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}(\mathbf{x}_1,\ldots,\mathbf{x}_N;\theta)$ that also depends on some parameter θ. A statistic $T = T(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)$ is* sufficient *for θ if and only if it is possible to identify two functions $g(T(\mathbf{x}_1,\ldots,\mathbf{x}_N);\theta)$ and $h(\mathbf{x}_1,\ldots,\mathbf{x}_N)$ such that the following holds.*

$$f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}(\mathbf{x}_1,\ldots,\mathbf{x}_N;\theta) = g(T(\mathbf{x}_1,\ldots,\mathbf{x}_N);\theta) \cdot h(\mathbf{x}_1,\ldots,\mathbf{x}_N)$$

Observe that function $g(T(\mathbf{x}_1,\ldots,\mathbf{x}_N);\theta)$ depends on θ, however function $h(\mathbf{x}_1,\ldots,\mathbf{x}_N)$ does not.

Proof. The logic of the proof is best illustrated in the **discrete case**, and it is helpful to start from there. Restricting the analysis to the discrete case, it best to begin by proving the "necessity" part of the theorem: if the above factorization exists, then T is sufficient for θ. Write the mass function of T as $q_T(T(\mathbf{x}_1,\ldots,\mathbf{x}_N);\theta)$. Furthermore, define the set of vectors spanning the same space as $(\mathbf{x}_1,\ldots,\mathbf{x}_N)$ and that result in the same values for T, as follows.

$$\mathbb{A}_T(\mathbf{x}_1,\ldots,\mathbf{x}_N) \equiv \{\mathbf{y}_1,\ldots,\mathbf{y}_N : T(\mathbf{x}_1,\ldots,\mathbf{x}_N) = T(\mathbf{y}_1,\ldots,\mathbf{y}_N)\}$$

By the property of probability functions, it is:

$$\begin{aligned} q_T\left(T\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right);\theta\right) &= \sum_{\mathbf{y}_1,\ldots,\mathbf{y}_N\in\mathbb{A}_T} f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}\left(\mathbf{y}_1,\ldots,\mathbf{y}_N;\theta\right) \\ &= \sum_{\mathbf{y}_1,\ldots,\mathbf{y}_N\in\mathbb{A}_T} g\left(T\left(\mathbf{y}_1,\ldots,\mathbf{y}_N\right);\theta\right)\cdot h\left(\mathbf{y}_1,\ldots,\mathbf{y}_N\right) \end{aligned}$$

and since $T\left(\mathbf{y}_1,\ldots,\mathbf{y}_N\right)$ is constant in $\mathbb{A}_T\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right)$, it is:

$$q_T\left(T\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right);\theta\right) = g\left(T\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right);\theta\right)\cdot \sum_{\mathbf{y}_1,\ldots,\mathbf{y}_N\in\mathbb{A}_T} h\left(\mathbf{y}_1,\ldots,\mathbf{y}_N\right)$$

where in both cases $\mathbb{A}_T$ is shorthand notation for $\mathbb{A}_T\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right)$. Then:

$$\begin{aligned} \frac{f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}\left(\mathbf{x}_1,\ldots,\mathbf{x}_N;\theta\right)}{q_T\left(T\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right);\theta\right)} &= \frac{g\left(T\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right);\theta\right)\cdot h\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right)}{q_T\left(T\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right);\theta\right)} \\ &= \frac{h\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right)}{\sum_{\mathbf{y}_1,\ldots,\mathbf{y}_N\in\mathbb{A}_T} h\left(\mathbf{y}_1,\ldots,\mathbf{y}_N\right)} \end{aligned}$$

as $g\left(T\left(\mathbf{y}_1,\ldots,\mathbf{y}_N\right);\theta\right)$ simplifies in the right hand side's ratio; the latter no longer depends on θ indicating that T is a sufficient statistic.

To prove that if T is sufficient the factorization of interest holds – the "sufficiency" part of the Theorem in the discrete case, a term which is unfortunate here – it is convenient to recall the interpretation of a joint mass function as a probability function:

$$\begin{aligned} f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}\left(\mathbf{x}_1,\ldots,\mathbf{x}_N;\theta\right) &= \mathbb{P}\left(\bigcup_{i=1}^{N}\boldsymbol{x}_i=\mathbf{x}_i;\theta\right) \\ &= \mathbb{P}\left(\bigcup_{i=1}^{N}\boldsymbol{x}_i=\mathbf{x}_i\,\middle|\,T=T\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right)\right) \\ &\quad\times\mathbb{P}\left(T=T\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right);\theta\right) \\ &= h\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right)\cdot q_T\left(T\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right);\theta\right) \end{aligned}$$

where the second line follows from the definition of conditional probability while the third just renames the previous probability function, noting that the conditional probability of the sample given T can be expressed as some generic function $h\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right)$ that does not depend on θ by definition of sufficient statistic.

(Sketched.) In the **continuous case** the logic of the proof is analogous; however, the proper demonstration requires the use of advanced measure

theory. Thus, the analysis is only outlined for a restricted case that can be related to the discrete case above, and is still general enough to allow many concrete situations. Suppose that there is a list of *bijective* and *differentiable* transformations that *does not depend on* θ denoted as:

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_N \end{pmatrix} = \begin{pmatrix} \boldsymbol{g}_1(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \\ \boldsymbol{g}_2(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \\ \vdots \\ \boldsymbol{g}_N(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \end{pmatrix}$$

where any element of this list, suppose the first element Y_{11} of $\boldsymbol{y}_1$, is fixed as $Y_{11} = T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ *by construction*. In addition, write the corresponding *inverse* transformation as follows.

$$\begin{pmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \\ \vdots \\ \boldsymbol{w}_N \end{pmatrix} = \begin{pmatrix} \boldsymbol{g}_1^{-1}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) \\ \boldsymbol{g}_2^{-1}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) \\ \vdots \\ \boldsymbol{g}_N^{-1}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) \end{pmatrix}$$

To show necessity, write the joint density of the transformation as:

$$\begin{aligned} f_{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N}(\mathbf{y}_1, \ldots, \mathbf{y}_N; \theta) &= f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}(\mathbf{w}_1, \ldots, \mathbf{w}_N; \theta) \cdot |\mathbf{J}^*| \\ &= g(T(\mathbf{w}_1, \ldots, \mathbf{w}_N); \theta) \cdot h(\mathbf{w}_1, \ldots, \mathbf{w}_N) \cdot |\mathbf{J}^*| \\ &= g(y_{11}; \theta) \cdot h(\mathbf{w}_1, \ldots, \mathbf{w}_N) \cdot |\mathbf{J}^*| \end{aligned}$$

where $|\mathbf{J}^*|$ is shorthand notation for the absolute value of the Jacobian of the inverse transformation, and the second line follows from hypothesis. It is obvious that the marginal distribution of Y_{11}, that is the density function $q_T(T(\mathbf{x}_1, \ldots, \mathbf{x}_N); \theta)$ of the statistic of interest T, inherits a factorization analogous to the above and thus since $y_{11} = T(\mathbf{x}_1, \ldots, \mathbf{x}_N)$, it can be shown that the ratio between the joint density of the sample and the density of T does not depend on θ, hence T is sufficient. To prove that if T is sufficient then a proper factorization can be expressed (the "sufficiency" part of the Theorem), apply the definition of conditional density function to show that:

$$f_{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N}(\mathbf{y}_1, \ldots, \mathbf{y}_N; \theta) = q_T(y_{11}; \theta) \cdot f_{\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N\} \setminus Y_{11}}(\{\mathbf{y}_1, \ldots, \mathbf{y}_N\} \setminus y_{11} | Y_{11})$$

where the notation $\{\cdot\} \setminus Y_{11}$ denotes a list that *excludes* Y_{11}. Dividing both sides of the above by $|\mathbf{J}^*|$ returns the desired factorization for:

$$h(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \frac{f_{\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N\} \setminus Y_{11}}(\{\mathbf{y}_1, \ldots, \mathbf{y}_N\} \setminus y_{11} | Y_{11})}{|\mathbf{J}^*|}$$

and for $g(T(\mathbf{x}_1, \ldots, \mathbf{x}_N); \theta) = q_T(T(\mathbf{x}_1, \ldots, \mathbf{x}_N); \theta)$. □

The factorization theorem can be easily applied to cases like the previous examples. However, it is particularly useful to show that multiple statistics are simultaneously sufficient for a given number of associated parameters. This is usually expressed through a *vector* of statistics $\boldsymbol{t}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$:

$$\boldsymbol{t}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \begin{pmatrix} T_1(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \\ T_2(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \\ \vdots \\ T_K(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \end{pmatrix}$$

which are said to be *simultaneously sufficient* for a *vector of parameters* $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_J \end{pmatrix}$$

where in general, it may as well be that $K \neq J$. The factorization theorem can be extended to allow for $g(\boldsymbol{t}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N); \boldsymbol{\theta})$ to be the joint density of all the statistics in question and for a multidimensional parameter vector.

Example 4.4. Two sufficient statistics for μ and σ^2 in the normal distribution. Let us revisit Example 4.2. There, the factorization theorem can be expressed for:

$$g(\bar{x}; \mu) = \exp\left(-\frac{N(\bar{x} - \mu)^2}{2\sigma^2}\right)$$

and:

$$h(x_1, \ldots, x_N) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} \exp\left(-\sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{2\sigma^2}\right)$$

and the product of these functions returns the joint density of the sample $X_1, \ldots, X_N$. Observe that, however, both expressions still incorporate the parameter σ^2. To obtain a sufficient statistic for it, it is intuitive to think of the sample variance S^2, whose realization is usually written as follows.

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

It is easy to verify that the factorization theorem here applies with:

$$g(\bar{x}, s^2; \mu, \sigma^2) = \left(\frac{1}{\sigma^2}\right)^{\frac{N}{2}} \exp\left(-\frac{N(\bar{x} - \mu)^2 + (N-1)S^2}{2\sigma^2}\right)$$

and $h(x_1, \ldots, x_N) = (2\pi)^{-N/2}$, implying that the pair of statistics $\left(\bar{X}, S^2\right)$ is *simultaneously* sufficient for the vector of parameters (μ, σ^2). This result is intuitive again, since σ^2 equals the variance in the normal case. ■

Example 4.5. Sufficient statistics for μ and Σ in the multivariate normal distribution. Suppose that a random (i.i.d.) sample is obtained from a random vector $\boldsymbol{x}$ following the *multivariate* normal distribution with parameters collected as $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$: $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The multivariate sample mean $\bar{\boldsymbol{x}} = \left(\bar{X}_1, \ldots, \bar{X}_K\right)$ is a collection of sufficient statistics for the mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, thus extending the result from Example 4.2 to the multivariate case. More accurately, one should say that the vector of sample means $\bar{\boldsymbol{x}}$ features K sufficient statistics that are *simultaneously* sufficient for the K parameters contained in the mean vector $\boldsymbol{\mu}$.

In order to show this one must observe that $\bar{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}/N)$ by the previous results on the multivariate sample mean and the linear properties of the normal distribution. Also define – similarly to the univariate case – the realization of the sample mean as follows.

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$$

The result from Example 4.2 can now be expressed as:

$$\frac{f_{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N}(\mathbf{x}_1, \ldots, \mathbf{x}_N; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{q_{\bar{\boldsymbol{x}}}(\bar{\mathbf{x}}; \boldsymbol{\mu}, \boldsymbol{\Sigma}/N)} = \frac{\exp\left(-\frac{1}{2} \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \right)}{\sqrt{\left[(2\pi)^K |\boldsymbol{\Sigma}| \right]^{N-1} N}}$$

where K is, as usual, the dimension of the random vector $\boldsymbol{x}$; the intermediate steps involve some tedious linear algebra. The intuition is that every random variable listed in $\boldsymbol{x}$, say X_k, follows a marginal distribution which is normal with location parameter μ_k; hence the sample mean $\bar{X}_k$ – which is listed in $\bar{\boldsymbol{x}}$ – exhausts all the information contained in the sample about that particular parameter, and this holds simultaneously for all $k = 1, \ldots, K$.

In analogy with in the univariate case, these observations can be pushed even further by claiming that the vector of sample means $\bar{\boldsymbol{x}}$ and the sample variance-covariance $\boldsymbol{S}$ are *simultaneously* sufficient for all the parameters of the multivariate normal distribution, $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The *realization* of the sample variance-covariance, written as $\mathbf{S}$, is as follows.

$$\mathbf{S} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}}$$

By means of some algebraic manipulation, one can show that the function:[3]

$$g\left(\bar{\mathbf{x}}, \mathbf{S}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \frac{1}{|\boldsymbol{\Sigma}|^{\frac{N}{2}}} \exp\left(-\frac{N}{2}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right) - \frac{N-1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\right)\right)$$

complies to the factorization theorem along with $h\left(\mathbf{x}_1, \ldots, \mathbf{x}_N\right) = (2\pi)^{-\frac{NK}{2}}$. To develop intuition, it is important to recall that the matrix $\boldsymbol{\Sigma}$ not only features the K variances of each normally distributed random variable listed in $\boldsymbol{x}$, but also the $K\left(K-1\right)/2$ covariances. The sample variance-covariance $\boldsymbol{S}$ provides appropriate sufficient statistics for all these parameters. ■

Example 4.6. Two sufficient statistics for the uniform distribution. Suppose that a random (i.i.d.) sample is obtained from a random variable X following the uniform distribution with *unknown support* $X \sim \mathcal{U}\left(\alpha, \beta\right)$. Here, the bounds of the support are written with the upright Greek letters α and β to remark that they shall be treated as *parameters*. The two extreme

[3]The calculations are as follows. The joint density of the sample is:

$$f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}\left(\mathbf{x}_1, \ldots, \mathbf{x}_N; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \left(\frac{1}{(2\pi)^K |\boldsymbol{\Sigma}|}\right)^{\frac{N}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{x}_i-\boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_i-\boldsymbol{\mu}\right)\right)$$

where, by recognizing that $\left(\mathbf{x}_i-\boldsymbol{\mu}\right) = \left(\mathbf{x}_i-\bar{\mathbf{x}}+\bar{\mathbf{x}}-\boldsymbol{\mu}\right)$, the term inside the exponential develops as follows.

$$\begin{aligned}\sum_{i=1}^{N}\left(\mathbf{x}_i-\boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_i-\boldsymbol{\mu}\right) = \sum_{i=1}^{N}\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_i-\bar{\mathbf{x}}\right) + N\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right) + \\ + \underbrace{\sum_{i=1}^{N}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)}_{=0} + \underbrace{\sum_{i=1}^{N}\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right)}_{=0}\end{aligned}$$

The last two terms are zero since $\sum_{i=1}^{N}\left(\mathbf{x}_i-\bar{\mathbf{x}}\right) = \mathbf{0}$. The first term in the decomposition instead develops, by the property of the trace operator, as:

$$\begin{aligned}\sum_{i=1}^{N}\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_i-\bar{\mathbf{x}}\right) &= \mathrm{tr}\left(\sum_{i=1}^{N}\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)\right) \\ &= \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N}\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)^{\mathrm{T}}\right) \\ &= \left(N-1\right)\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\right)\end{aligned}$$

where the last line follows from the definition of $\mathbf{S}$. Collecting terms allows to verify that the factorization $f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}\left(\mathbf{x}_1, \ldots, \mathbf{x}_N; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = g\left(\bar{\mathbf{x}}, \mathbf{S}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \cdot h\left(\mathbf{x}_1, \ldots, \mathbf{x}_N\right)$ holds with the expressions given in the text.

order statistics, the minimum $X_{(1)}$ (with $x_{(1)} = \min\{x_1, \ldots, x_N\}$) and the maximum $X_{(N)}$ (with $x_{(N)} = \max\{x_1, \ldots, x_N\}$ as in Example 4.3) are in fact simultaneously sufficient for (α, β). This is shown by observing that the joint density function of the sample is:

$$f_{X_1,\ldots,X_N}(x_1, \ldots, x_N; \alpha, \beta) = \left(\frac{1}{\beta - \alpha}\right)^N \cdot \mathbb{1}[\alpha \leq x_1, \ldots, x_N \leq \beta]$$

and by setting:

$$g\left(x_{(1)}, x_{(N)}; \alpha, \beta\right) = \left(\frac{1}{\beta - \alpha}\right)^N \cdot \mathbb{1}\left[\alpha \leq x_{(1)}\right] \cdot \mathbb{1}\left[x_{(N)} \leq \beta\right]$$

and $h(x_1, \ldots, x_N) = 1$. Hence, the factorization theorem applies trivially. One more time, the result is intuitive: if both bounds of the uniform distribution are unknown, there is no better information contained in the sample than the two extreme order statistics. ■

The factorization theorem allows to quickly verify that certain statistics are sufficient for the specified parameters of an important "macrofamily" of probability distributions, which is defined next.

Definition 4.20. Exponential Family. A family of probability distributions characterized by a vector of parameters $\theta = (\theta_1, \ldots, \theta_J)$ is said to belong to the *exponential* (macro)-family if the associated mass or density functions can be written, for $J \leq L$, as:

$$f_X(x; \theta) = h(x)\, c(\theta) \exp\left(\sum_{\ell=1}^{L} w_\ell(\theta)\, t_\ell(x)\right)$$

where $h(x)$ and $t_\ell(x)$ are functions of the realizations x while $c(\theta) \geq 0$ and $w_\ell(\theta)$ are functions of the parameters θ, and in both cases, $\ell = 1, \ldots, L$.

The exponential macrofamily is extensive: it comprises many of the distribution families analyzed in Lecture 2. In particular, the discrete Bernoulli, geometric, Poisson families; as well as the continuous normal, lognormal, Beta and Gamma families – including the special cases of the Gamma family, like the chi-squared and exponential distribution – are all sub-families of the exponential macro-family.[4] All these claims can be verified by manipulating of the density functions of interest. Other distributions are said to belong to the exponential family so long as certain parameters are "fixed" (i.e. not part of θ in the above definition): this is the case of the binomial and negative binomial families for a constant number of trials n or r.

[4]One must be careful at not mistaking the exponential (macro-)family for the more restricted subfamily of exponential distributions!

The connection between sufficient statistics and the exponential family is expressed through the following result.

Theorem 4.8. Sufficient statistics and the exponential family. *If a random sample is obtained from any random variable X whose distribution belongs to the exponential family, the L statistics in the vector:*

$$\boldsymbol{t}\left(X_1,\ldots,X_N\right)=\begin{pmatrix}\sum_{i=1}^{N}t_1\left(X_i\right)\\ \sum_{i=1}^{N}t_2\left(X_i\right)\\ \vdots\\ \sum_{i=1}^{N}t_L\left(X_i\right)\end{pmatrix}$$

are simultaneously sufficient for θ, where the functions $t_\ell(x)$ are as in the previous definition of the exponential family for $\ell=1,\ldots,L$.

Proof. The joint density of the sample can be expressed as:

$$f_{X_1,\ldots,X_N}\left(x_1,\ldots,x_N;\theta\right)=\left(\prod_{i=1}^{N}h\left(x_i\right)\right)\left[c\left(\theta\right)\right]^N\exp\left(\sum_{\ell=1}^{L}w_\ell\left(\theta\right)\sum_{i=1}^{N}t_\ell\left(x_i\right)\right)$$

and applying the factorization theorem is straightforward. □

Example 4.7. A sufficient statistic for the Bernoulli distribution, revisited. It might not appear too obvious at first, but the Bernoulli family of distributions for $p\in(0,1)$ is a full member of the exponential family. Its density function can be rewritten, for $x\in\{0,1\}$, as:

$$\begin{aligned}f_X\left(x,p\right)&=\left(1-p\right)\left(\frac{p}{1-p}\right)^x\\&=\left(1-p\right)\exp\left(\log\left(\frac{p}{1-p}\right)x\right)\end{aligned}$$

and so the sufficient statistic $T=\sum_{i=1}^{N}X_i$ from Example 4.1 complies with Theorem 4.8. ■

Example 4.8. Two sufficient statistics for the normal distribution, a member of the exponential family. Consider some random variable $X\sim\mathcal{N}\left(\mu,\sigma^2\right)$. Its density function can be rewritten as:

$$f_X\left(x;\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\mu^2}{2\sigma^2}\right)\exp\left(\frac{\mu}{\sigma^2}x-\frac{1}{2\sigma^2}x^2\right)$$

therefore, by Theorem 4.8 two statistics that are simultaneously sufficient for μ and σ^2 are $T_1=\sum_{i=1}^{N}X_i$ and $T_2=\sum_{i=1}^{N}X_i^2$. These are unfamiliar in the context of normal distributions; it is soon shown how they relate to the more frequent sample mean $\bar{X}$ and sample variance S^2. ■

Example 4.9. Two sufficient statistics for the Gamma distribution. The density function of random variable $X \sim \text{Gamma}(\alpha, \beta)$ can be written as:

$$f_X(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \exp[(\alpha - 1)\log(x) - \beta x]$$

for $x > 0$. By Theorem 4.8, two statistics that are simultaneously sufficient for α and β are $T_1 = \sum_{i=1}^N \log(X_i)$ and $T_2 = \sum_{i=1}^N X_i$. A simpler analysis extends to the two special cases of the Gamma distribution, the exponential and the chi-squared distributions. In both cases, $T = \sum_{i=1}^N X_i$ is a sufficient statistic for the unknown parameter (λ and κ respectively). ■

The treatment of sufficient statistics is concluded, along the entire Lecture, by observing that it is easy to obtain alternative sufficient statistics for the same parameters through appropriate transformations; if $T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ is a sufficient statistic for some parameter of interest θ, also the transformation $T'(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = g(T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N))$ results in a sufficient for θ if $g(\cdot)$ does not depend on θ. This follows from the definition of sufficient statistics and the theorems about the transformation of random variables. These considerations extend to transformations of *vectors* of sufficient statistics, $\boldsymbol{t}'(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \boldsymbol{g}(\boldsymbol{t}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N))$, as in the following examples.

Example 4.10. Two sufficient statistics for the normal distribution, a member of the exponential family, revisited. The two seemingly different results from Examples 4.4 and 4.8 can be reconciled by observing that, given $T_1 = \sum_{i=1}^N X_i$ and $T_2 = \sum_{i=1}^N X_i^2$, it is:

$$\bar{X} = \frac{1}{N} T_1 \quad \text{and} \quad S^2 = \frac{1}{N-1}\left(T_2 - \frac{T_1^2}{N}\right)$$

a transformation that does not depend on the parameters. ■

Example 4.11. Two sufficient statistics for the Gamma distribution, revisited. The two sufficient statistics for α and β in random samples drawn from the Gamma distribution are typically listed as:

$$T_1' = \prod_{i=1}^N X_i \quad \text{and} \quad T_2' = \sum_{i=1}^N X_i$$

which is easily be verified via direct application of the factorization theorem. This can be related to Example 4.9 since $T_1' = \exp(T_1)$ and $T_2' = T_2$. ■

With the help of examples, this section has displayed multiple methods to obtain the sufficient statistics of interest. The most appropriate method is generally context-dependent, and it is often useful to verify that alternative routes can lead to the same result.

Lecture 5

Statistical Inference

This lecture develops the core concepts of statistical inference: the theory and practice of the statistical evaluation of data. After having introduced the concept of point estimator and two chief methods for constructing estimators – the Method of Moments and Maximum Likelihood Estimation – this lecture discusses a framework and associated results for the evaluation of the statistical properties of different estimators. Finally, this lecture concludes with an outline of the theory and the practice of hypothesis testing in statistical inference and the associated methods to construct confidence intervals for estimators, the so-called interval estimation.

5.1 Principles of Estimation

The term *estimation* refers to a broad concept that includes different types of *evaluations* about certain features of the probability distributions that are hypothesized to generate a sample of data. Here the focus is on **parameter estimation**: the use of selected statistics for evaluating the *parameters* of such a distribution. There are multiple methods to perform parameter estimation, and each is motivated by some *statistical principle*. This section introduces two of them: the *Method of Moments* and *Maximum Likelihood Estimation*. Other methods, such as the so-called "Bayesian" ones, are instead outside the scope of this lecture. It must be mentioned that not every type of statistical estimation concerns the parameters of a distribution. The theory and the practice of *non-parametric estimation*, for example, concerns the direct evaluation of the density or the mass functions that are believed to generate the data, without making specific hypotheses about the functional form – and the parameters – of these distributions. Following this general introduction, a first definitions is in order.

Definition 5.1. (Point) estimators, and their estimates. Any statistic, if used to make evaluations about certain features of a probability distribution, is called a *point estimator* (or more simply, an *estimator*). The sample realization of such a statistic is called an *estimate*.

The notation $\widehat{\boldsymbol{\theta}}$, with the typical "hat," is typically used to denote a point estimator for some parameters $\boldsymbol{\theta}$ of a distribution (that are possibly multivalued). This notation is used for both estimators intended as statistics, that is random variables or vectors endowed with a sampling distribution, and for the estimates calculated in the data. The ensuing discussion treats the parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ that are sought after as vector-valued with dimension K; the univariate (scalar) case can be considered as a particular one (but with examples aplenty). Note that depending on the context, some values may or may not admissible estimates for certain parameters. For example, the scale parameter σ of location-scale families, or the parameters α and β of the Gamma distribution, cannot be negative. It is important to accurately define the set of values that are allowed in the estimation.

Definition 5.2. Parameter space. The set of admissible values for the parameters $\boldsymbol{\theta}$ is called *parameter space* and is usually denoted as $\boldsymbol{\Theta} \subseteq \mathbb{R}^K$.

The first of the two method to find or construct estimators that is introduced here is both the most intuitive and the oldest one (unsurprisingly). The **Method of Moments** is based on the following idea, formulated as a statistical principle.

Statistical Principle 2. Analogy. The *analogy principle* states that if the random variables that generate the sample and the parameters are related via some vector-valued function $\boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\theta})$ of dimension K, such that for $i = 1, \ldots, N$ a *zero moment condition* can be established:

$$\mathbb{E}\left[\boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\theta})\right] = \mathbf{0} \tag{5.1}$$

then a point estimator for $\boldsymbol{\theta}$ can be obtained as the solution to the so-called *sample analogue* of the zero moment condition, that is the condition that equates the sample mean of $\boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\theta})$ to zero.

$$\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{m}\left(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}_{MM}\right) = \mathbf{0} \tag{5.2}$$

Here, the estimator $\widehat{\boldsymbol{\theta}}_{MM}$ is denoted by the subscript that identifies it as a Method of Moments (MM) estimator.

The intuition behind the method of moments estimator is simple: because $\boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\theta})$ is a random vector with mean zero, the expectation of its sample mean must also be zero (Theorem 4.3). The most intuitive way to simulate this property with real data is to select a value for $\boldsymbol{\theta}$ that satisfies the mean zero requirement *in the sample*. Note that this definition is not restricted to random samples. Also observe that the method of moments does not exploit any characteristic of the joint distribution of the sample other than the zero moment condition (which is expressed as a function of the parameters).

Example 5.1. Estimation of the mean. There are several probability distributions such that their mean exactly equals one of the parameters. These include the Bernoulli, Poisson, normal, logistic, Laplace, exponential and others. Write the parameter in question as μ (which corresponds to p in the Bernoulli case, λ in the Poisson and exponential cases, etc.). Suppose that a random sample is obtained from one of these distributions. The zero moment condition here is simply:

$$\mathbb{E}[X_i - \mu] = \mathbb{E}[m(X_i; \mu)] = 0 \tag{5.3}$$

and the Method of Moments estimator is obtained as follows.

$$\widehat{\mu}_{MM} = \frac{1}{N}\sum_{i=1}^{N} X_i = \bar{X} \tag{5.4}$$

Consider next a multivariate distribution whose mean satisfies the following:

$$\mathbb{E}[\boldsymbol{x}_i - \boldsymbol{\mu}] = \mathbb{E}[\boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\mu})] = \mathbf{0} \tag{5.5}$$

as for the multivariate normal distribution, and others. Once again:

$$\widehat{\boldsymbol{\mu}}_{MM} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i = \bar{\boldsymbol{x}} \tag{5.6}$$

the Method of Moments estimator is the multivariate sample mean. ■

Example 5.2. Estimation of the variance (and covariance). The zero moment conditions naturally extend to moments higher than the mean, as all moments are ultimately expectations. For example, in random sampling from the normal distribution it holds that:

$$\begin{aligned} \mathbb{E}[X_i - \mu] &= \mathbb{E}\left[m_1\left(X_i; \mu, \sigma^2\right)\right] = 0 \\ \mathbb{E}\left[(X_i - \mathbb{E}[X])^2 - \sigma^2\right] &= \mathbb{E}\left[m_2\left(X_i; \mu, \sigma^2\right)\right] = 0 \end{aligned} \tag{5.7}$$

where the second condition can also be expressed as $\mathbb{E}[X_i^2] - \mu^2 - \sigma^2 = 0$. Here, the moment condition (5.3) is combined with another moment about the variance to result in a system of two equations and two unknowns.

The solution of this particular system is a pair of Method of Moments estimators expressed by (5.4) for the estimator of the location parameter μ, and the following expression for the estimator of the scale parameter σ^2.

$$\widehat{\sigma}^2_{MM} = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2 = \frac{N-1}{N}S^2 \tag{5.8}$$

Observe that this estimator differs from the sample variance S^2 by a factor $\frac{N-1}{N}$, hence its expectation does not equal the actual variance of X_i in a random sample. The method can be extended to other distributions; in the logistic case for example under the standard parametrization the variance is $\mathbb{V}\mathrm{ar}[X_i] = \sigma^2\pi^2/3$, therefore $\widehat{\sigma}_{MM} = S\sqrt{3(N-1)/N}/\pi$. Next, consider the multivariate normal distribution; there:

$$\begin{aligned} \mathbb{E}\left[\boldsymbol{x}_i - \boldsymbol{\mu}\right] &= \mathbb{E}\left[\boldsymbol{m}_{\boldsymbol{\mu}}\left(\boldsymbol{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\right] = \mathbf{0} \\ \mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} - \mathbb{E}\left[\boldsymbol{x}_i\right]\mathbb{E}\left[\boldsymbol{x}_i\right]^{\mathrm{T}} - \boldsymbol{\Sigma}\right] &= \mathbb{E}\left[\boldsymbol{m}_{\boldsymbol{\Sigma}}\left(\boldsymbol{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\right] = \mathbf{0} \end{aligned} \tag{5.9}$$

where the second set of conditions also writes as $\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right] - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\Sigma} = \mathbf{0}$ (note that this is a matrix-valued condition). The sample analogue of these moment conditions delivers as solution another set of Method of Moments estimators. These estimators are the sample mean as per (5.6) for $\boldsymbol{\mu}$, and:

$$\widehat{\boldsymbol{\Sigma}}_{MM} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}} = \frac{N-1}{N}\boldsymbol{S} \tag{5.10}$$

that is a rescaled version of the sample variance-covariance, for $\boldsymbol{\Sigma}$. ■

Example 5.3. Combined estimation of moments. Consider sampling from the Gamma distribution, where two zero moment conditions:

$$\mathbb{E}\left[X_i - \frac{\alpha}{\beta}\right] = \mathbb{E}\left[m_1\left(X_i; \alpha, \beta\right)\right] = 0 \tag{5.11}$$

$$\mathbb{E}\left[X_i^2 - \frac{\alpha(\alpha+1)}{\beta^2}\right] = \mathbb{E}\left[m_2\left(X_i; \alpha, \beta\right)\right] = 0 \tag{5.12}$$

deliver a system of two equations in two unknown parameters. The Method of Moments estimators of α and β are:

$$\widehat{\alpha}_{MM} = \frac{\bar{X}^2}{\frac{1}{N}\sum_{i=1}^{N}X_i^2 - \bar{X}^2} \tag{5.13}$$

$$\widehat{\beta}_{MM} = \frac{\bar{X}}{\frac{1}{N}\sum_{i=1}^{N}X_i^2 - \bar{X}^2} \tag{5.14}$$

where $\bar{X} = \frac{1}{N}\sum_{i=1}^{N}X_i$ is the sample mean. ■

Example 5.4. Estimation of the bivariate linear regression model. Remember the bivariate linear regression model from Example 3.11. There, the application of the analogy principle to the two parameters of interest is straightforward, because the covariance and the variance that define (3.11) have simple sample analogues.

$$\widehat{\beta}_{0,MM} = \bar{Y} - \bar{X} \cdot \widehat{\beta}_{1,MM} \tag{5.15}$$

$$\widehat{\beta}_{1,MM} = \frac{\sum_{i=1}^{N} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2} \tag{5.16}$$

Here, $\bar{X}$ and $\bar{Y}$ are the sample means of X_i and Y_i respectively, and the two estimators are also called the **least squares** estimators of the model, for reasons to be elaborated in Lecture 7. Note that to derive these estimators independence is not necessary, since the two quantities are obtained directly from (3.10) and (3.11). Furthermore, no specification of the joint density of Y_i and X_i was made, except that the two variables are related via a linear conditional expectation function $\mathbb{E}\left[Y_i \middle| X_i\right] = \beta_0 + \beta_1 X_i$. ■

The Method of Moments is simple and often convenient, since it requires fewer assumptions than its leading competitor for parametric estimation: Maximum Likelihood Estimation. Furthermore, the Method of Moments is virtually applicable to all statistical settings (those cases where the moments are infinite or undefined, like with the Cauchy distribution, are exceptions). As it is discussed later though, Maximum Likelihood has generally superior statistical properties than the Method of Moments: thus a trade-off between simplicity and flexibility on one side, and improved properties on the other side, arises. To understand this, it is necessary to introduce the Maximum Likelihood Estimator, starting from the definition of **likelihood function**.

Definition 5.3. The Likelihood Function. Suppose that some sample of observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ is observed. For fixed values of those realizations, the *likelihood* function is defined as the joint mass or density function of the sample, $f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}(\mathbf{x}_1, \ldots, \mathbf{x}_N; \boldsymbol{\theta})$, as a function *of the parameters*; it is generally written as follows.

$$\mathcal{L}\left(\boldsymbol{\theta} \middle| \mathbf{x}_1, \ldots, \mathbf{x}_N\right) = f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}(\mathbf{x}_1, \ldots, \mathbf{x}_N; \boldsymbol{\theta}) > 0$$

The likelihood function is by definition always positive because only values in the support of $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ can be observed.

The likelihood function cannot be interpreted as a sort of probability function, since it certainly does not integrate to 1 (and parameters are not

functions of events either). However, it bears an interpretation in terms of "how plausible" alternative parameter sets are for generating the observed realizations. This function is associated with the following principle.

Statistical Principle 3. Likelihood. Suppose that two samples of observations, $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, are observed, and they are obtained from distributions with the same unknown parameters $\boldsymbol{\theta}$. Suppose that the likelihood functions associated with the two realizations are proportional, in the sense that there exists a constant, expressed as a function of the observations $C(\mathbf{x}_1, \ldots, \mathbf{x}_N; \mathbf{y}_1, \ldots, \mathbf{y}_N)$, such that the two likelihood functions are always identical up to this constant:

$$\mathcal{L}(\boldsymbol{\theta} | \mathbf{x}_1, \ldots, \mathbf{x}_N) = C(\mathbf{x}_1, \ldots, \mathbf{x}_N; \mathbf{y}_1, \ldots, \mathbf{y}_N) \cdot \mathcal{L}(\boldsymbol{\theta} | \mathbf{y}_1, \ldots, \mathbf{y}_N)$$

where "always" means for every admissible value of the parameters $\boldsymbol{\theta}$. Then, any evaluation about the parameters should be identical in the two samples.

The likelihood principle has two main interpretations. The first is that if $C(\mathbf{x}_1, \ldots, \mathbf{x}_N; \mathbf{y}_1, \ldots, \mathbf{y}_N) = 1$ uniformly for all alternative pairs of samples, then two observations with the same value of the likelihood function imply identical "evaluations" (that is, estimations) about the parameters $\boldsymbol{\theta}$. The second implication is that for any two alternative values of the parameters, say $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}''$, the ratio

$$\frac{\mathcal{L}(\boldsymbol{\theta}' | \mathbf{x}_1, \ldots, \mathbf{x}_N)}{\mathcal{L}(\boldsymbol{\theta}'' | \mathbf{x}_1, \ldots, \mathbf{x}_N)} = \frac{\mathcal{L}(\boldsymbol{\theta}' | \mathbf{y}_1, \ldots, \mathbf{y}_N)}{\mathcal{L}(\boldsymbol{\theta}'' | \mathbf{y}_1, \ldots, \mathbf{y}_N)}$$

must be constant across any different sets of observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$. Consequently, if one treats the value expressed by the likelihood function as a measure of "plausibility" that the parameter in question is the one that generates the observations, the highest such value *does not depend on the particular observations*. The next logical step is to define an estimator which does select the value in question.

A **Maximum Likelihood Estimator** is a statistic that maximizes the observed likelihood function. Such estimator is usually specified as:

$$\widehat{\boldsymbol{\theta}}_{MLE} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\max}\, \mathcal{L}(\boldsymbol{\theta} | \mathbf{x}_1, \ldots, \mathbf{x}_N) \tag{5.17}$$

where the subscript "MLE" has an obvious meaning. Since the likelihood function is always positive, in practical settings it is often useful to maximize its logarithm instead, which is called the **log-likelihood function**. In other words, (5.17) is equivalent to the following.

$$\widehat{\boldsymbol{\theta}}_{MLE} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\max} \log \mathcal{L}(\boldsymbol{\theta} | \mathbf{x}_1, \ldots, \mathbf{x}_N) \tag{5.18}$$

If the sample has certain properties, the calculation of the MLE is simplified further. First, if the observations are independent the joint mass or density of the sample reduces to the product of the mass or density functions of all the observations, and so maximizing the log-likelihood function amounts to maximize a summation:

$$\widehat{\boldsymbol{\theta}}_{MLE} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max} \log\left[\prod_{i=1}^{N} f_{\boldsymbol{x}_i}\left(\mathbf{x}_i;\boldsymbol{\theta}\right)\right] = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max} \sum_{i=1}^{N} \log f_{\boldsymbol{x}_i}\left(\boldsymbol{\theta}\middle|\,\mathbf{x}_i\right) \quad (5.19)$$

where $\log f_{\boldsymbol{x}_i}\left(\mathbf{x}_i;\boldsymbol{\theta}\right)$ is the so-called "observation-specific component" of the log-likelihood function. Furthermore, if the observations are also identically distributed (the sample is random) it is $\log f_{\boldsymbol{x}_i}\left(\boldsymbol{\theta}\middle|\,\mathbf{x}_i\right) = \log f_{\boldsymbol{x}}\left(\boldsymbol{\theta}\middle|\,\mathbf{x}_i\right)$: the observation-specific component is identical for all $i = 1, \ldots, N$.

Example 5.5. Maximum Likelihood Estimation of N independent Bernoulli trials. Suppose one is interested to estimate the parameter p that generates the realizations $\{x_1, \ldots, x_N\}$ out of N independent Bernoulli trials. Note that here the parameter space of p is $\boldsymbol{\Theta} = [0, 1]$. The likelihood function is:

$$\mathcal{L}\left(p\middle|\,x_1, \ldots, x_N\right) = \prod_{i=1}^{N} p^{x_i}\left(1-p\right)^{1-x_i}$$

and the log-likelihood function is as follows.

$$\log\mathcal{L}\left(p\middle|\,x_1, \ldots, x_N\right) = \left(\sum_{i=1}^{N} x_i\right)\cdot\log\left(p\right) + \left(N - \sum_{i=1}^{N} x_i\right)\cdot\log\left(1-p\right)$$

The First Order Condition with respect to p is:

$$\frac{d\log\mathcal{L}\left(\widehat{p}_{MLE}\middle|\,x_1, \ldots, x_N\right)}{dp} = \frac{\sum_{i=1}^{N} x_i}{\widehat{p}_{MLE}} - \frac{N - \sum_{i=1}^{N} x_i}{1 - \widehat{p}_{MLE}} = 0$$

solving for which allows to verify that the Maximum Likelihood Estimator for this problem is the sample mean.

$$\widehat{p}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} X_i = \bar{X}$$

Two observations are in order. First, since $X_i \in \{0, 1\}$ it is $\bar{X} \in [0, 1]$, hence the MLE is restricted to valid values in the parameter space. Second, the Second Order Condition is as follows:

$$\frac{d^2\log\mathcal{L}\left(p\middle|\,x_1, \ldots, x_N\right)}{dp^2} = -\frac{\sum_{i=1}^{N} x_i}{p^2} - \frac{N - \sum_{i=1}^{N} x_i}{\left(1-p\right)^2} < 0$$

verifying that indeed $\widehat{p}_{MLE}$ is the maximizer of the likelihood function. ■

Example 5.6. Maximum Likelihood Estimation of the parameters of the normal distribution. Suppose now that some random sample is drawn from a normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$. The parameter space is $\boldsymbol{\Theta} = \mathbb{R} \times \mathbb{R}_{++}$: while the mean can take any real value, the variance is allowed to take only positive values. The likelihood function equals the usual joint density of a normally distributed sample:

$$\begin{aligned}\mathcal{L}\left(\mu, \sigma^2 \middle| x_1, \ldots, x_N\right) &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}\right)\end{aligned}$$

but in the log-likelihood form, it simplifies as follows.

$$\log \mathcal{L}\left(\mu, \sigma^2 \middle| x_1, \ldots, x_N\right) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\left(\sigma^2\right) - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}$$

The First Order Conditions, *evaluated at the solution*, are:

$$\begin{aligned}\frac{\partial \log \mathcal{L}\left(\widehat{\mu}_{MLE}, \widehat{\sigma}^2_{MLE} \middle| x_1, \ldots, x_N\right)}{\partial \mu} &= \sum_{i=1}^{N} \frac{x_i - \widehat{\mu}_{MLE}}{\widehat{\sigma}^2_{MLE}} = 0 \\ \frac{\partial \log \mathcal{L}\left(\widehat{\mu}_{MLE}, \widehat{\sigma}^2_{MLE} \middle| x_1, \ldots, x_N\right)}{\partial \sigma^2} &= -\frac{N}{2\widehat{\sigma}^2_{MLE}} + \sum_{i=1}^{N} \frac{\left(x_i - \widehat{\mu}_{MLE}\right)^2}{2\widehat{\sigma}^4_{MLE}} = 0\end{aligned}$$

which is a system of two equations in two unknowns. Solving it delivers the paired MLE's for the normal distribution, which fit the parameter space.

$$\begin{aligned}\widehat{\mu}_{MLE} &= \frac{1}{N}\sum_{i=1}^{N} X_i \\ \widehat{\sigma}^2_{MLE} &= \frac{1}{N}\sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2\end{aligned}$$

Note that while the solution looks like the paired Method of Moments estimators for μ and σ^2, this is not generally true. To verify that the likelihood function is indeed maximized, it is necessary to analyze the determinant of the Hessian matrix of the log-likelihood function *evaluated at the solution.* The Hessian matrix in question is the following.

$$\mathbf{H}\left(\mu, \sigma^2 \middle| x_1, \ldots, x_N\right) = \begin{bmatrix} \frac{\partial^2 \log \mathcal{L}\left(\mu, \sigma^2 \middle| x_1, \ldots, x_N\right)}{\partial \mu^2} & \frac{\partial^2 \log \mathcal{L}\left(\mu, \sigma^2 \middle| x_1, \ldots, x_N\right)}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 \log \mathcal{L}\left(\mu, \sigma^2 \middle| x_1, \ldots, x_N\right)}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 \log \mathcal{L}\left(\mu, \sigma^2 \middle| x_1, \ldots, x_N\right)}{\partial (\sigma^2)^2} \end{bmatrix}$$

Note that the two second-order partial derivatives outside the diagonal are symmetric and equal, and:

$$\frac{\partial^2 \log \mathcal{L}\left(\mu, \sigma^2 \middle| x_1, \ldots, x_N\right)}{\partial \mu^2} = -\frac{N}{\sigma^2}$$

$$\frac{\partial^2 \log \mathcal{L}\left(\mu, \sigma^2 \middle| x_1, \ldots, x_N\right)}{\partial \mu \partial \sigma^2} = -\sum_{i=1}^{N} \frac{(x_i - \mu)}{\sigma^4}$$

$$\frac{\partial^2 \log \mathcal{L}\left(\mu, \sigma^2 \middle| x_1, \ldots, x_N\right)}{\partial \left(\sigma^2\right)^2} = \frac{N}{2\sigma^4} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^6}$$

however, when evaluated at the solution, the cross-derivatives equals zero, because $\sum_{i=1}^{N} (x_i - \widehat{\mu}_{MLE}) = 0$, while the second derivative of σ^2 simplifies too. In fact, by the second of the two First Order Conditions:

$$\begin{aligned}\frac{\partial^2 \log \mathcal{L}\left(\widehat{\mu}_{MLE}, \widehat{\sigma}^2_{MLE} \middle| x_1, \ldots, x_N\right)}{\partial \left(\sigma^2\right)^2} &= \frac{N}{2\widehat{\sigma}^4_{MLE}} - \sum_{i=1}^{N} \frac{\left(x_i - \widehat{\mu}_{MLE}\right)^2}{\widehat{\sigma}^6_{MLE}} \\ &= \frac{N}{2\widehat{\sigma}^4_{MLE}} - \frac{N}{\widehat{\sigma}^4_{MLE}} \\ &= -\frac{N}{2\widehat{\sigma}^4_{MLE}}\end{aligned}$$

and it follows the Hessian matrix, evaluated at the solution, is:

$$\mathbf{H}\left(\widehat{\mu}_{MLE}, \widehat{\sigma}^2_{MLE} \middle| x_1, \ldots, x_N\right) = \begin{bmatrix} -\dfrac{N}{\widehat{\sigma}^2_{MLE}} & 0 \\ 0 & -\dfrac{N}{2\widehat{\sigma}^4_{MLE}} \end{bmatrix}$$

and its determinant is obviously always positive. Since at least one second order partial derivative (in particular, the second derivative for μ) is always negative, the solution is indeed a maximum. ■

Example 5.7. Maximum Likelihood Estimation of the parameters of the multivariate normal distribution. Move next to a multivariate environment, and consider sampling from a random vector $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The likelihood function is:

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma} \middle| \mathbf{x}_1, \ldots, \mathbf{x}_N\right) &= \prod_{i=1}^{N} \frac{1}{\sqrt{(2\pi)^K |\boldsymbol{\Sigma}|}} \exp\left(-\frac{(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})}{2}\right) \\ &= \frac{1}{\left[(2\pi)^K |\boldsymbol{\Sigma}|\right]^{\frac{N}{2}}} \exp\left(-\sum_{i=1}^{N} \frac{(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})}{2}\right)\end{aligned}$$

which becomes simpler again if transformed into a log-likelihood function.

$$\log \mathcal{L}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \, \mathbf{x}_1, \ldots, \mathbf{x}_N\right) = -\frac{NK}{2} \log\left(2\pi\right) - \\ - \frac{N}{2} \log\left(|\boldsymbol{\Sigma}|\right) - \sum_{i=1}^{N} \frac{\left(\mathbf{x}_i - \boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}\right)}{2}$$

To find the MLE estimator for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ it is easiest to split the problem into simpler bits: the estimation of $\boldsymbol{\mu}$ and that of $\boldsymbol{\Sigma}$ (note that this is not always possible). Here, the First Order Conditions with respect to $\boldsymbol{\mu}$:

$$\frac{\partial \log \mathcal{L}\left(\widehat{\boldsymbol{\mu}}_{MLE}, \boldsymbol{\Sigma} | \, \mathbf{x}_1, \ldots, \mathbf{x}_N\right)}{\partial \boldsymbol{\mu}} = \boldsymbol{\Sigma}^{-1} \sum_{i=1}^{N} \left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_{MLE}\right) = \mathbf{0}$$

constitute a system of K equations in K unknowns $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ whose solution does not depend on $\boldsymbol{\Sigma}$![1] It follows that the MLE estimator of the location parameters is, again, the vector of sample means.

$$\widehat{\boldsymbol{\mu}}_{MLE} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i = \bar{\boldsymbol{x}}$$

To obtain the maximum for $\boldsymbol{\Sigma}$ it is easiest to differentiate the log-likelihood function with respect to its *inverse* $\boldsymbol{\Sigma}^{-1}$; this must return the same solution. Differentiating a scalar with respect to a matrix returns yet another matrix; here this operation gives the following $K \times K$ matrix:[2]

$$\frac{\partial \log \mathcal{L}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \, \mathbf{x}_1, \ldots, \mathbf{x}_N\right)}{\partial \boldsymbol{\Sigma}^{-1}} = \frac{N}{2} \boldsymbol{\Sigma} - \sum_{i=1}^{N} \frac{\left(\mathbf{x}_i - \boldsymbol{\mu}\right) \left(\mathbf{x}_i - \boldsymbol{\mu}\right)^{\mathrm{T}}}{2}$$

[1] This is so because $\boldsymbol{\Sigma}$ is positive semi-definite, a property that extends to its inverse. Thus, those First Order Conditions can only be equal to zero if $\sum_{i=1}^{N} \left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_{MLE}\right) = \mathbf{0}$.

[2] To get into the algebraic details, observe that:

$$\frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}} \log\left(\left|\boldsymbol{\Sigma}^{-1}\right|\right) = \boldsymbol{\Sigma}$$

and that the derivative of the summation component is as follows.

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}} \sum_{i=1}^{N} \left(\mathbf{x}_i - \boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}\right) &= \frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}} \mathrm{tr}\left[\sum_{i=1}^{N} \left(\mathbf{x}_i - \boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}\right)\right] \\ &= \frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}} \mathrm{tr}\left[\sum_{i=1}^{N} \left(\mathbf{x}_i - \boldsymbol{\mu}\right) \left(\mathbf{x}_i - \boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\right] \\ &= \sum_{i=1}^{N} \left(\mathbf{x}_i - \boldsymbol{\mu}\right) \left(\mathbf{x}_i - \boldsymbol{\mu}\right)^{\mathrm{T}} \end{aligned}$$

which, if evaluated at the solution where $\boldsymbol{\mu} = \bar{\boldsymbol{x}}$ and set at zero, returns the MLE of the variance-covariance matrix.

$$\widehat{\boldsymbol{\Sigma}}_{MLE} = \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}\right)^{\mathrm{T}}$$

As in the Method of Moments and in the univariate MLE case, this estimator is a rescaled version of the sample variance-covariance, $\widehat{\boldsymbol{\Sigma}}_{MLE} = \frac{N-1}{N}\boldsymbol{S}$. Some tedious analysis, similar to that from the univariate case, would show that the MLE solutions $\widehat{\boldsymbol{\mu}}_{MLE}$ and $\widehat{\boldsymbol{\Sigma}}_{MLE}$ indeed identify a maximum of the (log-)likelihood function. ■

In the last few cases, the Method of Moments and Maximum Likelihood estimators are seen to coincide. This, however, is generally not true, as the following example shows.

Example 5.8. Maximum Likelihood Estimation of the parameters of the Gamma distribution. Consider again random sampling from the Gamma distribution as in Example 5.3. There, the likelihood function is:

$$\begin{aligned}\mathcal{L}\left(\alpha, \beta \middle| x_1, \ldots, x_N\right) &= \prod_{i=1}^{N}\frac{1}{\Gamma\left(\alpha\right)}\beta^{\alpha}x_i^{\alpha-1}\exp\left(-\beta x_i\right) \\ &= \left(\frac{\beta^{\alpha}}{\Gamma\left(\alpha\right)}\right)^{N}\left(\prod_{i=1}^{N}x_i^{\alpha-1}\right)\exp\left(-\beta\sum_{i=1}^{N}x_i\right)\end{aligned}$$

and the log-likelihood function is:

$$\begin{aligned}\log\mathcal{L}\left(\alpha, \beta \middle| x_1, \ldots, x_N\right) = N\alpha\log\left(\beta\right) - N\log\left[\Gamma\left(\alpha\right)\right] + \\ + \left(\alpha - 1\right)\sum_{i=1}^{N}\log\left(x_i\right) - \beta\sum_{i=1}^{N}x_i\end{aligned}$$

with the following First Order Conditions.[3]

$$\begin{aligned}\frac{\partial\log\mathcal{L}\left(\alpha, \beta \middle| x_1, \ldots, x_N\right)}{\partial\alpha} &= N\log\left(\beta\right) - \frac{N}{\Gamma\left(\alpha\right)}\frac{\partial\Gamma\left(\alpha\right)}{\partial\alpha} + \sum_{i=1}^{N}\log\left(x_i\right) \\ \frac{\partial\log\mathcal{L}\left(\alpha, \beta \middle| x_1, \ldots, x_N\right)}{\partial\beta} &= N\frac{\alpha}{\beta} - \sum_{i=1}^{N}x_i\end{aligned}$$

[3]The derivative of the logarithm of the Gamma function is known as the *polygamma function*, and unless the argument (e.g. α here) is an integer, it only admits an integral representation, which makes it difficult to solve the First Order Conditions for α and β.

There is no closed form solution to this problem. Even if the solution must clearly respect the property that:

$$\frac{\widehat{\alpha}_{MLE}}{\widehat{\beta}_{MLE}} = \frac{1}{N}\sum_{i=1}^{N} X_i = \bar{X}$$

as in the Method of Moments case, exact expressions of the estimators for α and β in terms of $(X_1, \ldots, X_2)$ – or of $(x_1, \ldots, x_2)$ – cannot be derived from the First Order Conditions. It is then necessary to employ *numerical methods* on a case-by-case basis in order to identify the estimates. ■

This example showed how difficult it can be to perform Maximum Likelihood Estimation in certain cases – and this is not uncommon! Sometimes, the MLE of interest does not even exist.

Example 5.9. Maximum Likelihood Estimation of the parameter of uniform distributions with fixed lower bound. Consider a random sample drawn from a uniformly distributed random variable $X_i \sim \mathcal{U}(0, \theta)$ with lower bound fixed at zero and *closed support*: $\mathbb{X} = [0, \theta]$. It is easy to see that $\mathbb{E}[X] = \theta/2$ and thus the Method of Moments estimator is:

$$\widehat{\theta}_{MM} = \frac{2}{N}\sum_{i=1}^{N} X_i = 2\bar{X}$$

while the MLE is the sample maximum.

$$\widehat{\theta}_{MLE} = X_{(N)}$$

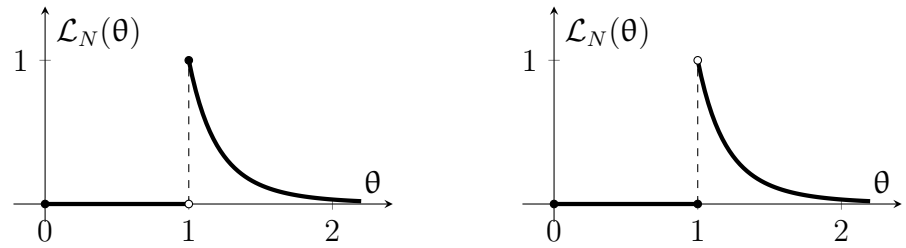
To see this, note that the likelihood function here is:

$$\mathcal{L}(\theta | x_1, \ldots, x_N) = \frac{1}{\theta^N} \cdot \mathbb{1}[0 \leq x_1, \ldots, x_N \leq \theta]$$

and there is no need to guess the log-likelihood function to see that the above is maximized for the *smallest* value of θ such that $\theta \geq \max\{x_1, \ldots, x_N\}$, hence the maximum. Suppose now that the support of X is *open*, at least on the right: $\mathbb{X} = [0, \theta)$. The Method of Moments estimator is unchanged, but the likelihood function now becomes:

$$\mathcal{L}(\theta | x_1, \ldots, x_N) = \frac{1}{\theta^N} \cdot \mathbb{1}[0 \leq x_1, \ldots, x_N < \theta]$$

with only an inequality being changed within the indicator function. Note that it is no longer possible to follow the reasoning above in order to identify a *statistic* that maximizes the likelihood function (to gain intuition, compare the two likelihood functions depicted next in Figure 5.1). In cases like that with open support, one typically says that *the MLE does not exist.* ■



Note: $N = 5$ and $x_{(5)} = 1$ in both cases; $\mathbb{X} = [0, \theta]$ in the left panel and $\mathbb{X} = [0, \theta)$ in the right panel. $\mathcal{L}_N(\theta)$ is shorthand notation for $\mathcal{L}(\theta | x_1, \ldots, x_N)$.

Figure 5.1: Compared likelihood functions for the uniform distribution

It thus might seem that thanks to its simplicity and flexibility, Method of Moments estimation trumps Maximum Likelihood as a more convenient method for constructing estimators, and this goes without mentioning that the latter possibly does not even exist. As anticipated, however, Maximum Likelihood estimators generally have conceptual and statistical advantages; one of these is illustrated next.

Theorem 5.1. Invariance of Maximum Likelihood Estimators. *Call* $\widehat{\boldsymbol{\theta}}_{MLE}$ *the Maximum Likelihood Estimator for some parameter vector* $\boldsymbol{\theta}$*. Let* $\boldsymbol{\varphi} = \boldsymbol{g}(\boldsymbol{\theta})$ *be some* transformation *of parameter vector* $\boldsymbol{\theta}$*. The Maximum Likelihood estimator of* $\boldsymbol{\varphi}$ *is simply the corresponding transformation of the Maximum Likelihood Estimator of* $\boldsymbol{\theta}$*.*

$$\widehat{\boldsymbol{\varphi}}_{MLE} = \boldsymbol{g}\left(\widehat{\boldsymbol{\theta}}_{MLE}\right)$$

Proof. The Maximum Likelihood Estimator of $\boldsymbol{\varphi}$ is obtained as the maximizer of the following, so-called *induced likelihood function.*

$$\mathcal{L}^*(\boldsymbol{\varphi} | \mathbf{x}_1, \ldots, \mathbf{x}_N) = \max_{\{\boldsymbol{\theta}: \boldsymbol{g}(\boldsymbol{\theta}) = \boldsymbol{\varphi}\}} \mathcal{L}(\boldsymbol{\theta} | \mathbf{x}_1, \ldots, \mathbf{x}_N)$$

Call such maximum $\widehat{\boldsymbol{\varphi}}_{MLE}$, and observe that:

$$\begin{aligned}
\mathcal{L}^*(\widehat{\boldsymbol{\varphi}}_{MLE} | \mathbf{x}_1, \ldots, \mathbf{x}_N) &= \max_{\boldsymbol{\varphi}} \max_{\{\boldsymbol{\theta}: \boldsymbol{g}(\boldsymbol{\theta}) = \boldsymbol{\varphi}\}} \mathcal{L}(\boldsymbol{\theta} | \mathbf{x}_1, \ldots, \mathbf{x}_N) \\
&= \max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta} | \mathbf{x}_1, \ldots, \mathbf{x}_N) \\
&= \mathcal{L}\left(\widehat{\boldsymbol{\theta}}_{MLE} \middle| \mathbf{x}_1, \ldots, \mathbf{x}_N\right) \\
&= \max_{\left\{\boldsymbol{\theta}: \boldsymbol{g}(\boldsymbol{\theta}) = \boldsymbol{g}\left(\widehat{\boldsymbol{\theta}}_{MLE}\right)\right\}} \mathcal{L}(\boldsymbol{\theta} | \mathbf{x}_1, \ldots, \mathbf{x}_N) \\
&= \mathcal{L}^*\left(\boldsymbol{g}\left(\widehat{\boldsymbol{\theta}}_{MLE}\right) \middle| \mathbf{x}_1, \ldots, \mathbf{x}_N\right)
\end{aligned}$$

where the first and last equalities follow from the definition of induced likelihood function, the second equality follows from the properties of iterated maximizations, and the remaining ones follow by the definition of MLE. □

Example 5.10. Maximum Likelihood Estimation of the "precision" parameter of the normal distribution. Recall that the normal distribution can be alternatively described in terms of the *precision* parameter $\phi^2 = \sigma^{-2}$, where the density function is expressed as in (2.37). In that case, the (induced) likelihood function would be as follows.

$$\mathcal{L}\left(\mu, \phi^2 \middle| x_1, \ldots, x_N\right) = \left(\frac{\phi^2}{2\pi}\right)^{\frac{N}{2}} \exp\left(-\sum_{i=1}^{N} \frac{\phi^2 (x_i - \mu)^2}{2}\right)$$

By an analysis similar to that of Example 5.6, the First Order Conditions of the log-likelihood function evaluated at the solution:

$$\frac{\partial \log \mathcal{L}\left(\widehat{\mu}_{MLE}, \widehat{\phi}^2_{MLE} \middle| x_1, \ldots, x_N\right)}{\partial \mu} = \widehat{\phi}^2_{MLE} \sum_{i=1}^{N} (x_i - \widehat{\mu}_{MLE}) = 0$$

$$\frac{\partial \log \mathcal{L}\left(\widehat{\mu}_{MLE}, \widehat{\phi}^2_{MLE} \middle| x_1, \ldots, x_N\right)}{\partial \phi^2} = \frac{N}{2\widehat{\phi}^2_{MLE}} - \sum_{i=1}^{N} \frac{(x_i - \widehat{\mu}_{MLE})^2}{2} = 0$$

would reveal that the MLE of μ is still the sample mean, while the MLE of the precision parameter is the following:

$$\widehat{\phi}^2_{MLE} = N \left[\sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2\right]^{-1}$$

which is obviously nothing else but the inverse of $\widehat{\sigma}^2_{MLE}$. ■

Example 5.11. Maximum Likelihood Estimation of an alternative parameter of the Gamma distribution. Also the Gamma distribution admits an alternative parametrization, for $\theta = \beta^{-1}$: the two parameters are distinguished by the names **rate** parameter for β and **scale** parameter for θ (while α is the **shape** parameter). The reparametrized density function is:

$$f_X(x; \alpha, \theta) = \frac{1}{\Gamma(\alpha)\, \theta^{\alpha}} x^{\alpha - 1} \exp\left(-\frac{1}{\theta} x\right) \quad \text{for } x > 0$$

hence in a random sample, the (induced) likelihood function is as follows.

$$\mathcal{L}\left(\alpha, \theta \middle| x_1, \ldots, x_N\right) = \left(\frac{1}{\Gamma(\alpha)\, \theta^{\alpha}}\right)^N \left(\prod_{i=1}^{N} x_i^{\alpha - 1}\right) \exp\left(-\frac{1}{\theta} \sum_{i=1}^{N} x_i\right)$$

Analyzing the First Order Conditions of the log-likelihood function:

$$\frac{\partial \log \mathcal{L}\left(\alpha, \theta | x_1, \ldots, x_N\right)}{\partial \alpha} = -N \log\left(\theta\right) - \frac{N}{\Gamma\left(\alpha\right)} \frac{\partial \Gamma\left(\alpha\right)}{\partial \alpha} + \sum_{i=1}^{N} \log\left(x_i\right)$$

$$\frac{\partial \log \mathcal{L}\left(\alpha, \theta | x_1, \ldots, x_N\right)}{\partial \theta} = -\frac{N\alpha}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{N} x_i$$

once again shows that a closed form solution cannot be identified; however, the solution must be such that $\widehat{\theta}_{MLE} = \widehat{\beta}_{MLE}^{-1}$. ■

The invariance property does not extend to Method of Moments estimators. While this is of little consequence in those cases where the latter coincide the MLE, as in the various estimators of the mean considered in the examples, this raises concerns about the Method of Moments in those cases where the two approaches differ.

5.2 Evaluation of Estimators

Whether obtained through the Method of Moments, Maximum Likelihood or other means, all estimators are *statistics* – functions of the random variables or vectors from which samples are drawn – and thus they are endowed of a sampling distribution. While it is often difficult to derive selected sampling distributions, it is often possible to analyze some of their properties, especially certain *moments* of the estimators, in order to inform the choice between different estimators. In fact, some estimators are better than others (that are meant for the same parameters), having better *statistical properties*: in practical settings, using the sample values of the "better" estimators results in more accurate conjectures about the parameters of interest.

This discussion is begun by introducing an important criterion that is used to compare different estimators.

Definition 5.4. Mean Squared Error (MSE). Consider an estimator $\widehat{\theta}$ for some parameters of interest θ, where both $\widehat{\theta}$ and θ have dimension K. The *mean squared error* is defined as the following quantity:

$$\text{MSE} \equiv \mathbb{E}\left[\left(\widehat{\theta} - \theta\right)^{\text{T}} \left(\widehat{\theta} - \theta\right)\right] = \sum_{k=1}^{K} \mathbb{E}\left[\left(\widehat{\theta}_k - \theta_k\right)^2\right]$$

where $k = 1, \ldots, K$ indexes the parameters and associated estimators listed in θ and $\widehat{\theta}$, respectively.

Thus, for any vector of parameters and associated estimates the MSE is simply the sum of K elements of the form:

$$\text{MSE}_k = \mathbb{E}\left[\left(\widehat{\theta}_k - \theta_k\right)^2\right]$$

where $\widehat{\theta}_k$ is a statistic, θ_k is a parameter, and both are unidimensional. As $\widehat{\theta}_k$ has a sampling distribution, the quantity expressed above is a measure of the average size of *squared* deviations from the parameter of interest that result from the particular estimator $\widehat{\theta}_k$. The use of a squared deviation, as mentioned in Lecture 1 for an analogous context, is intuitive and motivated on the fact that larger deviations are less desirable than smaller deviations, and if the MSE is used to compare estimators, those estimators that produce larger deviations are more penalized by this criterion.

Alternative criteria, such one that adopts *absolute deviations* like:

$$\text{MAE}_k = \mathbb{E}\left[\left|\widehat{\theta}_k - \theta_k\right|\right]$$

are certainly possible (the above is called **Mean Absolute Error** – MAE). Nevertheless, the overwhelming majority of practical applications adopts the MSE; among the various reasons (including analytical convenience) the following property plays a fundamental role.

$$\begin{aligned}
\text{MSE}_k &= \mathbb{E}\left[\left(\widehat{\theta}_k - \mathbb{E}\left[\widehat{\theta}_k\right] + \mathbb{E}\left[\widehat{\theta}_k\right] - \theta_k\right)^2\right] \\
&= \mathbb{E}\left[\left(\widehat{\theta}_k - \mathbb{E}\left[\widehat{\theta}_k\right]\right)^2\right] + \mathbb{E}\left[\left(\mathbb{E}\left[\widehat{\theta}_k\right] - \theta_k\right)^2\right] + \\
&\quad + 2\,\mathbb{E}\left[\left(\widehat{\theta}_k - \mathbb{E}\left[\widehat{\theta}_k\right]\right)\left(\mathbb{E}\left[\widehat{\theta}_k\right] - \theta_k\right)\right] \\
&= \mathbb{V}\text{ar}\left[\widehat{\theta}_k\right] + \left(\mathbb{E}\left[\widehat{\theta}_k\right] - \theta_k\right)^2
\end{aligned}$$

Above, the last element in the second line is easily shown to vanish to zero; this decomposition should reminiscent of the analysis conducted in Lecture 1 about the mean as the "best guess" of a random variable. In words, the MSE relative to a specific estimator $\widehat{\theta}_k$ can be decomposed in two parts: the *variance* of the estimator, and the squared deviation of its *mean* from the parameter of interest. The last concept warrants a definition.

Definition 5.5. Bias and unbiasedness. Consider a unidimensional estimator $\widehat{\theta}$ for some parameter of interest θ. Its *bias* is the quantity:

$$\text{Bias}_{\widehat{\theta}} \equiv \mathbb{E}\left[\widehat{\theta}\right] - \theta$$

and the estimator is *unbiased* if its bias is zero.

Unbiased estimators are certainly appealing, because they can be interpreted as estimators whose "average" value (in the population of all possible samples) equals the parameter estimation. However, an unbiased estimator might produce a MSE which is larger than that of a biased estimator. Thus, a researcher who wants to compare the MSE of different estimators, so to choose the one with the smallest MSE, must be aware that a **bias-variance trade-off** might arise: choosing an estimator with smaller bias might imply accepting a higher variance than the alternative. If a given estimator has a smaller variance relative to another estimator it is compared against, it is said that the former is **more efficient** than the latter, and vice versa.

Example 5.12. Estimation of the parameters of the normal distribution: the bias-variance trade-off. Consider a random sample drawn from some normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, and the following two alternative estimators of the parameters:

$$\begin{pmatrix} \widehat{\mu}_1 \\ \widehat{\sigma}_1^2 \end{pmatrix} = \begin{pmatrix} \bar{X} \\ S^2 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \widehat{\mu}_2 \\ \widehat{\sigma}_2^2 \end{pmatrix} = \begin{pmatrix} \bar{X} \\ \frac{N-1}{N} S^2 \end{pmatrix}$$

where $\widehat{\sigma}_2^2 = \frac{N-1}{N} S^2$ is motivated by either the Method of Moments or Maximum Likelihood. The estimators of the location parameter are identical, so they do not contribute to the difference between the MSE's of the two alternatives. It was shown in Theorem 4.3 that $\mathbb{E}\left[S^2 - \sigma^2\right] = 0$, implying that $\widehat{\sigma}_1^2$ is an *unbiased* estimator of σ^2. Consequently, $\mathbb{E}\left[\frac{N-1}{N} S^2 - \sigma^2\right] = -\frac{1}{N}\sigma^2$: $\widehat{\sigma}_2^2$ is a *biased* estimator of σ^2. Notably, the MSE associated with the second setup is lower! In fact, $\widehat{\sigma}_2^2$ is *more efficient* than $\widehat{\sigma}_1^2$:

$$\begin{aligned} \mathbb{V}\mathrm{ar}\left[\widehat{\sigma}_1^2\right] - \mathbb{V}\mathrm{ar}\left[\widehat{\sigma}_2^2\right] &= \mathbb{V}\mathrm{ar}\left[S^2\right] - \left(\frac{N-1}{N}\right)^2 \mathbb{V}\mathrm{ar}\left[S^2\right] \\ &= \frac{2N-1}{N^2}\, \mathbb{V}\mathrm{ar}\left[S^2\right] \\ &= \frac{(2N-1)\,\sigma^4}{N^2\,(N-1)^2}\, \mathbb{V}\mathrm{ar}\left[(N-1)\,\frac{S^2}{\sigma^2}\right] \\ &= \frac{2\,(2N-1)\,\sigma^4}{N^2\,(N-1)} \end{aligned}$$

where the last line follows from the fact that if $W \sim \chi_\kappa^2$, then $\mathbb{V}\mathrm{ar}[W] = 2\kappa$ (here, $\kappa = N-1$). Clearly:

$$\mathbb{V}\mathrm{ar}\left[\widehat{\sigma}_1^2\right] - \mathbb{V}\mathrm{ar}\left[\widehat{\sigma}_2^2\right] = \frac{2\,(2N-1)\,\sigma^4}{N^2\,(N-1)} > \frac{\sigma^4}{N^2} = \left\{\mathbb{E}\left[\widehat{\sigma}_2^2 - \sigma^2\right]\right\}^2$$

that is, the difference between the variances of the two alternative estimators of the variance is larger than the square of the bias of $\widehat{\sigma}_2^1$. ■

This example clarifies why the unbiasedness property does not guarantee that an estimator performs better than others in practical situations. When restricting their attention to unbiased estimators, researchers should at least ensure that these are also those with the smallest possible variance. These estimators have a proper name in the theory of statistical inference.

Definition 5.6. Best unbiased estimators. Consider the set of unbiased estimators $\widehat{\theta}$ of a certain parameter θ:

$$\mathbb{C}_\theta = \left\{ \widehat{\theta} : \mathbb{E}\left[\widehat{\theta}\right] = \theta \right\}$$

An estimator $\widehat{\theta}^*$ is called the *best unbiased estimator*, or the *uniform minimum variance unbiased estimator* of θ, if the following holds.

$$\mathbb{V}\text{ar}\left[\widehat{\theta}\right] - \mathbb{V}\text{ar}\left[\widehat{\theta}^*\right] \geq 0 \quad \text{for all } \widehat{\theta} \in \mathbb{C}_\theta \tag{5.20}$$

In a multidimensional environment, if $\widehat{\boldsymbol{\theta}}$ is a vector of estimators that are all unbiased for a vector of parameters $\boldsymbol{\theta}$, this definition is recast in terms of a vector $\widehat{\boldsymbol{\theta}}^*$ of best unbiased estimators such that:

$$\mathbb{V}\text{ar}\left[\widehat{\boldsymbol{\theta}}\right] - \mathbb{V}\text{ar}\left[\widehat{\boldsymbol{\theta}}^*\right] \geq \mathbf{0} \quad \text{for all } \widehat{\boldsymbol{\theta}} \in \mathbb{C}_{\theta_1} \times \ldots \times \mathbb{C}_{\theta_K} \tag{5.21}$$

where the inequality is interpreted in the sense that the matrix on the left-hand side is positive semi-definite, and $\mathbb{C}_{\theta_k}$ is the set of unbiased estimators of θ_k for $k = 1, \ldots, K$.

A property of best unbiased estimators is that they are unique.

Theorem 5.2. Uniqueness of best unbiased estimators. *Let $\widehat{\theta}^*$ be a best unbiased estimator for some parameter θ. In this setting $\widehat{\theta}^*$ is unique, in the sense that (5.20) holds sharply (without equality).*

Proof. Suppose that there is another estimator $\widehat{\theta}^{**}$ that is also a best unbiased estimator, in the sense that it has the same expectation and variance as $\widehat{\theta}^*$. Define the estimator:

$$\widehat{\theta}' \equiv \frac{1}{2}\widehat{\theta}^* + \frac{1}{2}\widehat{\theta}^{**}$$

it is clear that $\mathbb{E}\left[\widehat{\theta}'\right] = \theta$. As per the variance, it must be that:

$$\begin{aligned}
\mathbb{V}\text{ar}\left[\widehat{\theta}'\right] &= \frac{1}{4}\mathbb{V}\text{ar}\left[\widehat{\theta}^*\right] + \frac{1}{4}\mathbb{V}\text{ar}\left[\widehat{\theta}^{**}\right] + \frac{1}{2}\mathbb{C}\text{ov}\left[\widehat{\theta}^*, \widehat{\theta}^{**}\right] \\
&\leq \frac{1}{4}\mathbb{V}\text{ar}\left[\widehat{\theta}^*\right] + \frac{1}{4}\mathbb{V}\text{ar}\left[\widehat{\theta}^{**}\right] + \frac{1}{2}\left\{\mathbb{V}\text{ar}\left[\widehat{\theta}^*\right]\mathbb{V}\text{ar}\left[\widehat{\theta}^{**}\right]\right\}^{\frac{1}{2}} \\
&= \mathbb{V}\text{ar}\left[\widehat{\theta}^*\right]
\end{aligned}$$

where the inequality follows from the same argument as in Theorem 3.4, and the last line is due to the fact that $\widehat{\theta}^*$ and $\widehat{\theta}^{**}$ have the same variance. Note that the inequality must be replaced by an equality to avoid a contradiction! If the inequality were sharp, then $\widehat{\theta}^*$ would not be a best unbiased estimator, as $\widehat{\theta}'$ would improve it. To have an equality, it must be – again by Theorem 3.4 – that $\widehat{\theta}^{**}$ is a linear transformation of $\widehat{\theta}^*$, that is $\widehat{\theta}^{**} = a + b\widehat{\theta}^*$. But in this case it must also be that $a = 0$, or else $\widehat{\theta}^{**}$ would be biased, and $b = 1$, since the following chain of equalities must also hold.

$$\mathbb{V}\text{ar}\left[\widehat{\theta}^*\right] = \mathbb{C}\text{ov}\left[\widehat{\theta}^*, \widehat{\theta}^{**}\right] = \mathbb{C}\text{ov}\left[\widehat{\theta}^*, b\widehat{\theta}^*\right] = b\,\mathbb{V}\text{ar}\left[\widehat{\theta}^*\right]$$

Thus, $\widehat{\theta}^*$, $\widehat{\theta}^{**}$ and $\widehat{\theta}'$ are all identical estimators, that is, $\widehat{\theta}^*$ is the only best unbiased estimator. □

The search for unbiased estimators with good properties is facilitated by the following result, which is stated here in its multivariate version.

Theorem 5.3. The Rao-Blackwell Theorem. *Consider an environment where* $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ *is a sample drawn from some list of random vectors,* $\boldsymbol{\theta}$ *is some parameter vector of interest,* $\widehat{\boldsymbol{\theta}}$ *is* any *vector of unbiased estimators of* $\boldsymbol{\theta}$*, and* $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ *is a vector of statistics that are all simultaneously sufficient for* $\boldsymbol{\theta}$*. Define the following statistic as a conditional expectation function.*

$$\widehat{\boldsymbol{\theta}}^* \equiv \mathbb{E}\left[\widehat{\boldsymbol{\theta}} \middle| \boldsymbol{t}\right]$$

The statistic $\widehat{\boldsymbol{\theta}}^*$ *is a uniformly better unbiased estimator of* $\boldsymbol{\theta}$*, that is, it is an unbiased estimator with lower variance than* $\widehat{\boldsymbol{\theta}}$*.*

Proof. The Law of Iterated Expectations:

$$\boldsymbol{\theta} = \mathbb{E}\left[\widehat{\boldsymbol{\theta}}\right] = \mathbb{E}_{\boldsymbol{t}}\left[\mathbb{E}\left[\widehat{\boldsymbol{\theta}} \middle| \boldsymbol{t}\right]\right] = \mathbb{E}\left[\widehat{\boldsymbol{\theta}}^*\right]$$

along with the Law of Total Variance:

$$\begin{aligned}\mathbb{V}\text{ar}\left[\widehat{\boldsymbol{\theta}}\right] &= \mathbb{V}\text{ar}_{\boldsymbol{t}}\left[\mathbb{E}\left[\widehat{\boldsymbol{\theta}} \middle| \boldsymbol{t}\right]\right] + \mathbb{E}_{\boldsymbol{t}}\left[\mathbb{V}\text{ar}\left[\widehat{\boldsymbol{\theta}} \middle| \boldsymbol{t}\right]\right] \\ &= \mathbb{V}\text{ar}\left[\widehat{\boldsymbol{\theta}}^*\right] + \mathbb{E}_{\boldsymbol{t}}\left[\mathbb{V}\text{ar}\left[\widehat{\boldsymbol{\theta}} \middle| \boldsymbol{t}\right]\right] \\ &\geq \mathbb{V}\text{ar}\left[\widehat{\boldsymbol{\theta}}^*\right]\end{aligned}$$

simultaneously show that *if* $\widehat{\boldsymbol{\theta}}^*$ is an estimator of $\boldsymbol{\theta}$, it is unbiased and it also has a lower variance than $\widehat{\boldsymbol{\theta}}$. The definition of sufficiency and that of $\widehat{\boldsymbol{\theta}}^*$, however, jointly imply that the latter is a legitimate estimator of $\boldsymbol{\theta}$, as its joint distribution by construction does not depend on $\boldsymbol{\theta}$. □

This result allows to "improve" already known unbiased estimators by constructing an appropriate CEF (conditional expectation function) with sufficient statistics of the parameters of interest as arguments. This is generally not easy. Another use of this result is to show that certain unbiased estimators *cannot be further improved*. This is the case of well-known estimators that are already expressed as simple functions of sufficient statistics.

Example 5.13. Rao-Blackwell applied to the normal distribution. Consider again the pair of unbiased estimators for the parameters (μ, σ^2) of the normal distribution: the two statistics $\left(\bar{X}, S^2\right)$. It turns out that these two statistics are also *sufficient* (Example 4.4), hence they can be expressed as trivial conditional expectation functions of themselves. Therefore, one cannot find better *unbiased* estimators that are based on the same sufficient statistics (but estimators with a smaller MSE are possible, as shown). ■

According to the MSE criterion, evaluating the properties of estimators, whether they are unbiased or not, requires some understanding about how their variances compare with those of competing estimators. A fundamental result in Statistics, known as the **Cramér-Rao Inequality**, characterizes a **lower bound** on the variance that estimators can achieve. Therefore, the closer an estimator is to that value, called the *Cramér-Rao lower bound*, the more confident a researcher should be about using that estimator. In what follows, the Theorem is stated and proved in both its general version, and in the case restricted ti random samples. To facilitate understanding, the two statements are expressed both in the univariate and multivariate cases; the proofs begin by demonstrating the results in the univariate cases, and then discuss how they are modified or extended in the multivariate cases.

Theorem 5.4. Cramér-Rao Inequality (general) – univariate case. *Consider a sample drawn from a list of random variables $(X_1, \ldots, X_N)$ with a joint mass or density function written as $f(x_1, \ldots, x_N; \theta)$ with shorthand notation. Also consider a parameter of interest θ, as well as some estimator $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_N)$ for θ, such that its variance is finite and additionally – in the continuous case only – that the differentiation operation taken with respect to θ can pass through the expectation operator as shown below.*

$$\frac{\partial}{\partial \theta} \mathbb{E}\left[\widehat{\theta}\right] = \int_{\mathbb{X}_1} \cdots \int_{\mathbb{X}_N} \frac{\partial}{\partial \theta} \widehat{\theta}(x_1, \ldots, x_N) \cdot f(x_1, \ldots, x_N; \theta)\, dx_1 \ldots dx_N$$

In this environment, the variance of $\widehat{\theta}$ must satisfy the following inequality.

$$\mathbb{V}\text{ar}\left[\widehat{\theta}\right] \geq \frac{\left(\frac{\partial}{\partial \theta} \mathbb{E}\left[\widehat{\theta}\right]\right)^2}{\mathbb{E}\left[\left(\frac{\partial}{\partial \theta} \log f(X_1, \ldots, X_N; \theta)\right)^2\right]}$$

Proof. If one defines the following transformed random variables:

$$U = \widehat{\theta}\left(X_1, \ldots, X_N\right)$$
$$V = \frac{\partial}{\partial \theta} \log f\left(X_1, \ldots, X_N; \theta\right)$$

the result follows through as a simple implication of Theorem 3.4.

$$\mathbb{V}\mathrm{ar}\left[U\right] \geq \frac{\left[\mathbb{C}\mathrm{ov}\left[U, V\right]\right]^2}{\mathbb{V}\mathrm{ar}\left[V\right]}$$

Note that if $\mathbb{E}\left[V\right] = 0$ the above is recast as:

$$\mathbb{V}\mathrm{ar}\left[U\right] \geq \frac{\left[\mathbb{E}\left[UV\right]\right]^2}{\mathbb{E}\left[V^2\right]}$$

and in fact, in the continuous case (the discrete case is analogous) it is:

$$\begin{aligned}
\mathbb{E}\left[V\right] &= \mathbb{E}\left[\frac{\partial}{\partial \theta} \log f\left(X_1, \ldots, X_N; \theta\right)\right] \\
&= \mathbb{E}\left[\frac{1}{f\left(X_1, \ldots, X_N; \theta\right)} \frac{\partial}{\partial \theta} f\left(X_1, \ldots, X_N; \theta\right)\right] \\
&= \int_{\mathbb{X}_1} \ldots \int_{\mathbb{X}_N} \frac{\partial}{\partial \theta} f\left(x_1, \ldots, x_N; \theta\right) dx_1 \ldots dx_N \\
&= \frac{\partial}{\partial \theta} \int_{\mathbb{X}_1} \ldots \int_{\mathbb{X}_N} f\left(x_1, \ldots, x_N; \theta\right) dx_1 \ldots dx_N \\
&= \frac{\partial}{\partial \theta} \cdot 1 \\
&= 0
\end{aligned}$$

where the second line applies the chain rule while the fourth line is based on the hypotheses about differentiation and expectation. Similarly:

$$\begin{aligned}
\mathbb{E}\left[UV\right] &= \mathbb{E}\left[\widehat{\theta}\left(X_1, \ldots, X_N\right) \cdot \frac{\partial}{\partial \theta} \log f\left(X_1, \ldots, X_N; \theta\right)\right] \\
&= \mathbb{E}\left[\frac{\widehat{\theta}\left(X_1, \ldots, X_N\right)}{f\left(X_1, \ldots, X_N; \theta\right)} \frac{\partial}{\partial \theta} f\left(X_1, \ldots, X_N; \theta\right)\right] \\
&= \int_{\mathbb{X}_1} \ldots \int_{\mathbb{X}_N} \widehat{\theta}\left(x_1, \ldots, x_N\right) \cdot \frac{\partial}{\partial \theta} f\left(x_1, \ldots, x_N; \theta\right) dx_1 \ldots dx_N \\
&= \frac{\partial}{\partial \theta} \int_{\mathbb{X}_1} \ldots \int_{\mathbb{X}_N} \widehat{\theta}\left(x_1, \ldots, x_N\right) \cdot f\left(x_1, \ldots, x_N; \theta\right) dx_1 \ldots dx_N \\
&= \frac{\partial}{\partial \theta}\, \mathbb{E}\left[\widehat{\theta}\right]
\end{aligned}$$

and collecting terms, the postulated result is obtained. □

Multivariate case. *Consider a sample drawn from a list of random vectors* $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ *with joint mass or density function written as* $f(\mathbf{x}_1, \ldots, \mathbf{x}_N; \theta)$ *with shorthand notation. Also consider a vector of parameters of interest* $\boldsymbol{\theta}$ *with length* K*, as well as some estimator* $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ *for* $\boldsymbol{\theta}$*, such that its variance is finite and additionally – in the continuous case only – that the differentiation operation taken with respect to* $\boldsymbol{\theta}$ *can pass through the expectation operator as shown below.*

$$\frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \mathbb{E}\left[\widehat{\boldsymbol{\theta}}\right] = \int_{\mathbb{X}_1} \cdots \int_{\mathbb{X}_N} \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \widehat{\boldsymbol{\theta}}(\mathbf{x}_1, \ldots, \mathbf{x}_N) \cdot f(\mathbf{x}_1, \ldots, \mathbf{x}_N; \boldsymbol{\theta}) \, \mathrm{d}\mathbf{x}_1 \ldots \mathrm{d}\mathbf{x}_N$$

In this environment, the variance of $\widehat{\boldsymbol{\theta}}$ *must satisfy the following inequality:*

$$\mathbb{V}\mathrm{ar}\left[\widehat{\boldsymbol{\theta}}\right] - \left[\frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \mathbb{E}\left[\widehat{\boldsymbol{\theta}}\right]\right] \left[\boldsymbol{I}_N(\boldsymbol{\theta})\right]^{-1} \left[\frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \mathbb{E}\left[\widehat{\boldsymbol{\theta}}\right]\right]^{\mathrm{T}} \geq \mathbf{0}$$

which is to be interpreted in the sense that the $K \times K$ *matrix on the left hand side is positive semi-definite, and where* $\boldsymbol{I}_N(\boldsymbol{\theta})$ *is as follows.*

$$\boldsymbol{I}_N(\boldsymbol{\theta}) \equiv \mathbb{E}\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; \boldsymbol{\theta})\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; \boldsymbol{\theta})\right)^{\mathrm{T}}\right]$$

Proof. In analogy with the univariate case, define the random vectors:

$$\begin{aligned} \boldsymbol{u} &= \widehat{\boldsymbol{\theta}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \\ \boldsymbol{v} &= \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; \boldsymbol{\theta}) \end{aligned}$$

that are related as follows, by the properties of multivariate moments.

$$\mathbb{V}\mathrm{ar}[\boldsymbol{u}] - [\mathbb{C}\mathrm{ov}[\boldsymbol{u}, \boldsymbol{v}]] [\mathbb{V}\mathrm{ar}[\boldsymbol{v}]]^{-1} [\mathbb{C}\mathrm{ov}[\boldsymbol{u}, \boldsymbol{v}]]^{\mathrm{T}} \geq \mathbf{0}$$

If $\mathbb{E}[\boldsymbol{v}] = \mathbf{0}$, the above simplifies as:

$$\mathbb{V}\mathrm{ar}[\boldsymbol{u}] - \left[\mathbb{E}\left[\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}\right]\right] \left[\mathbb{E}\left[\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}\right]\right]^{-1} \left[\mathbb{E}\left[\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}\right]\right]^{\mathrm{T}} \geq \mathbf{0}$$

where $\mathbb{E}\left[\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}\right] = \boldsymbol{I}_N(\boldsymbol{\theta})$. Consequently, the stated result follows through if, in addition, the following holds too.

$$\mathbb{E}\left[\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}\right] = \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \mathbb{E}\left[\widehat{\boldsymbol{\theta}}\right]$$

Note that if the above relationship is proved, then $\mathbb{E}[\boldsymbol{v}] = \mathbf{0}$ follows easily by replacing $\boldsymbol{u}$ with the unit vector $\iota_K = (1, \ldots, 1)^{\mathrm{T}}$ having the same length

K as $\boldsymbol{\theta}$. To avoid repeating similar arguments as it was done (for illustrative purposes) in the univariate case, only the more complex case is developed.

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}\right] &= \mathbb{E}\left[\widehat{\boldsymbol{\theta}}\left(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\right)\cdot\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\log f\left(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N;\boldsymbol{\theta}\right)\right] \\
&= \mathbb{E}\left[\frac{\widehat{\boldsymbol{\theta}}\left(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\right)}{f\left(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N;\boldsymbol{\theta}\right)}\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}f\left(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N;\boldsymbol{\theta}\right)\right] \\
&= \int_{\mathbb{X}_1}\cdots\int_{\mathbb{X}_N}\widehat{\boldsymbol{\theta}}\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right)\cdot\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}f\left(\mathbf{x}_1,\ldots,\mathbf{x}_N;\boldsymbol{\theta}\right)\mathrm{d}\mathbf{x}_1\ldots\mathrm{d}\mathbf{x}_N \\
&= \frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\int_{\mathbb{X}_1}\cdots\int_{\mathbb{X}_N}\widehat{\boldsymbol{\theta}}\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right)\cdot f\left(\mathbf{x}_1,\ldots,\mathbf{x}_N;\boldsymbol{\theta}\right)\mathrm{d}\mathbf{x}_1\ldots\mathrm{d}\mathbf{x}_N \\
&= \frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\,\mathbb{E}\left[\widehat{\boldsymbol{\theta}}\right]
\end{aligned}$$

As a consequence, $\mathbb{E}\left[\boldsymbol{v}\right]=\mathbf{0}$ as well as the main result also follow. □

Once stated and proved, the expressions of the Cramér-Rao inequalities surely look formidable, and it is worthwhile to analyze them carefully. In the univariate case, the main determinant of the lower bound is the denominator of the ratio, which is called **Fisher information number**:

$$\mathcal{I}_N\left(\theta\right)=\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f\left(X_1,\ldots,X_N;\theta\right)\right)^2\right]$$

note that this number is a different function of the parameter θ for each possible distribution that generates the data. The name "information" is based on the interpretation of this number as the overall "amount of knowledge" that a certain distribution $f\left(X_1,\ldots,X_N;\theta\right)$ can provide about a parameter of interest θ (the higher the number, the lower the bound on the variance of θ). Its multivariate analogue is the matrix $\boldsymbol{I}_N\left(\boldsymbol{\theta}\right)$, which is unsurprisingly called **Fisher information matrix**.

As useful as this intuition can be, the expressions that characterize the bound appear still difficult to operationalize in practice. However, they can be simplified in a number of different ways.

1. If the estimators are *unbiased*, the two terms $\frac{\partial}{\partial\theta}\,\mathbb{E}\left[\widehat{\theta}\right]$ and $\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\,\mathbb{E}\left[\widehat{\boldsymbol{\theta}}\right]$ clearly reduce to 1 and the identity matrix $\mathbf{I}$ respectively.

2. If the sample is *random*, the information number and the information matrix can be simplified as expressed in the theorem stated next.

3. Additional simplifications are possible under some fairly general conditions that are detailed later.

Theorem 5.5. Cramér-Rao Inequality (i.i.d.) – univariate case. *In the (univariate) setup of Theorem 5.4, if the sample is random the inequality can be expressed as follows:*

$$\mathbb{V}\mathrm{ar}\left[\widehat{\theta}\right] \geq \frac{\left(\frac{\partial}{\partial\theta}\,\mathbb{E}\left[\widehat{\theta}\right]\right)^2}{N \cdot \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f_X\left(X;\theta\right)\right)^2\right]}$$

where $f_X\left(x;\theta\right)$ is the mass or density function that generates the sample.

Proof. Observe that:

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f\left(X_1,\ldots,X_N;\theta\right)\right)^2\right] &= \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log\prod_{i=1}^{N} f_X\left(X_i;\theta\right)\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{N}\frac{\partial}{\partial\theta}\log f_X\left(X_i;\theta\right)\right)^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{N}\left(\frac{\partial}{\partial\theta}\log f_X\left(X_i;\theta\right)\right)^2\right] \\
&= \sum_{i=1}^{N}\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f_X\left(X_i;\theta\right)\right)^2\right] \\
&= N \cdot \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f_X\left(X;\theta\right)\right)^2\right]
\end{aligned}$$

where the first line follows from random sampling, the second line is a simple manipulation, the third and fourth lines are based on the linear properties of expectations and independence, as terms of the following form for $i \neq j$:

$$\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f_X\left(X_i;\theta\right)\right)\left(\frac{\partial}{\partial\theta}\log f_X\left(X_j;\theta\right)\right)\right] = 0$$

must be equal to the product of the respective means and therefore to zero, while the fifth line follows from identically distributed observations. □

Multivariate case. *In the (multivariate) version of the setup of Theorem 5.4, if the sample is random the inequality is based on the following version of the information matrix:*

$$\boldsymbol{I}_N\left(\boldsymbol{\theta}\right) = N \cdot \mathbb{E}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right)^{\mathrm{T}}\right]$$

where $f_{\boldsymbol{x}}\left(\boldsymbol{x};\boldsymbol{\theta}\right)$ is the mass or density function that generates the sample.

Proof. The proof is all but an extension of the univariate case. The information matrix is developed as:

$$\begin{aligned}
\boldsymbol{I}_N(\boldsymbol{\theta}) &= \mathbb{E}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N;\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N;\boldsymbol{\theta})\right)^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log\prod_{i=1}^{N} f_{\boldsymbol{x}}(\boldsymbol{x}_i;\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log\prod_{i=1}^{N} f_{\boldsymbol{x}}(\boldsymbol{x}_i;\boldsymbol{\theta})\right)^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}(\boldsymbol{x}_i;\boldsymbol{\theta})\right)\left(\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}(\boldsymbol{x}_i;\boldsymbol{\theta})\right)^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{N}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}(\boldsymbol{x}_i;\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}(\boldsymbol{x}_i;\boldsymbol{\theta})\right)^{\mathrm{T}}\right] \\
&= \sum_{i=1}^{N}\mathbb{E}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}(\boldsymbol{x}_i;\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}(\boldsymbol{x}_i;\boldsymbol{\theta})\right)^{\mathrm{T}}\right] \\
&= N\cdot\mathbb{E}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}(\boldsymbol{x};\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}(\boldsymbol{x};\boldsymbol{\theta})\right)^{\mathrm{T}}\right]
\end{aligned}$$

where the crucial step is between the third and the fourth line, as the terms of the following form, for $i \neq j$:

$$\mathbb{E}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}(\boldsymbol{x}_i;\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}(\boldsymbol{x}_j;\boldsymbol{\theta})\right)^{\mathrm{T}}\right] = \mathbf{0}$$

disappear due to independence; the other steps are simple manipulations or other implications of random sampling. □

The mentioned additional simplifications are possible if the differentiation operation with respect to the parameters of interest can pass through the expectation operator *twice* (this is generally the case for a wide number of distributions, including all those in the exponential macro-family). This implies that in the univariate case, it is:

$$\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f_X(X;\theta)\right)^2\right] = -\,\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f_X(X;\theta)\right]$$

while in the multivariate case the following two $K\times K$ matrices are equal: a result known as the **information matrix equality**.

$$\mathbb{E}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}(\boldsymbol{x};\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}(\boldsymbol{x};\boldsymbol{\theta})\right)^{\mathrm{T}}\right] = -\,\mathbb{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\log f_{\boldsymbol{x}}(\boldsymbol{x};\boldsymbol{\theta})\right]$$

These results are essentially mathematical properties of the logarithm of density and mass functions; these properties can facilitate the calculation of the Cramér-Rao bound. Only the multivariate continuous case is proven here (the univariate and the discrete cases are respectively a particular and an analogous version of it). Observe that:

$$\begin{aligned}
\mathbf{0} &= \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \frac{\partial}{\partial \boldsymbol{\theta}} 1 \\
&= \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \frac{\partial}{\partial \boldsymbol{\theta}} \int_{\mathbb{X}} f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right) \mathrm{d}\mathbf{x} \\
&= \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \int_{\mathbb{X}} \frac{\partial f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \mathrm{d}\mathbf{x} \\
&= \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \int_{\mathbb{X}} \frac{\partial \log f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right) \mathrm{d}\mathbf{x} \\
&= \int_{\mathbb{X}} \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \left[\frac{\partial \log f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right)\right] \mathrm{d}\mathbf{x} \\
&= \int_{\mathbb{X}} \left[\frac{\partial^2 \log f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right) + \frac{\partial \log f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \frac{\partial f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right] \mathrm{d}\mathbf{x} \\
&= \int_{\mathbb{X}} \frac{\partial^2 \log f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right) \mathrm{d}\mathbf{x} + \\
&\quad + \int_{\mathbb{X}} \frac{\partial \log f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \frac{\partial \log f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}^{\mathrm{T}}} f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right) \mathrm{d}\mathbf{x}
\end{aligned}$$

where the first line is almost a tautology, the second line follows from the definition of joint density function, the third and fourth lines are just manipulations, the fifth line takes advantage of the fact that the differentiation operator can pass through the integral twice, the sixth line applies the chain rule, and finally the seventh and last line applies one last manipulation that makes the two sides of the information matrix equality distinct and visible. As all the lines equal a $K \times K$ matrix of zeros, the result must hold.

Collecting all these results together, if all the simplifications described apply the Cramér-Rao Inequality can be written in the univariate case as:

$$\mathbb{V}\mathrm{ar}\left[\widehat{\theta}\right] \geq -\frac{1}{N}\left\{\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \log f_X\left(X; \theta\right)\right]\right\}^{-1} \tag{5.22}$$

and in the multivariate case as follows.

$$\mathbb{V}\mathrm{ar}\left[\widehat{\boldsymbol{\theta}}\right] + \frac{1}{N}\left\{\mathbb{E}\left[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \log f_{\boldsymbol{x}}\left(\boldsymbol{x}; \boldsymbol{\theta}\right)\right]\right\}^{-1} \geq \mathbf{0} \tag{5.23}$$

Some examples help illustrate the usefulness of these conclusions.

Example 5.14. Comparison of estimators for the parameter of the Poisson distribution. Recall that if $X \sim \text{Pois}(\lambda)$, then $\mathbb{E}[X] = \lambda$ and $\mathbb{V}\text{ar}[X] = \lambda$. This fact straightforwardly suggests two *unbiased* estimators for the Poisson parameter λ: the sample mean $\bar{X}$ and the sample variance S^2. A natural subsequent question is: which of the two estimators is better according to the MSE criterion, that is, which of the two has the smallest *sampling variance*? To answer this question, one could proceed by calculating the variance associated with either statistic. In the case of the sample mean this is simple: $\mathbb{V}\text{ar}\left[\bar{X}\right] = \lambda/N$ by Theorem 4.3. However, calculating the variance of the sample variance is not as straightforward.

If one is working with a random sample, however, a shortcut is possible: the question can be answered in favor of the sample mean $\bar{X}$ by calculating the Fisher information number. Recall that $\mathbb{E}[X^2] = \lambda + \lambda^2$ and note that:

$$\begin{aligned}
\mathcal{I}_N(\lambda) &= N \cdot \mathbb{E}\left[\left(\frac{\partial}{\partial\lambda}\log\frac{\exp(-\lambda)\cdot\lambda^X}{X!}\right)^2\right] \\
&= N \cdot \mathbb{E}\left[\left(\frac{\partial}{\partial\lambda}\left(-\lambda + X\log(\lambda) - \log(X!)\right)\right)^2\right] \\
&= N \cdot \mathbb{E}\left[\left(-1 + \frac{X}{\lambda}\right)^2\right] \\
&= N\left(1 - \frac{2}{\lambda}\cdot\mathbb{E}[X] + \frac{1}{\lambda^2}\cdot\mathbb{E}\left[X^2\right]\right) \\
&= N\left(\frac{\lambda+\lambda^2}{\lambda^2} - 1\right) \\
&= \frac{N}{\lambda}
\end{aligned}$$

the Cramér-Rao bound is $\mathbb{V}\text{ar}\left[\widehat{\lambda}\right] \geq \lambda/N$ and is *attained* by $\widehat{\lambda} = \bar{X}$. Also:

$$\begin{aligned}
\mathcal{I}_N(\lambda) &= -N \cdot \mathbb{E}\left[\frac{\partial^2}{\partial\lambda^2}\log\frac{\exp(-\lambda)\cdot\lambda^X}{X!}\right] \\
&= -N \cdot \mathbb{E}\left[\frac{\partial}{\partial\lambda}\left(-1 + \frac{X}{\lambda}\right)\right] \\
&= -N \cdot \mathbb{E}\left[-\frac{X}{\lambda^2}\right] \\
&= \frac{N}{\lambda}
\end{aligned}$$

showcasing the alternative procedure used to calculate the information number, which is often more straightforward. ■

Example 5.15. The information matrix of the normal distribution. The normal distribution has two parameters: hence, the Cramér-Rao bound is evaluated in a multivariate setting. Suppose one is working with a random sample; in this case, the information matrix is most easily calculated through the Hessian matrix of the logarithmic density function:

$$\boldsymbol{I}_N\left(\mu, \sigma^2\right) = -N \cdot \mathbb{E}\begin{bmatrix} \frac{\partial^2}{\partial \mu^2} \log\left[\frac{1}{\sigma}\phi\left(\frac{X-\mu}{\sigma}\right)\right] & \frac{\partial^2}{\partial \mu \partial \sigma^2} \log\left[\frac{1}{\sigma}\phi\left(\frac{X-\mu}{\sigma}\right)\right] \\ \frac{\partial^2}{\partial \sigma^2 \partial \mu} \log\left[\frac{1}{\sigma}\phi\left(\frac{X-\mu}{\sigma}\right)\right] & \frac{\partial^2}{\partial (\sigma^2)^2} \log\left[\frac{1}{\sigma}\phi\left(\frac{X-\mu}{\sigma}\right)\right] \end{bmatrix}$$

where $\phi(z)$, as usual, is the density function of the standard normal distribution. In analogy with the calculation of the Hessian matrix from Example 5.6, the above information matrix is as follows.

$$\boldsymbol{I}_N\left(\mu, \sigma^2\right) = -N \cdot \mathbb{E}\begin{bmatrix} -\frac{1}{\sigma^2} & -\frac{X-\mu}{\sigma^4} \\ -\frac{X-\mu}{\sigma^4} & \frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6} \end{bmatrix} = \begin{bmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{N}{2\sigma^4} \end{bmatrix}$$

Clearly, $\bar{X}$ is an unbiased estimator and its variance $\mathbb{V}\mathrm{ar}\left[\bar{X}\right] = \sigma^2/N$ attains the Cramér-Rao bound. However, while the estimator S^2 is unbiased:

$$\mathbb{V}\mathrm{ar}\left[S^2\right] = \frac{\sigma^4}{(N-1)^2}\,\mathbb{V}\mathrm{ar}\left[(N-1)\,\frac{S^2}{\sigma^2}\right] = \frac{2\sigma^4}{N-1}$$

it does not attain the Cramér-Rao bound, which is calculated as $2\sigma^4/N$ for unbiased estimators of σ^2. Consider instead the rescaled, biased estimator of the variance $\widehat{\sigma}^2 = \frac{N-1}{N}S^2$. Its variance is:

$$\mathbb{V}\mathrm{ar}\left[\frac{N-1}{N}S^2\right] = \frac{\sigma^4}{N^2}\,\mathbb{V}\mathrm{ar}\left[(N-1)\,\frac{S^2}{\sigma^2}\right] = \frac{2\,(N-1)\,\sigma^4}{N^2}$$

and to verify whether the Cramér-Rao bound is attained, one must calculate the latter by taking into account the bias. Calling $\mathcal{I}_N(\sigma^2)$ the bottom-right element of the information matrix $\boldsymbol{I}_N(\mu, \sigma^2)$, the bound is expressed as:

$$\begin{aligned} \mathbb{V}\mathrm{ar}\left[\frac{N-1}{N}S^2\right] &\geq \frac{1}{\mathcal{I}_N(\sigma^2)}\left(\frac{\partial}{\partial \sigma^2}\,\mathbb{E}\left[\frac{N-1}{N}S^2\right]\right)^2 \\ &= \frac{2\sigma^4}{N}\left[\frac{\partial}{\partial \sigma^2}\left(\frac{N-1}{N}\sigma^2\right)\right]^2 \\ &= \frac{2\,(N-1)^2\,\sigma^4}{N^3} \end{aligned}$$

and not even in this case it is attained. For both estimators, the actual value of the Cramér-Rao bound is equal to $\frac{N-1}{N}$ times their effective variance. ■

Following this discussion, one may be left wondering whether any mathematical result can help identify whether an unbiased estimator attains the Cramér-Rao bound or not. Such a result exists and is the following.

Theorem 5.6. Attainment of the Cramér-Rao Bound – univariate case. *In the (univariate) setup of Theorem 5.4, if $\widehat{\theta}$ is an unbiased estimator of θ, it attains the Cramér-Rao bound if and and only if:*

$$a_N(\theta)\left[\widehat{\theta}-\theta\right]=\frac{\partial}{\partial\theta}\log f_{X_1,\ldots,X_N}(x_1,\ldots,x_N;\theta)$$

for some function $a_N(\theta)$ of the parameter.

Proof. Recall the proof of Theorem 5.4 as well as Theorem 3.4: the equality is attained only if U (the estimator) is a linear function of V (the derivative of the logarithmic joint mass or density function of the sample, i.e. the log-likelihood function). By the Cauchy-Schwarz Inequality this can be phrased as $a(U-\mathbb{E}[U])=V$. As a can be a function of θ, write it as $a_N(\theta)$. □

Multivariate case. *In the (multivariate) setup of Theorem 5.4, if $\widehat{\boldsymbol{\theta}}$ is an unbiased estimator of $\boldsymbol{\theta}$, it attains the Cramér-Rao bound if and and only if:*

$$\mathbf{A}_N(\boldsymbol{\theta})\left[\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right]=\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}(\mathbf{x}_1,\ldots,\mathbf{x}_N;\boldsymbol{\theta})$$

for some $K\times K$ matrix $\mathbf{A}_N(\boldsymbol{\theta})$ which is a function of the parameters.

Proof. Similarly to the univariate case, the equality is only attained if $\boldsymbol{u}$ is a linear function of $\boldsymbol{v}$, i.e. $\mathbf{A}(\boldsymbol{u}-\mathbb{E}[\boldsymbol{u}])=\boldsymbol{v}$ where $\mathbf{A}=\mathbf{A}_N(\boldsymbol{\theta})$. □

Example 5.16. Attainment of the Cramér-Rao bound for estimators of the normal distribution. Consider again random sampling from the normal distribution. The derivative of the joint density corresponds to the MLE First Order Conditions as in Example 5.6; write them as:

$$\begin{bmatrix}\displaystyle\sum_{i=1}^{N}\frac{(x_i-\mu)}{\sigma^2}\\ \displaystyle\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^4}-\frac{N}{2\sigma^2}\end{bmatrix}=\underbrace{\begin{bmatrix}\frac{N}{\sigma^2}&0\\0&\frac{N}{2\sigma^4}\end{bmatrix}}_{=\mathbf{A}_N(\mu,\sigma^2)}\begin{bmatrix}\displaystyle\frac{1}{N}\sum_{i=1}^{N}x_i-\mu\\ \displaystyle\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{N}-\sigma^2\end{bmatrix}$$

as per Theorem 5.6. This decomposition not only shows again that $\bar{X}$ is an unbiased estimator of μ that attains the bound; it also reveals that the *only* unbiased estimator of σ^2 that attains the bound is $\widetilde{\sigma}^2=\sum_{i=1}^{N}(X_i-\mu)^2/N$: an estimator that is usually *unfeasible* because it requires an ex-ante perfect knowledge of the location parameter μ, an unlikely occurrence. ■

5.3 Tests of Hypotheses

The estimation of a parameter (or of other features of a probability distribution), which is performed by calculating the associated estimate obtained in the data, is usually only the first step of statistical analysis. In practical contexts, researchers usually aim to perform **statistical inference**, that is, probabilistic evaluations that concern the estimates and that are aimed at answering some real world questions of interest. In particular, researcher might be interested to evaluate the implications of their estimates on some **hypotheses** that they have formulated. The methods by which these evaluations are performed fall under the name of **tests of hypotheses**.

Tests of hypotheses are formulated as follows. Researchers first formulate some **null hypothesis** about their parameters of interest, that is, some statements about the baseline scenario that represents some initial belief or scenario to be evaluated. In general, this is written as:

$$\mathrm{H}_0 : \ \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$$

where H_0 is a common notation to represent the null hypothesis, $\boldsymbol{\theta}$ are the possibly multidimensional parameters of interest, and $\boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$ is the set of values in the parameter space $\boldsymbol{\Theta}$ that are permitted by the null hypothesis. By contrast, the **alternative hypothesis** is a statement that negates the null hypothesis:

$$\mathrm{H}_1 : \ \boldsymbol{\theta} \in \boldsymbol{\Theta}_0^c$$

where the H_1 is the usual notation for the alternative hypothesis, while $\boldsymbol{\Theta}_0^c$ is the complement of $\boldsymbol{\Theta}_0$ in the parameter space. It is helpful to represent the dichotomy between null and alternative hypotheses via some examples.

Example 5.17. Test on the mean of the normal distribution. The simplest case of an hypothesis test is that about the value of a single parameter, say the mean of the normal distribution. In this case, the null and alternative hypothesis read respectively as:

$$\mathrm{H}_0 : \ \mu = C \qquad\qquad \mathrm{H}_1 : \ \mu \neq C$$

where $|C| < \infty$ is a finite value. In a slightly more nuanced case, the two hypotheses are represented by two complementary inequalities. If $\boldsymbol{\Theta} = \mathbb{R}$, these can for example be:

$$\mathrm{H}_0 : \ \mu \geq C \qquad\qquad \mathrm{H}_1 : \ \mu < C$$

or vice versa. The two cases are usually referred to as **two-sided test** and **one-sided test**, respectively. A common scenario is the one for $C = 0$; in this case, the one-sided test is a test about the sign of the parameter. ■

Example 5.18. Test on the regression slope. A very common example of test is the one about the slope parameter of the linear regression model. In this case, the two hypotheses read, respectively:

$$\mathrm{H}_0: \ \beta_1 = C \qquad\qquad \mathrm{H}_1: \ \beta_1 \neq C$$

for the two-sided test, and:

$$\mathrm{H}_0: \ \beta_1 \geq C \qquad\qquad \mathrm{H}_1: \ \beta_1 < C$$

or vice versa for the one-sided test. In this specific case, the test for $C = 0$ is about the relevance of the exogenous variable X_i as an explanation of the endogenous variable Y_i (or as a "predictor" as it is sometimes said). ■

Example 5.19. Test on the parameter of the exponential distribution. The parameter space for the parameter λ of the exponential distribution is the set of positive values. Thus, the test with:

$$\mathrm{H}_0: \ 0 < \lambda \leq C \qquad\qquad \mathrm{H}_1: \ \lambda > C$$

is a proper formulation for testing whether the waiting time of some phenomenon of interest that can be modeled through the exponential distribution is lower or higher than some positive number $C > 0$. ■

Example 5.20. Test on the equality of the means of the multivariate normal distribution. Suppose that one is analyzing a phenomenon that can be modeled via the bivariate normal distribution, and is wondering whether the means of the two random variables involved (call them X and Y) are equal. In this case, the two hypotheses are formulated as:

$$\mathrm{H}_0: \ \mu_X - \mu_Y = 0 \qquad\qquad \mathrm{H}_1: \ \mu_X - \mu_Y \neq 0$$

which is a well-defined restriction in the parameter space. Next, consider the multivariate case, where it might be interesting to verify if all location parameters are equal to some specified value. Here, the hypotheses are:

$$\mathrm{H}_0: \ \mu_k = C_k \qquad\qquad \mathrm{H}_1: \ \mu_k \neq C_k$$

for $k = 1, \ldots, K$, where the restricted set $\boldsymbol{\Theta}_0$ is a specific point in $\mathbb{R}^K$. ■

Example 5.21. Test on the variance of the normal distribution. A researcher who aims to test the parameter representing the variance of a distribution, say the normal, must be conscious of the associated parameter space, e.g. $\sigma^2 > 0$. A sensible test in this case is:

$$\mathrm{H}_0: \ 0 < \sigma^2 \leq C \qquad\qquad \mathrm{H}_1: \ \sigma^2 > C$$

where, again, the constant $C > 0$ must be positive. ■

Example 5.22. Test on the variance ratio of two independent normal distributions. One may wonder about the relationship between the variances of two independent normal random variables X and Y. An test which is adequate for this environment is:

$$\mathrm{H}_0 : \frac{\sigma_X^2}{\sigma_Y^2} \leq C \qquad\qquad \mathrm{H}_1 : \frac{\sigma_X^2}{\sigma_Y^2} > C$$

where σ_X^2 and σ_Y^2 are the variances of X and Y, respectively. Here, $C < \infty$ must be finite but otherwise unrestricted. If $C = 1$ the test has an obvious interpretation: the null hypothesis represents the scenario where X has a variance smaller or equal than that of Y, while the alternative hypothesis states that the variance of X is larger than that of Y. Naturally, two-sided tests about specific values of the ratio are perfectly possible. ■

A test of hypothesis is conventionally conducted as follows.

1. The researcher establishes the two alternative hypotheses, H_0 and H_1.
2. The researcher identifies *ex-ante* those values of the sample realizations that are associated with acceptance of the null hypothesis, and rejection of the alternative hypothesis – this is called the **acceptance region** – as well as those values that are associated with acceptance of the alternative hypothesis, and rejection of the null hypothesis – the **rejection region**. The two sets must be complementary in the support of the sample.
3. The researcher examines the sample and performs a decision according to the criteria established at point 2. above.

Naturally, specifying the acceptance and rejection regions for large samples can be quite complicated, and maybe not extremely useful. Therefore, it is common to use univariate **test statistics** for this purpose.

Definition 5.7. Test statistic. In the context of some test of hypothesis, a *test* statistic $T = T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ is a statistic with support $\mathbb{T}$ and whose sample realization value is written as $t = T(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ which is such that, given a binary partition of the support $\mathbb{T}_0 \cup \mathbb{T}_0^c = \mathbb{T}$, the test is resolved as follows.

$$t \in \begin{cases} \mathbb{T}_0 & \Rightarrow \mathrm{H}_0 \text{ is accepted, and } \mathrm{H}_1 \text{ is rejected} \\ \mathbb{T}_0^c & \Rightarrow \mathrm{H}_1 \text{ is accepted, and } \mathrm{H}_0 \text{ is rejected} \end{cases}$$

In this setting, $\mathbb{T}_0$ and $\mathbb{T}_0^c$ are respectively called the **acceptance region** and the **rejection region** associated with the test statistic.

A proper test statistic is one whose probability distribution varies along with the parameters being tested, so that it is possible to associate probabilities for the acceptance and rejection regions that also vary as a function of the parameters under examination. The choice of a specific test statistic is usually test dependent. Before illustrating – with the aid of some examples – different test statistics for different hypotheses, it is necessary to discuss how the acceptance and rejection regions are determined. Ultimately, these are always arbitrary choices of researchers, that are however typically conducted according to certain conventions well grounded in statistical theory. This discussion requires some additional definitions.

Definition 5.8. Type I Error. In the framework of a test of hypotheses, the *type I* error is the circumstance whereby the null hypothesis is *rejected*, and the alternative hypothesis is *accepted*, while the null hypothesis is *true*.

Definition 5.9. Type II Error. In the framework of a test of hypotheses, the *type II* error is the circumstance whereby the null hypothesis is *accepted*, and the alternative hypothesis is *rejected*, while the null hypothesis is *false*.

The various outcomes of a test are commonly schematized as follows.

	$t \in \mathbb{T}_0$	$t \in \mathbb{T}_0^c$
$\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$	Correct decision	Type I error
$\boldsymbol{\theta} \in \boldsymbol{\Theta}_0^c$	Type II error	Correct decision

In an ideal test, both types of errors never occur; clearly, this ideal cannot be attained as otherwise it would not be necessary to conduct tests in the first place. At the same time, it is not possible to identify a criterion which is useful to simultaneously shrink both types of errors; since the probability to commit either depends on the acceptance and rejection regions, reducing one increases the other and vice versa. The following concept well represents the trade-off in question.

Definition 5.10. Power Function. The probability that the test statistic falls in the rejection region, as a function of the parameters $\boldsymbol{\theta}$, is the *power* function of a test.

$$\mathbb{P}_T(\boldsymbol{\theta}) = \mathbb{P}(t \in \mathbb{T}_0^c; \boldsymbol{\theta}) = 1 - \mathbb{P}(t \in \mathbb{T}_0; \boldsymbol{\theta})$$

Clearly, a power function expresses the probability to commit a Type I error if $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, and equals one minus the probability to commit a Type II error if $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0^c$. This notion, in turn, is instrumental in the following definitions.

Definition 5.11. Level of a test. Given a number $\alpha \in [0,1]$, a test with power function $\mathbb{P}_T(\boldsymbol{\theta})$ has *confidence level* α if $\mathbb{P}_T(\boldsymbol{\theta}) \leq \alpha$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$.

Definition 5.12. Size of a test. Given a number $\alpha \in [0,1]$, a test with power function $\mathbb{P}_T(\boldsymbol{\theta})$ has *size* α if $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \mathbb{P}_T(\boldsymbol{\theta}) = \alpha$.

The distinction between level and size is subtle, but it highlights aspects of the testing procedure. Given that a trade-off between Type I and Type II errors exists, the convention in statistical analysis is to restrict the attention to tests that have a sufficiently small probability of Type I errors (rejecting the null hypothesis when it is true), a value which is fixed at some α. These tests are said to have *confidence level* α. The confidence level is a *always* a discretionary choice of the researcher, but conventionally, α is chosen to be equal to one value between 0.1, 0.05, and 0.01. The smaller is the confidence level, the more credible is the outcome of the test when the null hypothesis is rejected (since that outcome has a smaller probability to occur *under the null hypothesis*). Once a confidence level is decided, a conscious researcher must recognize that attempting to further reduce the probability of Type I errors might be counterproductive, due to an increased probability of Type II errors. Thus, the attention is restricted to those tests whose maximum probability of rejecting the null hypothesis when it is true is exactly α: the *size* of the test.[4] In most practical applications, this nominal distinction is of little consequence, but it is important to make a correct use of terminology.

Example 5.23. Testing for the mean: level and size. Consider a test about the mean of a certain distribution, say the normal. In the two-sided case, $\boldsymbol{\Theta}_0 = \{C\}$ and $\boldsymbol{\Theta}_0^c = \mathbb{R} \setminus \{C\}$, hence there is no practical distinction between level and size. In the one-sided case, however, if the null hypothesis is that the mean is smaller or equal than some constant C, $\boldsymbol{\Theta}_0 = (-\infty, C]$ and $\boldsymbol{\Theta}_0^c = (C, \infty)$, and vice versa. Consequently, for a fixed level α there are different rejection probabilities for different values in $\boldsymbol{\Theta}_0$. In typical testing procedures, the maximum rejection probability is achieved at $\mu = C$. ■

After conducting tests, researchers usually report the following information enclosed to their statistical analyses: the confidence level, the decision outcome (acceptance vs. rejection) and often, a statistic called ***p*-value**.

Definition 5.13. The *p*-value. In a test of hypothesis with given size α, a *p-value* is a statistic $P = P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ such that for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, it is $\mathbb{P}(P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \leq \alpha) \leq \alpha$.

[4] Some advanced statistical theory of tests helps identify criteria to obtain the "optimal" tests, that is, those tests that minimize the probability of Type II errors for a fixed level α. This analysis is outside the scope of this discussion.

The definition of p-value is cumbersome and recursive, but delivers an intuition: the smaller the p-value associated with a sample, the smaller is the probability to observe that sample when *the null hypothesis is true*. Hence, a smaller p-value is interpreted in terms of less favorable evidence in favor of the null hypothesis. This concept allows researchers to evaluate the outcomes of tests on a more continuous scale, instead of being constrained by the "acceptance" vs. "rejection" dichotomy. Usually, p-values are obtained through the test statistics, by measuring the probability of observing realizations of the test statistic that are *even less favorable to the null hypothesis* than the actual realization t. This is best illustrated via some examples.

Example 5.24. Testing for the mean μ of the normal distribution: test statistics, error types and p-values. Suppose that some researcher is conducting a test about the mean of a population which is described by a normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$. To conduct the test, the researcher has collected a random sample drawn from X. Suppose *for the moment* that the researcher knows the variance σ^2 of the distribution of X. Additionally suppose, again *for the moment*, that the null hypothesis is that the mean is not a positive number.

$$\mathrm{H}_0 : \mu \leq 0 \qquad\qquad \mathrm{H}_1 : \mu > 0$$

As discussed in Lecture 4, in this environment the *standardized* sample mean is a statistic which follows the standard normal distribution: thus, a logical test protocol is to reject the null hypothesis if the observed standardized sample mean surpasses a certain **critical value**, call it z^*. This means that the test, for a given $\mu_0 \leq 0$, is resolved as follows:

$$\sqrt{N}\,\frac{\bar{X} - \mu_0}{\sigma} \begin{cases} \leq z^* & \Rightarrow \mu \leq 0 \\ > z^* & \Rightarrow \mu > 0 \end{cases}$$

where the arrows directed to the right indicate alternative conclusions about the value of μ. The ideal test for these hypotheses would be based on $\mu_0 = 0$: the reason is best illustrated by the following two observations.

1. The probability to conduct a Type I error:

$$\mathbb{P}\left(\text{Type I error}\right) = \mathbb{P}\left(\sqrt{N}\,\frac{\bar{X} - \mu_0}{\sigma} > z^* \middle| \mathrm{H}_0 \text{ is true}\right) = \alpha$$

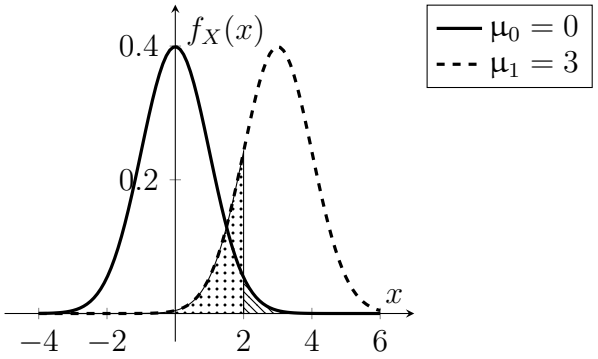
depends on the actual value of μ_0 in the expression above. It is obvious that the highest Type I error probability is attained for the supremum value of μ in the set defined by the null hypothesis: this value is clearly $\mu_0 = 0$. The probability of rejecting the null hypothesis associated with that value is the *size* α of the test.

2. Similarly, the probability to conduct a Type II error:

$$\mathbb{P}\left(\text{Type II error}\right) = \mathbb{P}\left(\sqrt{N}\,\frac{\bar{X}-\mu_1}{\sigma} \leq z^* \,\middle|\, \mathrm{H}_1 \text{ is true}\right)$$

is a function of the actual value of $\mu_1 > 0$ *if the alternative hypothesis is true*. The researcher is ignorant about this value, but it is clear that whatever the value, the probability increases along with z^*.

To illustrate the trade-off between the two error types probabilities, suppose that the alternative hypothesis is true and that the actual mean is $\mu_1 = 3$, while the decision cutoff is set by the researcher at $\sqrt{N}\cdot\bar{x}/\sigma > z^* = 2$. The standardized sample mean is centered at $\mu_0 = 0$ since that is the value that maximizes the probability of a Type I error, and simultaneously minimizes the probability of a Type II error. The two probabilities are respectively a decreasing and an increasing function of the threshold value z^*, and they are illustrated in Figure 5.2 below for the given threshold $z^* = 2$.



Note: this figure represents the probabilities of both a type I and a type II error when $\mathrm{H}_0 : \mu \leq 0$, the alternative hypothesis is true for $\mu_1 = 3$, and the testing protocol of the researcher is to reject the null hypothesis if the realized standardized sample mean centered at $\mu_0 = 0$ exceeds 2. The probability of a Type I error is thus the shaded area below the continuous density function centered at $\mu_0 = 0$ while the probability of a Type II error is the dotted area below the dashed density function centered at $\mu_1 = 3$.

Figure 5.2: Test on the mean of a normal distribution: error types I & II

Consider now the general case of a one-sided test, for some given C.

$$\mathrm{H}_0 : \mu \leq C \qquad\qquad \mathrm{H}_1 : \mu > C$$

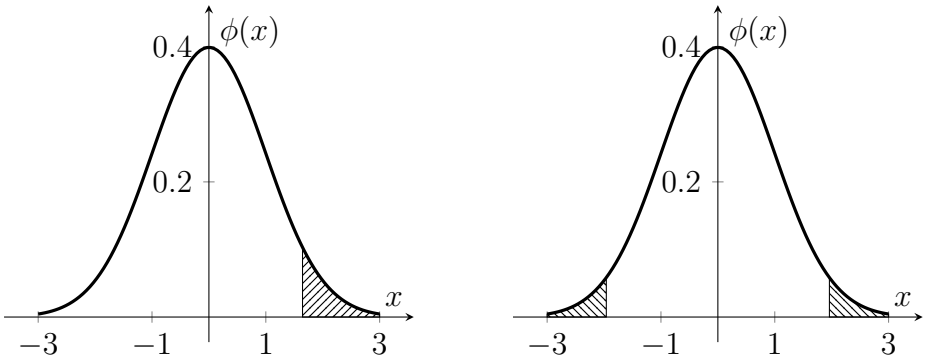
In order to use the standardized sample mean as an appropriate test statistic with a given size α, one must solve the following equation in terms of the critical value z^*_α, where the subscript indicates the size of the test.

$$\mathbb{P}\left(\bar{X} > C + \frac{\sigma}{\sqrt{N}}\,z^*_\alpha\right) = \mathbb{P}\left(\sqrt{N}\,\frac{\bar{X}-C}{\sigma} > z^*_\alpha\right) = \alpha$$

In the above expression, the second probability is evaluated with reference to a standard normal cumulative distribution $\Phi(z)$.[5] This procedure implies that the null hypothesis is rejected if the observed realization of the sample mean is such that $\sqrt{N}(\bar{x} - C)/\sigma > z^*_\alpha$; whatever the outcome of the test, the p-value is calculated as follows:

$$p(\bar{x}) = \mathbb{P}\left(\bar{X} \geq \bar{x}\right)$$

where again, the calculation is performed through the cumulative standard normal. To illustrate, consider the left panel of Figure 5.3, which depicts a standard normal's density function. The shaded area in the right tail represents the rejection region, the area corresponding to the realizations of the standardized sample mean that lead to the rejection of the null hypothesis, if $\alpha = 0.05$. In this case, the critical value is $z^*_{0.05} \approx 1.64$.



Note: the left panel depicts a one-sided test, the right-panel a two-sided test, both with size $\alpha = 0.05$. The shaded areas represent the corresponding rejection regions. The random variable X represented in both panels is the standardized sample mean centered at C; it follows the standard normal distribution. The critical values are, respectively, $z^*_{0.05} \approx 1.64$ in the left panel and $z^*_{0.025} \approx 1.96$ in the right panel.

Figure 5.3: Mean of the normal distribution: rejection region

Now suppose that the test is two-sided: the null hypothesis allows for only one value C, while the alternative hypothesis admits all other values.

$$\mathrm{H}_0 : \mu = C \qquad\qquad \mathrm{H}_1 : \mu \neq C$$

The researcher must now look for **two** *symmetric* **critical values**: $z^*_{\alpha/2} > 0$ and its mirror image $-z^*_{\alpha/2} < 0$. Intuitively, the researcher is agnostic about the sign of the deviation from C in case the alternative hypothesis is true;

[5] Once again, the standardized sample mean is centered at $\mu = C$ because with this choice, and for a fixed level α, the probability of a Type I error is maximized while the probability of a Type II error is minimized.

hence, given a level α the probabilities of *both* the Type I and the Type II errors are minimized when:

$$\mathbb{P}\left(\left|\bar{X}-C\right|>\frac{\sigma}{\sqrt{N}}z^*_{\alpha/2}\right)=\mathbb{P}\left(\sqrt{N}\,\frac{\left|\bar{X}-C\right|}{\sigma}>z^*_{\alpha/2}\right)=\frac{\alpha}{2}$$

and the null hypothesis is rejected if $\sqrt{N}\,|\bar{x}-C|\,/\sigma > z^*_{\alpha/2}$.[6] This is visually represented in the right panel of Figure 5.3, where the rejection region is composed by two symmetric tails of the standard normal distribution. Here, the p-value is calculated as *the sum of two symmetric probabilities*.

$$\begin{aligned}p\left(\bar{x}\right)&=\mathbb{P}\left(\bar{X}>\bar{x}\right)+\mathbb{P}\left(\bar{X}<-\bar{x}\right)\\&=2\cdot\mathbb{P}\left(\bar{X}>|\bar{x}|\right)\\&=2\cdot\mathbb{P}\left(\bar{X}>\bar{x}\right)\\&=2\cdot\mathbb{P}\left(\bar{X}<-\bar{x}\right)\end{aligned}$$

Two-sided tests about the mean of the normal distribution are perhaps the most common kinds of tests of hypotheses. It is thus useful to memorize the critical values associated with conventional confidence levels: $z^*_{0.05}\approx 1.64$ if $\alpha=0.1$, $z^*_{0.025}\approx 1.96$ if $\alpha=0.05$, and $z^*_{0.005}\approx 2.33$ if $\alpha=0.01$.

Suppose instead that the variance σ^2 is unknown by the researcher. In this case, the test statistic is unsurprisingly the t-statistic, where "t" stands for *test*. For example, in the previous case of a one-sided test the researcher should derive a critical value t^*_α according to the expression:

$$\mathbb{P}\left(\bar{X}>C+\frac{S}{\sqrt{N}}t^*_\alpha\right)=\mathbb{P}\left(\sqrt{N}\,\frac{\bar{X}-C}{S}>t^*_\alpha\right)=\alpha$$

where the second probability in the above display is evaluated in terms of a Student's t-distribution with $N-1$ degrees of freedom. Similarly as above, the test is rejected if $\sqrt{N}\,(\bar{x}-C)\,/s > t^*_\alpha$, and the p-value is calculated as the following function of both the *observed* sample mean and variance.

$$p\left(\bar{x},s^2\right)=\mathbb{P}\left(\frac{\bar{X}-C}{S}>\frac{\bar{x}-C}{s}\right)$$

The two-sided case bears symmetric analogies. One could graphically represent the two scenarios similarly as in Figure 5.3, but using the Student's t-distribution instead of the standard normal. ■

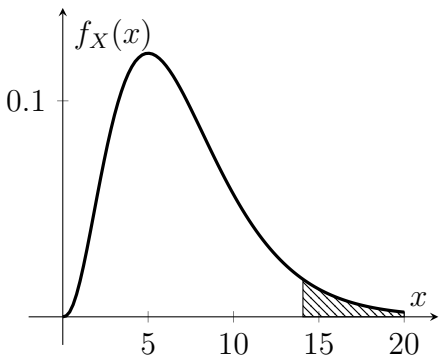
[6] Here, the standardized sample mean calculated for evaluating the test is centered at $\mu=C$ because this is the only value allowed by the null hypothesis.

Example 5.25. Testing for the variance σ^2 of the normal distribution: test statistic and *p*-values. To additionally elaborate the previous example, suppose that the researcher is testing the variance parameter σ^2 of the normally distributed population, with the same hypotheses outlined in Example 5.21 in the introduction to this section.

$$\mathrm{H}_0 : \ 0 < \sigma^2 \leq C \qquad\qquad \mathrm{H}_1 : \ \sigma^2 > C$$

Here, the test statistic is the rescaled sample variance $(N-1)\,S^2/C$, and the critical value k^*_α for a test of size α is identified through the chi-squared distribution with $N-1$ degrees of freedom (see Figure 5.4 below).

$$\mathbb{P}\left(S^2 > \frac{C}{N-1}\,k^*_\alpha\right) = \mathbb{P}\left((N-1)\,\frac{S^2}{C} > k^*_\alpha\right) = \alpha$$



Note: the shaded area represents the rejection region for a test with size $\alpha = 0.05$ on the variance of a normal distribution if $N = 8$. The represented random variable $X \sim \chi^2_7$ is the rescaled sample variance.

Figure 5.4: Variance of the normal distribution: rejection region

Thus, the null hypothesis is rejected if $(N-1)\,s^2/C > k^*_\alpha$, and the p-value is calculated as $p\,(s^2) = \mathbb{P}\,(S^2 \geq s^2)$. ■

Example 5.26. Testing for the variance ratio of two normal distributions: test statistic and *p*-values. Suppose now that the interest of the analyst falls the variances of two independent normally distributed populations. The relevant hypotheses are as in Example 5.22:

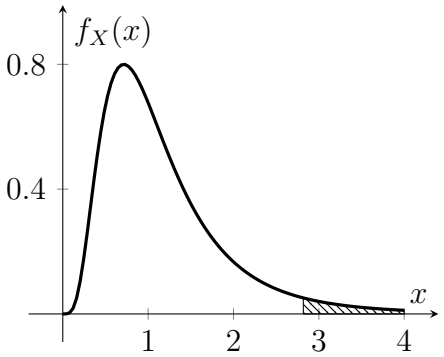
$$\mathrm{H}_0 : \ \frac{\sigma^2_X}{\sigma^2_Y} \leq C \qquad\qquad \mathrm{H}_1 : \ \frac{\sigma^2_X}{\sigma^2_Y} > C$$

and by the analysis conducted in Lecture 4, the relevant test statistic is the F-statistic $\mathrm{F} = S^2_X/S^2_Y C$. Thus, the critical value k^*_α for a test of size α is

obtained by evaluating an F-distribution with paired $N_X - 1$ and $N_Y - 1$ degrees of freedom, as per the expression:

$$\mathbb{P}\left(\frac{S_X^2}{S_Y^2} > Ck_\alpha^*\right) = \mathbb{P}\left(\frac{S_X^2}{S_Y^2}\frac{1}{C} > k_\alpha^*\right) = \alpha$$

and the illustration is given in Figure 5.5 below.



Note: the shaded area represents the rejection region for a test with size $\alpha = 0.05$ on the normal variance ratio if $N_X = N_Y = 12$. The represented random variable $X \sim \mathcal{F}_{11,11}$ is the F-statistic.

Figure 5.5: Ratio of two normal distributions' variances: rejection region

In this scenario, the null hypothesis is rejected if $(s_X^2/s_Y^2)/C > k_\alpha^*$ and the p-value is calculated as $p(s_X^2, s_Y^2) = \mathbb{P}(S_X^2/S_Y^2 > s_X^2/s_Y^2)$. ■

Example 5.27. Testing multiple means of the multivariate normal distribution: test statistic and *p*-values. Consider now some *composite* hypotheses about multiple parameters – specifically, multiple means – of the multivariate normal distribution (recall Example 5.20):

$$\mathrm{H}_0: \ \mu_k = C_k \qquad\qquad \mathrm{H}_1: \ \mu_k \neq C_k$$

for $k = 1, \ldots, K$. This test is best expressed in vectorial form:

$$\mathrm{H}_0: \ \boldsymbol{\mu} = \mathbf{c} \qquad\qquad \mathrm{H}_1: \ \boldsymbol{\mu} \neq \mathbf{c}$$

where $\boldsymbol{\mu}$ is the vector of means and $\mathbf{c} = (C_1, \ldots, C_K)^{\mathrm{T}}$. A researcher who knows the matrix $\boldsymbol{\Sigma}$ of variance-covariance parameters is in the position to compute the following so-called u-statistic:

$$\mathrm{u} = N(\bar{\boldsymbol{x}} - \mathbf{c})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{x}} - \mathbf{c}) \sim \chi_K^2$$

which, by the properties of the sample mean $\bar{\boldsymbol{x}}$ drawn from a multivariate normal distribution, follows the chi-squared distribution with K degrees of freedom. The u-statistic can thus be used to test the hypothesis of interest when the parameter matrix $\boldsymbol{\Sigma}$ is known.

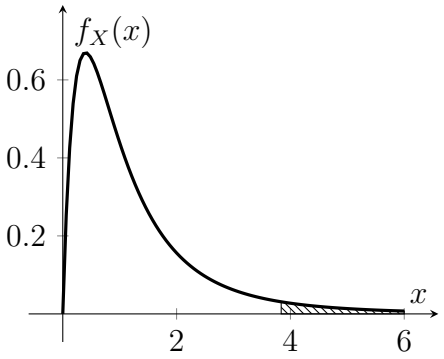
In general, however, $\boldsymbol{\Sigma}$ is unknown and the sample variance-covariance matrix $\boldsymbol{S}$ is used in its place, hence the test statistic adopted is the rescaled Hotelling's t-squared statistic, already introduced in Lecture 4.

$$\frac{N-K}{K\,(N-1)}\,\mathrm{t}^2 = \frac{(N-K)\,N}{K\,(N-1)}\,(\bar{\boldsymbol{x}}-\mathbf{c})^{\mathrm{T}}\,\boldsymbol{S}^{-1}\,(\bar{\boldsymbol{x}}-\mathbf{c}) \sim \mathcal{F}_{K,N-K}$$

This test statistic follows the F-distribution with paired degrees of freedom K and $N-K$. While this is a two-sided test, the null-hypothesis is rejected if the observed test statistic surpasses a certain critical values k^*_α, an indication that the true means are effectively likely larger than $\mathbf{c}$.[7] Given a size α, this critical value is defined as follows.

$$\mathbb{P}\left(\frac{N-K}{K\,(N-1)}\,\mathrm{t}^2 > k^*_\alpha\right) = \alpha$$

This critical value is obtained as a quantile of an appropriate F-distribution, as shown in Figure 5.6 below.



Note: the shaded area represents the rejection region for a test with size $\alpha = 0.05$ about $K = 4$ means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ of a multivariate normal distribution, with $N = 12$. The represented random variable $X \sim \mathcal{F}_{4,8}$ is the rescaled Hotelling's t-squared statistic which is discussed above in the text.

Figure 5.6: Mean of the multivariate normal distribution: rejection region

The p-value in this case is the probability to observe an Hotelling's t-squared statistic which is larger than the actual realization.

$$p\,(\bar{\mathbf{x}}, \mathbf{S}) = \mathbb{P}\left(\mathrm{t}^2 > N\,(\bar{\mathbf{x}}-\mathbf{c})^{\mathrm{T}}\,\mathbf{S}^{-1}\,(\bar{\mathbf{x}}-\mathbf{c})\right)$$

Clearly, it is perfectly possible to test only a subset $L < K$ of the means of the multivariate normal distribution, for L of its variables. In this case, the relevant sample variance-covariance $\boldsymbol{S}$ has dimension $L \times L$, and the test statistic follows an F-distribution with degrees of freedom L and $N - L$.

[7] Negative deviations of the kind $\bar{X}_k - C_k$ for $k = 1, \ldots, K$ contribute positively to the calculation of the test statistic, since they are squared.

The logic of the test can be generalized. For example, if the hypotheses of interest were about the equality of the means of a random vector (X, Y) that follows the bivariate normal distribution, like:

$$\mathrm{H}_0 : \ \mu_X - \mu_Y = 0 \qquad\qquad \mathrm{H}_1 : \ \mu_X - \mu_Y \neq 0$$

the appropriate test-statistic is the following version of Hotelling's t-squared statistic:[8]

$$\mathrm{t}^2 = \frac{N\left(\bar{X} - \bar{Y}\right)^2}{S_X^2 + S_Y^2 - 2S_{XY}} \sim \mathcal{F}_{1,N-1}$$

which follows the F-distribution with paired degrees of freedom 1 and $N-1$, and where:

$$S_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

is the sample covariance between X and Y. This obtains from applying the testing procedure to the following transformed random variable.

$$W = X - Y \sim \mathcal{N}\left(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2 - 2\rho_{XY}\sigma_X\sigma_Y\right)$$

Note that it would be inappropriate to test this specific hypothesis by simply analyzing the difference between the two standardized sample means while disregarding their covariance, *unless* there are good reasons to believe that the two random variables X and Y are independent. ■

Example 5.28. Testing for the parameter λ of the exponential distribution: test statistic and *p*-values. Abandon for once the familiar framework of a random sample drawn from the normal distribution, and consider instead that of a random sample drawn from a random variable $X \sim \mathrm{Exp}\left(\lambda\right)$ that follows the exponential distribution; In this setting the sample mean $\bar{X}$ has an easily identifiable distribution. By the properties of moment generating functions:

$$M_{\bar{X}}\left(t\right) = \left[M_X\left(\frac{1}{N}\,t\right)\right]^N = \left(1 - \frac{\lambda}{N}\,t\right)^{-N}$$

and therefore:

$$\bar{X} \sim \Gamma\left(N, \frac{N}{\lambda}\right)$$

that is, $\bar{X}$ follows the Gamma distribution with the given parameters.

[8]The name *t-squared* associated with Hotelling's statistic comes from the relationship between the Student's t-distribution and the F-distribution. As it was already observed in Lecture 2, if $X \sim \mathcal{T}\left(\nu\right)$ and $Y = X^2$, it is $Y \sim \mathcal{F}\left(1, \nu\right)$.

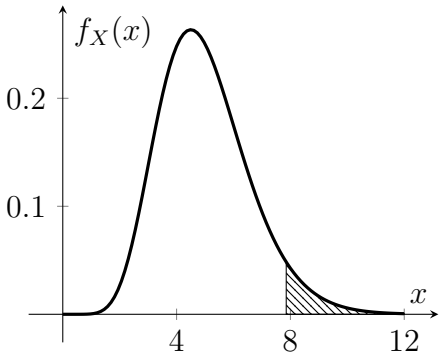
It is thus intuitive to use $\bar{X}$ as a test statistic for λ based on that Gamma distribution, should a researcher be interested about testing that parameter. Specifically, let the hypotheses at hand be the following (Example 5.19).

$$\mathrm{H}_0 : \ 0 < \lambda \leq C \qquad\qquad \mathrm{H}_1 : \ \lambda > C$$

Given a test size α, the critical value g_α^* is evaluated as:

$$\mathbb{P}\left(\bar{X} > C\left(\frac{g_\alpha^*}{\sqrt{N}} + 1\right)\right) = \mathbb{P}\left(\sqrt{N}\,\frac{\bar{X} - C}{C} > g_\alpha^*\right) = \alpha$$

and the null hypothesis is rejected if $\bar{x} > Cg_\alpha^*/\sqrt{N} + C$; the illustration is given in Figure 5.7 below.



Note: the shaded area represents the rejection region for a test with size $\alpha = 0.05$ on the parameter λ of an exponential distribution if $N = 10$ and $C = 5$. The represented random variable $X \sim \Gamma(10, 2)$ is the rescaled sample mean under the null hypothesis that $\lambda = C$. Note that both parameters depend on C.

Figure 5.7: Exponential distribution's parameter λ: rejection region

Here, the p-value is calculated similarly as in the one-sided normal test, that is $p(\bar{x}) = \mathbb{P}\left(\bar{X} \geq \bar{x}\right)$. ■

This example almost exhausts the analysis of the alternative hypotheses introduced at the beginning of this section: only the test about the linear regression slope parameter β_1 (Example 5.18) has been left out. The reason is that the distribution of any estimator of β_1 (say, the Method of Moments estimator from Example 5.4) depends on the underlying assumptions about the conditional distribution of $Y_i | X_i$. If, for example, such a distribution is normal, one can show that the distribution of the estimator is also normal, and therefore tests would proceed as in Example 5.24. Observe that most of the previous examples analyze one-sided tests; a useful exercise is to recast them as two-sided tests, derive expressions for their critical values, and plot the rejection regions. Finally, it must be observed that some of these tests can be simplified in an asymptotic environment, as discussed in Lecture 6.

5.4 Interval Estimation

However precise, any exercise in parameter estimation is always uncertain. Acknowledging this fact, statistical analysts usually supplement point estimates with other "likely" values of their parameters of interest, so to better inform the understanding of real world phenomena. This exercise falls under the name of **interval estimation**. A formal definition follows.

Definition 5.14. Interval estimators. Consider the statistical inference about a *scalar* parameter θ. An *interval* estimator is a pair of statistics that are functions of the sample: the "lower" bound statistic $L = L(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ and the "upper" bound statistic $U = U(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$, such that $L \leq U$ and that if the values $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ are observed, the conclusion of the statistical inference is that θ falls in the interval defined by the two realized statistics.

$$\theta \in [L(\mathbf{x}_1, \ldots, \mathbf{x}_N), U(\mathbf{x}_1, \ldots, \mathbf{x}_N)]$$

The interval in question is also called *confidence interval*.

Naturally, a confidence interval can be made however large so to increase the chances that the "true" parameter θ falls inside it; however, the larger the interval the less informative it is! At the extreme, the confidence interval can encompass the entire parameter space for θ, which clearly makes the entire exercise moot. Therefore, a good confidence interval is one that is *as small as possible* while having a probability to *include the true parameter* θ which is *as high as possible*. To evaluate this property, one must take into account the following concepts.

Definition 5.15. Coverage probability. The *coverage* probability that is associated with an interval estimator is the probability that the associated confidence interval covers the true parameter, for a *given* parameter θ.

$$\text{Coverage Probability} = \mathbb{P}(L(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \leq \theta \leq U(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N))$$

Definition 5.16. Confidence coefficient. The confidence *coefficient* that is associated with an interval estimator is the infimum of all the confidence probabilities in the parameter space of θ (write it as Θ).

$$\text{Confidence Coefficient} = \inf_{\theta \in \Theta} \mathbb{P}(L(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \leq \theta \leq U(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N))$$

Note that the probabilities defined above depend on the chosen statistics L and U (two random variables), and are evaluated in the sample space defined by the support of the sample. In practice, the distinction between the two definitions is often irrelevant, because in many common cases the coverage probability does not vary in the parameter space. When it varies, however, an interval estimator is evaluated in terms of the confidence coefficient.

This problem should be reminiscent of the analogous issue discussed in the setting of tests of hypothesis: how to select the acceptance and rejection regions while minimizing the combined chance of errors? The analogy is self-evident and thus, it should not be too surprising that the statistical methods for constructing confidence intervals are intimately related to techniques for the conduction of hypothesis tests. In fact, all methods for the construction of confidence intervals are related to tests; here, only one of these methods is described as all the others can be related to it. This method goes by the name of *inversion of test statistics* and it proceeds as follows.

1. Start from a *two-sided* hypothesis about θ.

$$\mathrm{H}_0 : \ \theta = C \qquad\qquad \mathrm{H}_1 : \ \theta \neq C$$

2. Construct an *acceptance region* for C based on a test statistic T that is a function of C:

$$\mathbb{T}_0 = \left\{ T\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; C\right) \in \left[k^{**}_{1-\alpha/2}, k^{*}_{\alpha/2} \right] \right\}$$

where $k^{**}_{1-\alpha/2}$ and $k^{*}_{\alpha/2}$ are two suitable critical values associated with, respectively, the $(\alpha/2)$-th and $(1-\alpha/2)$-th quantiles of the distribution of the test statistic; if the latter is symmetric around zero it holds $k^{**}_{1-\alpha/2} = -k^{*}_{\alpha/2}$. The notation here is somewhat counterintuitive, since $k^{**}_{1-\alpha/2}$ corresponds to the $(\alpha/2)$-th quantile:

$$\mathbb{P}\left(T\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; C\right) > k^{**}_{1-\alpha/2} \right) = 1 - \frac{\alpha}{2}$$

and symmetrically $k^{*}_{\alpha/2}$ corresponds to the $(1-\alpha/2)$-th quantile:

$$\mathbb{P}\left(T\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; C\right) > k^{*}_{\alpha/2} \right) = \frac{\alpha}{2}$$

This notation is chosen for the sake of consistency with the more general treatment of tests. The above acceptance region $\mathbb{T}_0$ is associated with a size α which is defined in terms of the following probability.

$$\mathbb{P}\left(k^{**}_{1-\alpha/2} \leq T\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; C\right) \leq k^{*}_{\alpha/2} \right) = 1 - \alpha$$

Note that this equals one minus the probability of a Type I error.

3. Construct the following two statistics by *inverting* the function that defines the test statistic, $T\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; C\right)$, with respect to C, and by evaluating the inverse at the two critical values.

$$I_1 = T^{-1}\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; k^{**}_{1-\alpha/2}\right)$$
$$I_2 = T^{-1}\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; k^{*}_{\alpha/2}\right)$$

4. The interval estimator is finally obtained as:

$$L(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \min\{I_1, I_2\}$$
$$U(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \max\{I_1, I_2\}$$

where typically, $L = I_2$ and $U = I_1$. Note that the coverage probability associated with the interval estimator is also $1 - \alpha$, since for any $C = \theta$ the procedure just described implies the following.

$$\mathbb{P}(L(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \leq \theta \leq U(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)) = 1 - \alpha$$

This procedure appears abstract and convoluted, but since most test statistics are simple function of the parameters, the inversion is generally straightforward and intuitive. The method is best illustrated via examples.

Example 5.29. The confidence interval for the mean of the normal distribution. As described in Example 5.24, in two-sided tests about the mean of the normal distribution in the case where the variance σ^2 is known, the acceptance region is defined in terms of the following interval:

$$\mathbb{T}_{0,\mu} = \left\{ \sqrt{N} \frac{\bar{X} - C}{\sigma} \in \left[-z^*_{\alpha/2}, z^*_{\alpha/2} \right] \right\}$$

since the null hypothesis is rejected if $\sqrt{N} |\bar{x} - C| / \sigma > z^*_{\alpha/2}$. Note that in this settings one can seamlessly convert open intervals into closed ones and vice versa, since realizations equal to a specific value have probability zero. The two critical values $-z^*_{\alpha/2}$ and $z^*_{\alpha/2}$ are evaluated in terms of the standard normal distribution, which is symmetric around zero. Thus, the confidence interval for μ is:

$$\mu \in \left[\bar{X} - \frac{\sigma}{\sqrt{N}} z^*_{\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{N}} z^*_{\alpha/2} \right]$$

which is obtained easily, since the test statistic is a simple linear function of the parameter. Note that the function is also a monotonically decreasing one, hence $L = I_2$ and $U = I_1$ according to the procedure described earlier. If instead the variance σ^2 is unknown, the analogous procedure based on the t-statistic results in the analogous confidence interval:

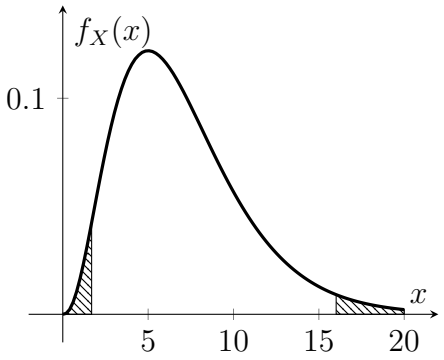
$$\mu \in \left[\bar{X} - \frac{S}{\sqrt{N}} t^*_{\alpha/2}, \bar{X} + \frac{S}{\sqrt{N}} t^*_{\alpha/2} \right]$$

where $t^*_{\alpha/2}$ is evaluated in terms of the Student's t-distribution with $N - 1$ degrees of freedom, which is also symmetric around zero. ■

Example 5.30. The confidence interval for the variance of the normal distribution. By extending Example 5.25, a *two-sided* test about the variance of the of the normal distribution would have acceptance region:

$$\mathbb{T}_{0,\sigma^2} = \left\{ (N-1)\,\frac{S^2}{C} \in \left[k^{**}_{1-\alpha/2}, k^{*}_{\alpha/2}\right] \right\}$$

where $k^{**}_{1-\alpha/2}$ and $k^{*}_{\alpha/2}$ are evaluated in terms of the chi-squared distribution with $N-1$ degrees of freedom. To appreciate the difference with Example 5.25, see Figure 5.8 below.



Note: the shaded area displays the rejection region for a *two-sided* version of the test in Example 5.25.

Figure 5.8: Variance of the normal distribution: two-sided rejection region

Therefore, the confidence interval for σ^2 is:

$$\sigma^2 \in \left[(N-1)\,\frac{S^2}{k^{*}_{\alpha/2}}, (N-1)\,\frac{S^2}{k^{**}_{1-\alpha/2}} \right]$$

and again, it is $L = I_2$ and $U = I_1$ because the test statistic is decreasing in the parameter of interest σ^2. ■

Example 5.31. The confidence interval for the variance ratio from two normal distributions. The case of the variance ratio σ^2_X/σ^2_Y from two samples drawn from two independent normal distributions is analogous; the *two-sided* version of Example 5.26, gives the following acceptance region:

$$\mathbb{T}_{0,\frac{\sigma^2_X}{\sigma^2_Y}} = \left\{ \frac{S^2_X}{S^2_Y}\frac{1}{C} \in \left[k^{**}_{1-\alpha/2}, k^{*}_{\alpha/2}\right] \right\}$$

where $k^{**}_{1-\alpha/2}$ and $k^{*}_{\alpha/2}$ are evaluated as quantiles of the F-distribution with paired degrees of freedom $N_X - 1$ and $N_Y - 1$. Consequently, the confidence

interval for the variance ratio is:

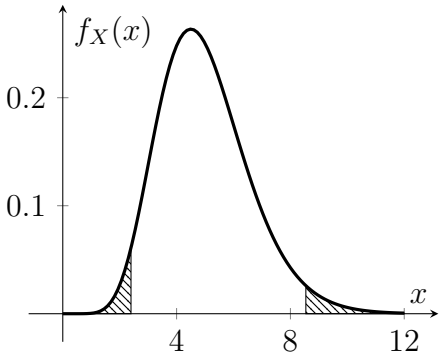
$$\frac{\sigma_X^2}{\sigma_Y^2} \in \left[\frac{S_X^2}{S_Y^2} \frac{1}{k^*_{\alpha/2}}, \frac{S_X^2}{S_Y^2} \frac{1}{k^{**}_{1-\alpha/2}} \right]$$

and it is once again $L = I_2$ and $U = I_1$. ■

Example 5.32. The confidence interval for the parameter λ of the exponential distribution. The inference about the parameter λ of the exponential distribution bears similarities with that about the mean of the normal distribution, the differences are that the parameter λ characterizes both the mean and the variance of the exponential distribution, and that the distribution of the test statistic is asymmetric and has support on positive values only. By elaborating the analysis from Example 5.28 as a *two-sided* test, the acceptance region becomes:

$$\mathbb{T}_{0,\lambda} = \left\{ \sqrt{N} \frac{\bar{X} - C}{C} \in \left[g^{**}_{1-\alpha/2}, g^*_{\alpha/2} \right] \right\}$$

where $g^{**}_{1-\alpha/2}$ and $g^*_{\alpha/2}$ are appropriate quantiles of the Gamma distribution with parameters $\alpha = N$ and $\beta = N/\lambda$, as shown in Figure 5.9.



Note: the shaded area displays the rejection region for a *two-sided* version of the test in Example 5.28.

Figure 5.9: Exponential distribution's par. λ: rejection region, two-sided

As a result, by the usual procedure one obtains:

$$\lambda \in \left[\frac{\bar{X}}{1 + N^{-1/2} g^*_{\alpha/2}}, \frac{\bar{X}}{1 + N^{-1/2} g^{**}_{1-\alpha/2}} \right]$$

that is, a confidence interval for λ with coverage probability $1 - \alpha$. ■

These examples show that, in general, the expression of the confidence interval as a function of the test statistic is contextual: it must be derived on a case-by-case basis. However, applications about the normal distribution dominate, thanks to the asymptotic results developed in the next Lecture.

Lecture 6

Asymptotic Analysis

This lecture introduces the fundamental concepts of asymptotic probability theory, and associated results that allow to expand and simplify statistical analysis in settings where data samples of sufficiently large size are available. More specifically, this lecture builds up the set of definitions and properties that are necessary for an appropriate analysis of the Laws of Large Numbers and the Central Limit Theorems; subsequently, it proves a relatively simple version of both sets of results and discusses their implications in terms of the asymptotic behavior of simple sample statistics and other estimators.

6.1 Convergence in Probability

The concept of *convergence in probability* relates to the idea of some random variables "approaching" specific values in the support of their distribution when the size of the sample grows very large. Convergence in probability is suited to characterize the "asymptotic" behavior of statistics and estimators in so-called *large samples*. Most typically, interest falls on those statistics that converge in probability to certain population parameters or moments of interest. To better introduce this concept it is necessary to formalize the notion of *sequences* of random variables and vectors.

Definition 6.1. Random sequence. Any random vector expressed as an N-indexed sequence, write it as $\boldsymbol{x}_N = (X_{1N}, \ldots, X_{KN})^{\mathrm{T}}$, is called a random *sequence*. In the univariate context ($K = 1$), one can write it simply as X_N.

The definition can be further extended to sequences of *random matrices* with dimension $J \times K$, that combine J vectorial sequences $\boldsymbol{x}_{jN}$ of length K for $j = 1, \ldots, J$. Such a matrix is indicated for example as follows.

$$\boldsymbol{X}_N = \begin{bmatrix} \boldsymbol{x}_{1N} & \boldsymbol{x}_{2N} & \ldots & \boldsymbol{x}_{JN} \end{bmatrix}^{\mathrm{T}}$$

Example 6.1. Common random sequences. Both the sample mean (a random vector) and the sample variance-covariance (a random matrix):

$$\bar{\boldsymbol{x}}_N = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \quad \text{and} \quad \boldsymbol{S}_N = \frac{1}{N-1}\sum_{i=1}^{N} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_N)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_N)^{\mathrm{T}}$$

are two random sequences, as they are statistics that depend on the sample size N. Their univariate versions are usually written as $\bar{X}_N$ and S_N^2.

Endowed with the definition of random sequence, it is possible to express the concept of convergence more formally. A first, intuitive requirement for convergent random sequences is that the latter are somehow "bounded," in the sense that as N grows the probability distribution of $\boldsymbol{x}_N$ concentrates around a subset of the support. This is expressed with the following concept.

Definition 6.2. Boundedness in Probability. A sequence $\boldsymbol{x}_N$ of random vectors is *bounded in probability* if and only if, for any $\varepsilon > 0$, there exists some number $\delta_\varepsilon < \infty$ and an integer N_ε such that

$$\mathbb{P}(\|\boldsymbol{x}_N\| \geq \delta_\varepsilon) < \varepsilon \quad \forall N \geq N_\varepsilon$$

which is also written as $\boldsymbol{x}_N = \mathcal{O}_p(1)$ and read as "$\boldsymbol{x}_N$ is big p-oh one."

However this is not quite enough, because this definition still allows for the probability distribution of $\boldsymbol{x}_N$ to remain "dense" within some specific interval but without shrinking into a unique point, even if N grows very large. In fact, the concept of convergence in probability that is most frequently adopted in these lectures is stronger than boundedness in probability.

Definition 6.3. Convergence in Probability. A sequence $\boldsymbol{x}_N$ of random vectors converges in probability to a constant vector $\mathbf{c}$ if

$$\lim_{N\to\infty} \mathbb{P}(\|\boldsymbol{x}_N - \mathbf{c}\| > \delta) = 0$$

for *any* positive real number $\delta > 0$.

The above definition formalizes the idea that as the sample size N grows increasingly larger, the probability distribution of $\boldsymbol{x}_N$ concentrates within an increasingly smaller neighborhood of $\mathbf{c}$. In terms of notation, convergence in probability is usually denoted in the two following alternative ways:

$$\boldsymbol{x}_N \xrightarrow{p} \mathbf{c}$$
$$\operatorname{plim} \boldsymbol{x}_N = \mathbf{c}$$

among which the former is preferred in these lectures. In fact, it is easy to see that in analogy with real sequences, convergence in probability is just a special case of boundedness in probability.

Theorem 6.1. Convergent Random Sequences are also Bounded. *If some sequence $\boldsymbol{x}_N$ of random vectors converges in probability to some constant $\mathbf{c}$, that is $\boldsymbol{x}_N \xrightarrow{p} \mathbf{c}$, then it is also bounded: $\boldsymbol{x}_N = \mathcal{O}_p(1)$.*

Proof. By the definition of convergence in probability, for any $\varepsilon > 0$ there is always an integer N_ε such that

$$\mathbb{P}\left(\|\boldsymbol{x}_N - \mathbf{c}\| > \delta\right) < \varepsilon \quad \forall N \geq N_\varepsilon$$

thus by setting $\delta_\varepsilon = \delta + \|\boldsymbol{x}_{N_\varepsilon}\| - \|\boldsymbol{x}_{N_\varepsilon} - \mathbf{c}\|$ one gets that $\boldsymbol{x}_N = \mathcal{O}_p(1)$. □

This statement and its proof properly clarify the difference between boundedness and convergence in probability: while the former is valid for a specific constant δ_ϵ so long as N large enough (and so long it exists), the latter must be true for any δ instead.

In the specific case of convergence in probability (Definition 6.3) where $\mathbf{c} = \mathbf{0}$, one can also write:

$$\boldsymbol{x}_N = o_p(1)$$

which is read as "$\boldsymbol{x}_N$ is little p-oh one." The use of the "probability" version of the the big-oh and little-oh notation facilitates outlining the properties of probability limits with respect to real sequences, which are analogous to the non-stochastic case.

Definition 6.4. Convergence of Random to Real Sequences. Consider a random sequence $\boldsymbol{x}_N$ and a *non*-random sequence $\mathbf{a}_N$ of the same dimension K. Moreover, define the random sequence $\boldsymbol{z}_N = (Z_{1N}, \ldots, Z_{KN})^{\mathrm{T}}$ where $Z_{kn} = X_{kn}/a_{kn}$ for $k = 1, \ldots, K$ and for $n = 1, 2, \ldots$ to infinity.

1. If $\boldsymbol{z}_N = \mathcal{O}_p(1)$, then $\boldsymbol{x}_N$ is said to be bounded in probability by $\mathbf{a}_N$, which one can write as $\boldsymbol{x}_N = \mathcal{O}_p(\mathbf{a}_N)$.
2. If $\boldsymbol{z}_N = o_p(1)$, then $\boldsymbol{x}_N$ is said to converge in probability to $\mathbf{a}_N$, which one can write as $\boldsymbol{x}_N = o_p(\mathbf{a}_N)$.

There are further definitions of "convergence" in probabilistic sense that are even stronger than convergence in probability.

Definition 6.5. Convergence in r-th Mean. A sequence $\boldsymbol{x}_N$ of random vectors is said to converge in r-th mean to a constant vector $\mathbf{c}$ under the following condition.

$$\lim_{N\to\infty} \mathbb{E}\left[\|\boldsymbol{x}_N - \mathbf{c}\|^r\right] = 0$$

In the special case where $r = 2$, this concept is known as **Convergence in Quadratic Mean** and is also expressed as follows.

$$\boldsymbol{x}_N \xrightarrow{qm} \mathbf{c}$$

The following two useful results show that whenever some random sequence converges in quadratic or higher mean to some specific vector (such as, say, its mean), it also converges in probability to it.

Theorem 6.2. Convergence in Lower Means. *A random sequence $\boldsymbol{x}_N$ that converges in r-th mean to some constant vector $\mathbf{c}$ also converges in s-th mean to $\mathbf{c}$ for $s < r$.*

Proof. The proof is based on Jensen's Inequality:

$$\lim_{N\to\infty} \mathbb{E}\left[\|\boldsymbol{x}_N - \mathbf{c}\|^s\right] = \lim_{N\to\infty} \mathbb{E}\left[\left(\|\boldsymbol{x}_N - \mathbf{c}\|^r\right)^{\frac{s}{r}}\right] \leq \lim_{N\to\infty} \left\{\mathbb{E}\left[\|\boldsymbol{x}_N - \mathbf{c}\|^r\right]\right\}^{\frac{s}{r}} = 0$$

since $\lim_{N\to\infty} \mathbb{E}\left[\|\boldsymbol{x}_N - \mathbf{c}\|^r\right] = 0$. □

Theorem 6.3. Convergence in Quadratic Mean and Probability. *If a random sequence $\boldsymbol{x}_N$ converges in r-th mean to a constant vector $\mathbf{c}$ for $r \geq 2$ (that is,* at least *$\boldsymbol{x}_N \xrightarrow{qm} \mathbf{c}$), then it also converges in probability to $\mathbf{c}$.*

Proof. Define the (one-dimensional) *nonnegative* random sequence Q_N as:

$$Q_N = \|\boldsymbol{x}_N - \mathbf{c}\| = \sqrt{(\boldsymbol{x}_N - \mathbf{c})^{\mathrm{T}} (\boldsymbol{x}_N - \mathbf{c})} \in \mathbb{R}_+$$

and notice that by Theorem 6.3 it must converge in *first* mean:

$$\lim_{N\to\infty} \mathbb{E}\left[Q_N\right] = \lim_{N\to\infty} \mathbb{E}\left[\|\boldsymbol{x}_N - \mathbf{c}\|\right] = 0$$

and therefore, quadratic mean convergence also implies the following.

$$\lim_{N\to\infty} \mathbb{V}\mathrm{ar}\left[Q_N\right] = \lim_{N\to\infty} \mathbb{E}\left[Q_N^2\right] = \lim_{N\to\infty} \mathbb{E}\left[\|\boldsymbol{x}_N - \mathbf{c}\|^2\right] = 0$$

At the same time, by Čebyšëv's Inequality:

$$\mathbb{P}\left(|Q_N - \mathbb{E}\left[Q_N\right]| > \delta\right) \leq \frac{\mathbb{V}\mathrm{ar}\left[Q_N\right]}{\delta^2}$$

therefore, taking limits on both sides gives:

$$\lim_{N\to\infty} \mathbb{P}\left(\|\boldsymbol{x}_N - \mathbf{c}\| > \delta\right) = \lim_{N\to\infty} \mathbb{P}\left(|Q_N - \mathbb{E}\left[Q_N\right]| > \delta\right) \leq \lim_{N\to\infty} \frac{\mathbb{V}\mathrm{ar}\left[Q_N\right]}{\delta^2} = 0$$

implying convergence in probability: $\boldsymbol{x}_N \xrightarrow{p} \mathbf{c}$. □

This result is useful for verifying that in random samples drawn from some random vector $\boldsymbol{x}$ with finite variance $\mathbb{V}\mathrm{ar}\left[\boldsymbol{x}\right] < \infty$, the sample mean $\boldsymbol{x}_N$ converges in probability to the mean of the population, $\mathbb{E}\left[\boldsymbol{x}\right]$.

Example 6.2. Convergence in Probability of the Sample Mean. In a random sample drawn from some random variable X:

$$\lim_{N\to\infty} \mathbb{E}\left[\bar{X}_N\right] = \lim_{N\to\infty} \frac{N}{N}\,\mathbb{E}\left[X\right] = \mathbb{E}\left[X\right]$$

and in addition, if $\mathbb{V}\mathrm{ar}\left[X\right] < \infty$:

$$\lim_{N\to\infty} \mathbb{E}\left[\left(\bar{X}_N - \mathbb{E}\left[\bar{X}_N\right]\right)^2\right] = \lim_{N\to\infty} \mathbb{V}\mathrm{ar}\left[\bar{X}_N\right] = \lim_{N\to\infty} \frac{\mathbb{V}\mathrm{ar}\left[X\right]}{N} = 0$$

and therefore, $\bar{X}_N \xrightarrow{qm} \mathbb{E}\left[X\right]$ which also implies $\bar{X}_N \xrightarrow{p} \mathbb{E}\left[X\right]$. This is easily generalized to a multivariate context: for an N-dimensional random sample drawn from a random vector $\boldsymbol{x}$ with $\mathbb{V}\mathrm{ar}\left[\boldsymbol{x}\right] < \infty$, it holds that:

$$\bar{\boldsymbol{x}}_N = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \xrightarrow{qm} \mathbb{E}\left[\boldsymbol{x}\right]$$

which also implies convergence in probability, $\bar{\boldsymbol{x}}_N \xrightarrow{p} \mathbb{E}\left[\boldsymbol{x}\right]$. ■

The last concept of convergence defined here, "almost sure convergence," is also stronger than convergence in probability, and it is mentioned here for the sake of completeness. This notion is typically employed in the analysis of time series, which deals with "sequences" of observations over time.

Definition 6.6. Almost Sure Convergence. A sequence $\boldsymbol{x}_N$ of random vectors converges *almost surely*, or *with probability one* to a constant vector $\mathbf{c}$ if it holds that:

$$\mathbb{P}\left(\lim_{N\to\infty} \boldsymbol{x}_N = \mathbf{c}\right) = 1$$

where $\lim_{N\to\infty} \boldsymbol{x}_N$ is a random vector. This is also expressed as $\boldsymbol{x}_N \xrightarrow{a.s.} \mathbf{c}$.

With the aid of some measure theory, it is possible to prove the intuitive result that almost sure convergence implies convergence in probability.

All concepts, definitions and results about convergence that have been discussed thus far apply to sequences of random *matrices* as well. A random sequence $\boldsymbol{X}_N$ of matrices converges in probability to some matrix $\mathbf{C}$ if:

$$\lim_{N\to\infty} \mathbb{P}\left(\|\boldsymbol{X}_N - \mathbf{C}\| > \delta\right) = 0$$

(where for any matrix $\mathbf{B}$, $\|\mathbf{B}\| = \sqrt{\mathrm{tr}\left(\mathbf{B}^{\mathrm{T}}\mathbf{B}\right)}$). This is denoted as follows.

$$\boldsymbol{X}_N \xrightarrow{p} \mathbf{C}$$

The following result about convergent random sequences is fundamental to easily derive the asymptotic properties of many statistics and estimators, and it is applied extensively in econometrics. The result is stated in terms of vectorial random sequences, but it applies to matricial ones too.

Theorem 6.4. Continuous Mapping Theorem. *Consider a vectorial random sequence* $\boldsymbol{x}_N \in \mathbb{X}$*, a vector* $\mathbf{c} \in \mathbb{X}$ *with the same length as* $\boldsymbol{x}_N$*, and a vector-valued continuous function* $\boldsymbol{g}(\cdot)$ *with a set of discontinuity points* $\mathbb{D}_{\boldsymbol{g}}$ *such that:*

$$\mathbb{P}(\mathbf{x} \in \mathbb{D}_{\boldsymbol{g}}) = 0$$

(the probability mass at the discontinuities is zero). Then, it holds that:

$$\boldsymbol{x}_N \xrightarrow{p} \mathbf{c} \Rightarrow \boldsymbol{g}(\boldsymbol{x}_N) \xrightarrow{p} \boldsymbol{g}(\mathbf{c})$$
$$\boldsymbol{x}_N \xrightarrow{a.s.} \mathbf{c} \Rightarrow \boldsymbol{g}(\boldsymbol{x}_N) \xrightarrow{a.s.} \boldsymbol{g}(\mathbf{c})$$

that is, convergence in probability and almost sure convergence are preserved when functions are applied to random sequences.

Proof. (Sketched.) Only the case about convergence in probability is proved here, with the purpose of illustrating the core argument (which is essentially an extension of the properties of limits for continuous functions). For a given positive number $\delta > 0$, define the set

$$\mathbb{G}_\delta = \{\mathbf{x} \in \mathbb{X} | \, \mathbf{x} \notin \mathbb{D}_{\boldsymbol{g}} : \exists \mathbf{y} \in \mathbb{X} : \|\mathbf{x} - \mathbf{y}\| < \delta, \|\boldsymbol{g}(\mathbf{x}) - \boldsymbol{g}(\mathbf{y})\| > \varepsilon\}$$

this is the set of points in $\mathbb{X}$ where $\boldsymbol{g}(\cdot)$ "amplifies" the distance with some other point $\mathbf{y}$ beyond a small neighborhood of ε. By this definition:

$$\mathbb{P}(\|\boldsymbol{g}(\boldsymbol{x}_N) - \boldsymbol{g}(\mathbf{c})\| > \varepsilon) \leq \mathbb{P}(\|\boldsymbol{x}_N - \mathbf{c}\| \geq \delta) + \mathbb{P}(\mathbf{c} \in \mathbb{G}_\delta) + \mathbb{P}(\mathbf{c} \in \mathbb{D}_{\boldsymbol{g}})$$

and notice that upon taking the limit of the right-hand side as $N \to \infty$, the second term vanishes by definition of a continuous function, while the third term is zero by hypothesis. Therefore:

$$\lim_{N\to\infty} \mathbb{P}(\|\boldsymbol{g}(\boldsymbol{x}_N) - \boldsymbol{g}(\mathbf{c})\| > \varepsilon) \leq \lim_{N\to\infty} \mathbb{P}(\|\boldsymbol{x}_N - \mathbf{c}\| \geq \delta)$$

which proves the theorem in the case of convergence in probability. □

The importance of this result is that it allows to easily derive the asymptotic properties of many statistical estimators. For example, it is generally not possible to derive the expected value of *some function* $g(\widehat{\mu}_N)$ of a given unbiased estimator $\widehat{\mu}_N$ such that $\mathbb{E}[\widehat{\mu}_N] = \mu_0$ for some μ_0. In fact, as it has been observed in Lecture 1, the best one can do about $\mathbb{E}[g(\widehat{\mu}_N)]$ is to make

approximations based on Jensen's Inequality. However, if $\widehat{\mu}_N$ also converges in probability to μ_0, the continuous mapping theorem ensures that in large samples $g(\widehat{\mu}_N)$ converges in probability to $g(\mu_0)$.

The most frequent applications of the Continuous Mapping Theorem are the simple transformations that are summarized in what follows separately for scalar, vectorial and matricial random sequences.[1]

1. **Scalars.** Given two scalar random sequences $X_N \xrightarrow{p} x$ and $Y_N \xrightarrow{p} y$, the following holds.
$$\begin{aligned}(X_N + Y_N) &\xrightarrow{p} x + y\\ X_N Y_N &\xrightarrow{p} xy\\ X_N / Y_N &\xrightarrow{p} x/y \quad \text{if } y \neq 0\end{aligned}$$

2. **Vectors.** Given two vector random sequences $\boldsymbol{x}_N \xrightarrow{p} \mathbf{x}$ and $\boldsymbol{y}_N \xrightarrow{p} \mathbf{y}$ of equal length, the following holds.
$$\begin{aligned}\boldsymbol{x}_N^{\mathrm{T}} \boldsymbol{y}_N &\xrightarrow{p} \mathbf{x}^{\mathrm{T}}\mathbf{y}\\ \boldsymbol{x}_N \boldsymbol{y}_N^{\mathrm{T}} &\xrightarrow{p} \mathbf{x}\mathbf{y}^{\mathrm{T}}\end{aligned}$$

3. **Matrices.** Given two matrix random sequences $\boldsymbol{X}_N \xrightarrow{p} \mathbf{X}$ and $\boldsymbol{Y}_N \xrightarrow{p} \mathbf{Y}$ of appropriate dimension it holds that:
$$\boldsymbol{X}_N \boldsymbol{Y}_N \xrightarrow{p} \mathbf{X}\mathbf{Y}$$
while for sequences of random full rank square matrices $\boldsymbol{Z}_N \xrightarrow{p} \mathbf{Z}$, it is as follows.
$$\boldsymbol{Z}_N^{-1} \xrightarrow{p} \mathbf{Z}^{-1}$$

4. **Combinations of the Above.** Consider the three random sequences X_N, $\boldsymbol{x}_N$ and $\boldsymbol{X}_N$ above, and suppose that the column dimension of $\boldsymbol{X}_N$ corresponds with the row dimension of $\boldsymbol{x}_N$. Then, the following holds.
$$X_N \boldsymbol{X}_N \boldsymbol{x}_N \xrightarrow{p} x\mathbf{X}\mathbf{x}$$

Clearly, all these properties apply to almost sure convergence as well. It is fairly easy to construct examples about these properties that involve, say, convergent sample means. Once again, comparing these properties to those of expectations sheds an unfavorable light upon the latter: for example, it is difficult to calculate the expectation of a ratio, $\mathbb{E}[X/Y]$, even if both $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ are known quantities!

[1] Observe that the random sequence $\boldsymbol{x}_N$ in the statement of theorem can be arrayed however wished, hence the Theorem equally applies to all desired combinations of scalars, vectors and matrices.

6.2 Laws of Large Numbers

The definitions and results about convergence presented thus far allow to formulate some fundamental results in probability theory, with crucial implications for estimation and statistical analysis. Surely, the most important of these results are the various theorems known as *Laws of Large Numbers*; these posit that in a random sample, any scalar-, vector- or matrix-valued sample mean converges in probability to its corresponding population mean, and it does so under conditions that are more general than the previously discussed result about convergence in quadratic mean. The Laws of Large Numbers are only presented here for the case of vector-valued sample means, of which scalars are a special case and matrices a more general one.

Theorem 6.5. Weak Law of Large Numbers (Khinčin's). *The sample mean* $\bar{\boldsymbol{x}}_N = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i$ *associated with a* random *(i.i.d.) sample drawn from the distribution of a random vector* $\boldsymbol{x}$ *with finite mean* $\mathbb{E}[\boldsymbol{x}] < \infty$ *converges in probability to the population mean of* $\boldsymbol{x}$*.*

$$\bar{\boldsymbol{x}}_N = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \xrightarrow{p} \mathbb{E}[\boldsymbol{x}]$$

Proof. (Sketched.) The analysis is restricted to random vectors $\boldsymbol{x}$ for which the moment-generating function $M_{\boldsymbol{x}}(\mathbf{t})$ is defined. The moment-generating function of the sample mean $\bar{\boldsymbol{x}}_N$ is, *for a given* N:

$$\begin{aligned} M_{\bar{\boldsymbol{x}}_N}(\mathbf{t}) &= \mathbb{E}\left[\exp\left(\mathbf{t}^{\mathrm{T}}\bar{\boldsymbol{x}}_N\right)\right] \\ &= \mathbb{E}\left[\exp\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{t}^{\mathrm{T}}\boldsymbol{x}_i\right)\right] \\ &= \prod_{i=1}^{N}\mathbb{E}\left[\exp\left(\frac{1}{N}\mathbf{t}^{\mathrm{T}}\boldsymbol{x}_i\right)\right] \\ &= \left[M_{\boldsymbol{x}}\left(\frac{1}{N}\mathbf{t}\right)\right]^N \end{aligned}$$

where the third line follows from independence between observations, while the fourth line relies on observations being identically distributed (so that they have the same moment-generating function); essentially, this analysis is an extension of Theorem 3.6. From a Taylor expansion around $\mathbf{t}_0 = \mathbf{0}$:

$$M_{\bar{\boldsymbol{x}}_N}(\mathbf{t}) = \left[1 + \frac{\mathbf{t}^{\mathrm{T}}\,\mathbb{E}[\boldsymbol{x}]}{N} + o\left(\frac{\mathbf{t}^{\mathrm{T}}\boldsymbol{\iota}}{N}\right)\right]^N$$

hence, taking the limit gives the following result.

$$\lim_{N\to\infty} M_{\bar{\boldsymbol{x}}_N}(\mathbf{t}) = \exp\left(\mathbf{t}^{\mathrm{T}}\,\mathbb{E}[\boldsymbol{x}]\right)$$

This is a trivial moment-generating function: that of a *degenerate* discrete random vector where the entire probability mass is concentrated in $\mathbb{E}[\boldsymbol{x}]$! Therefore, exploiting the result that moment-generating functions uniquely characterize their distributions, one can actually conclude that the sample mean converges in probability to its mean as N grows larger. If the random vector $\boldsymbol{x}$ lacks a moment-generating function, one can extend an analogous proof based on the characteristic function $\varphi_{\bar{\boldsymbol{x}}_N}(\mathbf{t})$ of the sample mean; this proof is obviously more complex and the pun here is intended. □

Unlike the result about convergence of sample means in quadratic mean, the Weak law of Large Numbers does not impose finite variances of $\boldsymbol{x}$, and is thus more general. The "stronger" version of the Law of Large Numbers is presented next, but in this case without proof. This version shows that, under stricter conditions, the sample mean approaches the population mean increasingly more closely, without deviating from it to an appreciable extent and with appreciable probability: it "converges with probability one."

Theorem 6.6. Strong Law of Large Numbers (Kolmogorov's). *If in a* random *(i.i.d.) sample drawn from the distribution of some random vector $\boldsymbol{x}$ it simultaneously holds that:* i. $\mathbb{E}[\boldsymbol{x}] < \infty$, ii. $\mathbb{V}\mathrm{ar}[\boldsymbol{x}] < \infty$, *and* iii. $\sum_{n=1}^{\infty} n^{-2}\,\mathbb{V}\mathrm{ar}[\boldsymbol{x}_n] < \infty$, *the sample mean $\bar{\boldsymbol{x}}_N = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i$ converges almost surely to its population mean.*

$$\bar{\boldsymbol{x}}_N = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i \xrightarrow{a.s.} \mathbb{E}[\boldsymbol{x}]$$

An even stronger version allows for independently, but not identically distributed observations (i.n.i.d.).

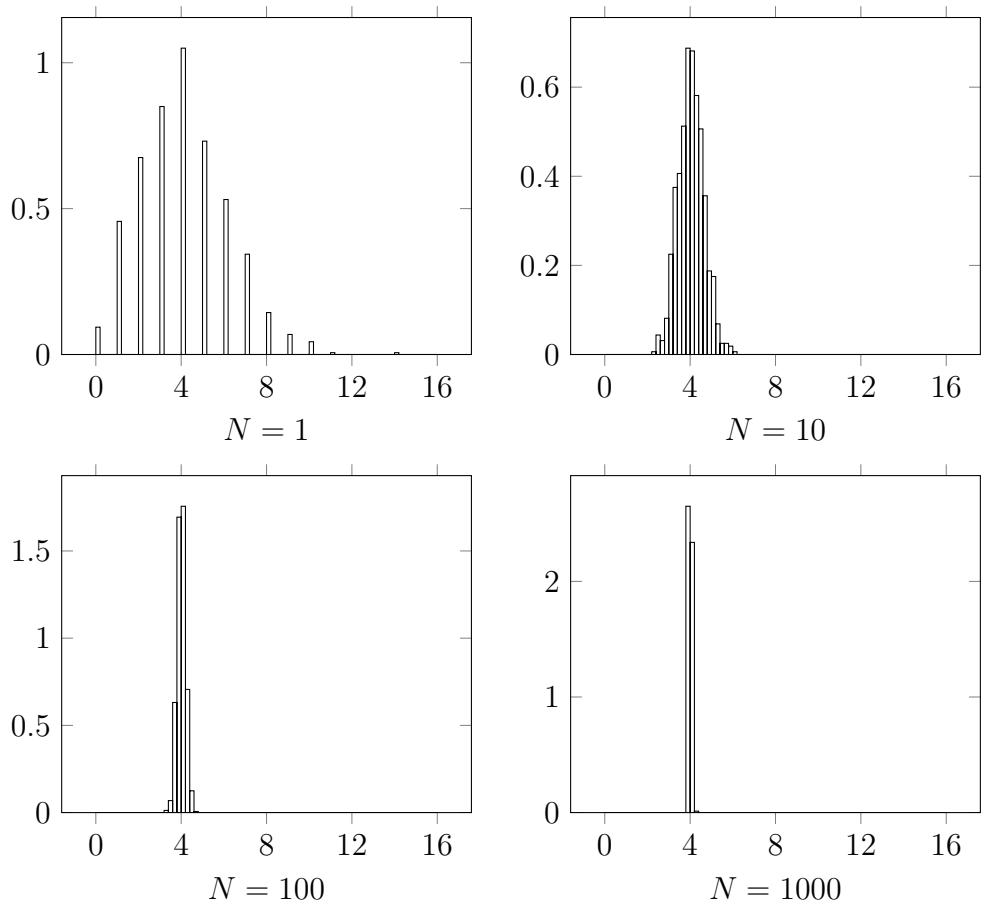
Theorem 6.7. Strong Law of Large Numbers (Markov's). *Consider a non-random sample with independent, non identically distributed observations (i.n.i.d.) where the random vectors $\boldsymbol{x}_i$ that generate it have possibly heterogeneous moments $\mathbb{E}[\boldsymbol{x}_i]$ and $\mathbb{V}\mathrm{ar}[\boldsymbol{x}_i]$. If for some $\delta > 0$ it holds that:*

$$\lim_{N\to\infty}\sum_{i=1}^{\infty}\frac{1}{i^{1+\delta}}\,\mathbb{E}\left[\left|\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i]\right|^{1+\delta}\right] < \infty$$

then the following almost sure convergence result holds.

$$\frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i]\right) \xrightarrow{a.s.} \mathbf{0}$$

Observe that Markov's version of the Strong Law of Large Numbers does not impose finite second moments, but only that the absolute moments of order slightly larger than one, i.e. $1 + \delta > 1$, are finite. This is a seemingly complex, but actually weaker condition (an analogue of which is also used in certain versions of the Central Limit Theorem, as it is discussed later). Other, more general versions of the Law of Large Numbers also allow for **weakly dependent** observations – that is, n.i.n.i.d. samples – which are a prominent feature of socio-economic settings. These results are extensively applied in econometrics, but are not elaborated here. To give more intuition about the working of the Law of Large Numbers, Figure 6.1 below displays the results of multiple simulations about the sample mean calculated from random samples of increasing size drawn from $X \sim \text{Pois}\,(4)$ – see the notes.



Note: histograms of realizations of $\bar{X}_N$ obtained from multiple i.i.d. samples drawn from $X \sim \text{Pois}\,(4)$. Each histogram is obtained with 800 samples of the indicated size N. The realizations of $\bar{X}_N$ are binned on the x-axes with bins of length 0.02. For all histograms, the y-axes measure the density of their bins.

Figure 6.1: Simulation of the Law of Large Numbers for $X \sim \text{Pois}\,(4)$

It is useful to exemplify how the Laws of Large Numbers can be extended to estimators that can be expressed as functions of sample means. To this end, a first definition is in order.

Definition 6.7. Consistent Estimators. An estimator $\widehat{\boldsymbol{\theta}}_N$ is *consistent* if it converges in probability to the *true* population parameters $\boldsymbol{\theta}_0$ which it is meant to estimate.

$$\widehat{\boldsymbol{\theta}}_N \xrightarrow{p} \boldsymbol{\theta}_0$$

Here the subscript "0" is used again to denote the true value of the parameter of interest. This is a standard convention in asymptotic analysis.

Example 6.3. Consistency of the linear regression estimators. Consider the bivariate linear regression model from Example 3.11 and the subsequent references. Suppose that a researcher has access to a random sample drawn from (X_i, Y_i). The Method of Moments (MM) estimator of the *true* slope parameter β_1 is defined – see Example 5.4 – as:

$$\widehat{\beta}_{1,MM} = \frac{\sum_{i=1}^{N} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2}$$

where $\bar{X} = N^{-1}\sum_{i=1}^{N} X_i$ and $\bar{Y} = N^{-1}\sum_{i=1}^{N} Y_i$. This estimator can also be obtained via Maximum Likelihood under certain assumptions.

Observe that this estimator is defined as the ratio between what is defined as the "sample covariance" between X_i and Y_i, and the sample variance of X_i. These sample statistics are obvious extensions of the sample mean; by the Weak Law of Large Numbers, their probability limits are:

$$\frac{1}{N}\sum_{i=1}^{N} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) \xrightarrow{p} \mathbb{C}\text{ov}\left[X_i, Y_i\right]$$

$$\frac{1}{N}\sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2 \xrightarrow{p} \mathbb{V}\text{ar}\left[X_i\right]$$

that is, the corresponding *population* moments. Therefore, by the properties of probability limits that derive from the Continuous Mapping Theorem, it follows that the MM estimator of the regression slope is consistent.

$$\widehat{\beta}_{1,MM} \xrightarrow{p} \beta_1$$

An extension of this analysis shows that the MM estimator of the regression constant β_0:

$$\widehat{\beta}_{0,MM} = \bar{Y} - \widehat{\beta}_{1,MM} \cdot \bar{X}$$

is also consistent; by the Continuous Mapping Theorem, $\widehat{\beta}_{0,MM} \xrightarrow{p} \beta_0$. ■

One can show that if the assumptions that motivate Method of Moments or Maximum Likelihood estimators are correct, these are consistent. This is shown next by some "heuristic" (i.e. intuitive, not too rigorous) proofs.

Theorem 6.8. Consistency of the Method of Moments. *An estimator $\widehat{\boldsymbol{\theta}}_{MM}$ defined as the solution of a set of sample moments (5.2) is consistent for the parameter set $\boldsymbol{\theta}_0$ that solves the corresponding population moments (5.1), if such a solution exists (i.e. if the estimation problem is well defined).*

Proof. (Heuristic.) By some applicable Law of Large Numbers:

$$\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{m}\left(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{MM}\right)\xrightarrow{p}\mathbb{E}\left[\boldsymbol{m}\left(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{MM}\right)\right]=\mathbf{0}$$

where the equality to $\mathbf{0}$ is given by the definition of a Method of Moments estimator, which is maintained throughout the sequence as $N\to\infty$. Since by hypothesis the zero moment conditions have only one admissible solution, at the probability limit it is $\operatorname{plim}\widehat{\boldsymbol{\theta}}_{MM}=\boldsymbol{\theta}_0$. □

For Maximum Likelihood, an analysis that is as general as in the Method of Moments case above would hardly be simple. Thus, the (heuristic) proof is given here for random samples only. Extensions of this result are possible, and they apply more generalized versions of the Law of Large Numbers.

Theorem 6.9. Consistency of Maximum Likelihood Estimators. *In a random sample, an estimator $\widehat{\boldsymbol{\theta}}_{MLE}$ which is defined as the maximizer of a log-likelihood function as per (5.19) is consistent for the parameter set $\boldsymbol{\theta}_0$ that maximizes the corresponding population moment function.*

$$\boldsymbol{\theta}_0=\arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\mathbb{E}\left[\log f_{\boldsymbol{x}}\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right]$$

If such a maximum exists, by the likelihood principle it corresponds to the true parameter of the distribution under analysis.

Proof. (Heuristic.) By the Weak Law of Large Numbers, for any $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ including $\widehat{\boldsymbol{\theta}}_{MLE}$ and $\boldsymbol{\theta}_0$:

$$\frac{1}{N}\sum_{i=1}^{N}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\boldsymbol{\theta}\right)\xrightarrow{p}\mathbb{E}\left[\log f_{\boldsymbol{x}}\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right]$$

moreover, by the definition of MLE the following holds for all $N\in\mathbb{N}$.

$$\frac{1}{N}\sum_{i=1}^{N}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{MLE}\right)\geq\frac{1}{N}\sum_{i=1}^{N}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\xrightarrow{p}\mathbb{E}\left[\log f_{\boldsymbol{x}}\left(\boldsymbol{x};\boldsymbol{\theta}_0\right)\right]$$

Consequently, given that $\boldsymbol{\theta}_0$ maximizes the expected log-density or log-mass function in the population:

$$\lim_{N\to\infty} \mathbb{P}\left(\mathbb{E}\left[\log f_{\boldsymbol{x}}\left(\boldsymbol{x};\boldsymbol{\theta}_0\right)\right] \geq \mathbb{E}\left[\log f_{\boldsymbol{x}}\left(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_{MLE}\right)\right]\right) = 1$$

all these facts can be reconciled only if, *at the limit*:

$$\frac{1}{N}\sum_{i=1}^{N} \log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{MLE}\right) \xrightarrow{p} \mathbb{E}\left[\log f_{\boldsymbol{x}}\left(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_{MLE}\right)\right] = \mathbb{E}\left[\log f_{\boldsymbol{x}}\left(\boldsymbol{x};\boldsymbol{\theta}_0\right)\right]$$

and *at the limit*, $\operatorname{plim} \widehat{\boldsymbol{\theta}}_{MLE} = \boldsymbol{\theta}_0$ by the Continuous Mapping Theorem. □

6.3 Convergence in Distribution

So far, all concepts of convergence that have been discussed have involved a random sequence converging to some constant $\mathbf{c}$ (or a matrix $\mathbf{C}$). However, one can extend the concept of convergence to a random sequence converging to *another random vector* $\boldsymbol{x}$ (or another random matrix $\boldsymbol{X}$). For example, writing the statement

$$\boldsymbol{x}_N \xrightarrow{p} \boldsymbol{x}$$

is equivalent to saying that the random sequence $\boldsymbol{x}_N$ converges in probability to the random vector $\boldsymbol{x}$ in the following sense.

$$\lim_{N\to\infty} \mathbb{P}\left(\|\boldsymbol{x}_N - \boldsymbol{x}\| > \delta\right) = 0$$

Intuitively, in the limit the probabilistic behavior of $\boldsymbol{x}_N$ "becomes similar" to that of $\boldsymbol{x}$ up to δ.[2] A relevant question is: "How similar?" – that is, do all moments and the distribution of $\boldsymbol{x}_N$ converge to that of $\boldsymbol{x}$? In general, the answer to this kind of questions is "No" unless one also invokes the concept of *convergence in distribution*.[3]

[2] Parallel notions exist for r-th moment and almost sure convergence.

[3] One can show – but it is intuitive – that if $\boldsymbol{x}_N$ converges in r-th mean to $\boldsymbol{x}$, that is

$$\lim_{N\to\infty} \mathbb{E}\left[\|\boldsymbol{x}_N - \boldsymbol{x}\|^r\right] = 0$$

then the r-th moment of $\boldsymbol{x}$ (and by extension all lower moments) converge to those of $\boldsymbol{x}$:

$$\lim_{N\to\infty} \mathbb{E}\left[|\boldsymbol{x}_N|^r\right] \to \mathbb{E}\left[|\boldsymbol{x}|^r\right] < \infty$$

so long as they are finite. However, this is not enough to guarantee that higher moments, not to mention the distribution function itself, converge to those of $\boldsymbol{x}$.

Definition 6.8. Convergence in Distribution. Consider a sequence of random vectors $\boldsymbol{x}_N$, whose each element has a cumulative distribution function $F_{\boldsymbol{x}_N}(\mathbf{x}_N)$, as well as a random vector $\boldsymbol{x}$ with cumulative distribution function $F_{\boldsymbol{x}}(\mathbf{x})$. The random sequence $\boldsymbol{x}_N$ is said to converge *in distribution* to $\boldsymbol{x}$ if:

$$\lim_{N\to\infty} |F_{\boldsymbol{x}_N}(\mathbf{x}_N) - F_{\boldsymbol{x}}(\mathbf{x})| = 0$$

at all *continuity* points $\mathbf{x} \in \mathbb{X}$ in the support of $\boldsymbol{x}$. This is usually expressed with the following formalism.

$$\boldsymbol{x}_N \xrightarrow{d} \boldsymbol{x}$$

Definition 6.9. Limiting Distribution. If $\boldsymbol{x}_N \xrightarrow{d} \boldsymbol{x}$, that is some random sequence $\boldsymbol{x}_N$ converges in distribution to a random vector $\boldsymbol{x}$, then $F_{\boldsymbol{x}}(\mathbf{x})$ is said to be the *limiting* distribution of $\boldsymbol{x}_N$.

Intuitively, convergence in distribution requires that the probability density of $\boldsymbol{x}_N$ tends to become identical to that of $\boldsymbol{x}$ everywhere in the support of the random vectors in question. This is a stronger condition than convergence in probability to a random vector, which more simply requires that the random sequence $\boldsymbol{x}_N$ and its limit $\boldsymbol{x}$ are "very likely to produce close observations" as the index of the sequence grows very large.

Some important examples of convergence in distribution are intimately related to certain relationships between probability distributions, where one distribution is identified as the "limit case" of another distribution when some parameter that characterizes the latter tends to a specific limit value. These relationships not only serve as excellent illustrations of convergence in distribution, but are also important by themselves. They are illustrated as follows in the form of *observations* about common probability distributions.

Observation 6.1. Asymptotics of Student's t-distribution. Consider a random variable that follows the Student's t-distribution with parameter ν, $X \sim \mathcal{T}(\nu)$. As $\nu \to \infty$, the probability distribution of X tends to that of the standard normal distribution, i.e. $\lim_{\nu\to\infty} X = Z \sim \mathcal{N}(0,1)$.

Proof. Taking the limit of the probability density function of the Student's t-distribution as $\nu \to \infty$:

$$\lim_{\nu\to\infty} \frac{1}{\mathrm{B}\left(\frac{1}{2},\frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu}} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

as $\lim_{\nu\to\infty} \sqrt{\nu}\,\mathrm{B}\left(\frac{1}{2},\frac{\nu}{2}\right) = \sqrt{2\pi}$ by the properties of the Beta function; while by more standard arguments, $\lim_{\nu\to\infty} (1 + x^2/\nu)^{-(\nu+1)/2} = \exp(-x^2/2)$. □

This observation substantiates the claim, already put forward in Lecture 2 that the Student's t-distribution becomes increasingly more similar to the standard normal distribution as ν increases. It is useful to report again the graphical intuition, similarly to Figure 2.12. In Figure 6.2 below, however, the result is instead represented in terms of *cumulative* distributions.

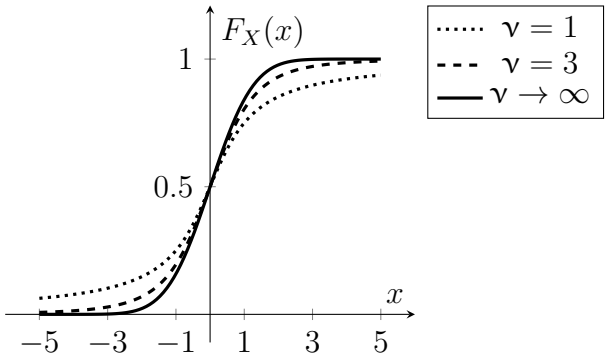


Figure 6.2: Convergence of the Student's t-distribution as $\nu \to \infty$

Example 6.4. Convergence in distribution of the t-statistic. Consider a random sample drawn from some normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$. By the arguments advanced in Lecture 4, the random sequence of t-statistics:

$$\mathrm{t}_N = \sqrt{N}\,\frac{\bar{X}_N - \mu}{S_N} \sim \mathcal{T}_{(N-1)}$$

follows the Student's t-distribution with degrees of freedom given by $N-1$. Therefore, by Observation 6.1 it follows that:

$$\mathrm{t}_N \xrightarrow{d} \mathcal{N}(0,1)$$

that is, the sequence t_N converges in distribution to a random variable that follows the standard normal distribution (note in the expression above that the notation of the limiting distribution is represented on the right-hand side, this is conventional). Again, the intuition is developed in Figure 6.2. An implication of this result is that all tests of hypotheses – as well as all interval estimators – that are based on the Student's t-distribution become increasingly similar, as the sample size increases, to those that are based on the standard normal distribution. In fact, the difference becomes negligible already for $N > 20$, which motivates the ubiquitous use of the critical values derived from the standard normal in applied statistical analysis. ■

Observation 6.2. Asymptotics of Snedecor's F-distribution. Consider a random variable that follows Snedecor's F-distribution with parameters ν_1 and ν_2, $X \sim \mathcal{F}(\nu_1, \nu_2)$. As $\nu_2 \to \infty$, the probability distribution of $W = \nu_1 X$ tends to that of a chi-squared distribution with parameter ν_1, i.e. $\lim_{\nu_2 \to \infty} \nu_1 X = W \sim \chi^2(\nu_1)$.

Proof. It is easy to derive the probability density function $f_W(w)$ of the transformation $W = \nu_1 X$. Taking its limit as $\nu_2 \to \infty$ gives:

$$\begin{aligned}
\lim_{\nu_2 \to \infty} f_W(w) &= \lim_{\nu_2 \to \infty} \frac{1}{\mathrm{B}\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \left(\frac{1}{\nu_2}\right)^{\frac{\nu_1}{2}} w^{\frac{\nu_1}{2}-1} \left(1 + \frac{w}{\nu_2}\right)^{-\frac{\nu_1+\nu_2}{2}} \\
&= \lim_{\nu_2 \to \infty} \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_2}{2}\right)} \frac{1}{\Gamma\left(\frac{\nu_1}{2}\right)} \left(\frac{1}{\nu_2 + w}\right)^{\frac{\nu_1}{2}} w^{\frac{\nu_1}{2}-1} \left(1 + \frac{w}{\nu_2}\right)^{-\frac{\nu_2}{2}} \\
&= \frac{1}{\Gamma\left(\frac{\nu_1}{2}\right) \cdot 2^{\frac{\nu_1}{2}}} w^{\frac{\nu_1}{2}-1} \exp\left(-\frac{w}{2}\right)
\end{aligned}$$

where:

$$\lim_{\nu_2 \to \infty} \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{1}{\nu_2 + w}\right)^{\frac{\nu_1}{2}} = 2^{-\frac{\nu_1}{2}}$$

follows by the properties of the Gamma function. □

Example 6.5. Convergence in distribution of Hotelling's t-squared statistic. Recall the formulation of Hotelling's *rescaled* t-squared statistic for a given K, and express it as a random sequence.

$$\frac{N-K}{K(N-1)} \mathrm{t}_N^2 = \frac{(N-K)N}{K(N-1)} (\bar{\boldsymbol{x}} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{S}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \sim \mathcal{F}_{K,N-K}$$

For a given N, this statistic follows the F-distribution with paired degrees of freedom K and $N-K$. By Observation 6.2, however, as the sample size N grows large one obtains the following result.

$$\mathrm{t}_N^2 \xrightarrow{d} \chi_K^2$$

In words, Hotelling's t-squared statistic (non-rescaled) converges in distribution to a chi-squared distribution with K degrees of freedom.[4] Similarly as in the univariate case, this has important implications for multivariate tests of hypothesis, say about the means of a multivariate normal distribution. As the sample grows large, these can be based on the relatively simple chi-squared distribution, instead of the more involved F-distribution. ■

[4]The rescaling factor is removed for two reasons. First, to apply Observation 6.2 one should multiply the sequence by $\nu_1 = K$. Second, as $N \to \infty$ the term $(N-K)/(N-1)$ becomes irrelevant, and the asymptotic result holds irrespectively of it.

Observation 6.3. Asymptotics of the Gamma distribution. Consider a random variable that follows the Gamma distribution with parameters α and β, $X \sim \Gamma(\alpha, \beta)$. Let $\mu = \alpha/\beta$ and $\sigma^2 = \alpha/\beta^2$. As $\alpha \to \infty$, the probability distribution of X tends to that of a normal distribution with parameters μ and σ^2, i.e. $\lim_{\alpha\to\infty} X \sim \mathcal{N}(\mu, \sigma^2)$.

Proof. Unlike the previous two observations, this one can be proved via the moment generating function (which both the Student's t-distribution and the F-distribution lack) and more handily so. Define the standardized random variable $Z = (X - \mu)/\sigma = \left(\beta/\sqrt{\alpha}\right) X - \sqrt{\alpha}$; by the properties of moment generating functions, and recalling (2.88), it is:

$$M_Z(t) = \exp\left(-\sqrt{\alpha}t\right) \cdot M_X\left(\frac{\beta}{\sqrt{\alpha}}t\right) = \exp\left(-\sqrt{\alpha}t\right)\left(1 - \frac{t}{\sqrt{\alpha}}\right)^{-\alpha}$$

and after some manipulation, the limit as $\alpha \to \infty$ gives:

$$\lim_{\alpha\to\infty} M_Z(t) = \lim_{\alpha\to\infty} \exp\left(-\sqrt{\alpha}t\right)\left(1 - \frac{t}{\sqrt{\alpha}}\right)^{-\alpha} = \exp\left(\frac{t^2}{2}\right)$$

showing that *at the limit*, $Z \sim \mathcal{N}(0, 1)$ and therefore $X \sim \mathcal{N}(\mu, \sigma^2)$. □

Example 6.6. Convergence in distribution of the sample mean $\bar{X}$ drawn from the exponential distribution. As elaborated in previous lectures, in a random sample drawn from a random variable $X \sim \text{Exp}(\lambda)$ the sample mean follows the Gamma distribution, $\bar{X} \sim \Gamma(N, N/\lambda)$. Thus, by Observation 6.3:

$$\sqrt{N}\left(\bar{X}_N - \lambda\right) \xrightarrow{d} \mathcal{N}\left(0, \lambda^2\right)$$

a statement interpreted in the sense that *for a fixed value of* N:

$$\bar{X}_N \overset{A}{\sim} \mathcal{N}\left(\lambda, \frac{\lambda^2}{N}\right)$$

where A stands for "approximation" (by definition, the sequence index N cannot appear in the formulation of the limiting distribution, therefore the former expression is a more rigorous characterization of convergence in distribution). This result allows to use of the normal distribution in statistical tests about the exponential distribution. As it is discussed later at length, this result is more general and goes by the name of Central Limit Theorem; what is interesting about the exponential setting is that the *exact* distribution of the sample mean is known to be the Gamma, and by Observation 6.3, this fact is reconciled with the Central Limit Theorem. ■

The following results about convergence in distribution are necessary to derive the asymptotic properties of many econometric estimators.

Theorem 6.10. Continuous Mapping Theorem (continued). *Under the hypotheses of Theorem 6.4:*

$$\boldsymbol{x}_N \xrightarrow{d} \boldsymbol{x} \Rightarrow \boldsymbol{g}(\boldsymbol{x}_N) \xrightarrow{d} \boldsymbol{g}(\boldsymbol{x})$$

that is, a random sequence which is obtained from the application of a transformation $\boldsymbol{g}(\cdot)$ to some original random sequence $\boldsymbol{x}_N$, converges in distribution to the distribution resulting from applying the transformation $\boldsymbol{g}(\cdot)$ to the random vector $\boldsymbol{x}$ associated with the limiting distribution of $\boldsymbol{x}_N$.

The proof of this statement is omitted as it involves some advanced measure theory. The continuous mapping theorem for convergence in distribution is an important result, as it allows to prove the following properties of random sequences which are heavily exploited in statistics and econometrics.

Theorem 6.11. Slutskij's Theorem. *Consider any two (scalar) random sequences X_N and Y_N such that:*

$$X_N \xrightarrow{d} X$$
$$Y_N \xrightarrow{p} c$$

that is, X_N converges in distribution *to that of the random variable X, while Y_N converges* in probability *to a constant c. Then, the following holds.*

$$(X_N + Y_N) \xrightarrow{d} X + c$$
$$X_N Y_N \xrightarrow{d} cX$$
$$X_N / Y_N \xrightarrow{d} X/c \quad \text{if } c \neq 0$$

Proof. It is enough to recognize that, as $Y_N \xrightarrow{p} c$, then Y_N has a degenerate limiting distribution, and thus the (vector) random sequence (X_N, Y_N) converges in distribution to that of the random vector (X, c). Then the results above follow from the application of the Continuous Mapping Theorem to three given continuous functions of X_N and Y_N. □

Corollary. Cramér-Wold Device. *Given a random sequence $\boldsymbol{x}_N$ and a constant vector $\mathbf{a}$ of the same dimension:*

$$\boldsymbol{x}_N \xrightarrow{d} \boldsymbol{x} \Rightarrow \mathbf{a}^{\mathrm{T}} \boldsymbol{x}_N \xrightarrow{d} \mathbf{a}^{\mathrm{T}} \boldsymbol{x}$$

that is, if a vectorial random sequence has a limiting distribution, any linear combination of its elements will converge in distribution to the distribution of the corresponding "limiting" linear combination.

Before moving to the extensive treatment of the Central Limit Theorem, this section is concluded with the analysis of another important result about convergence in distribution. This result is foundational for an entire branch of statistics called **Extreme Value Theory**, which concerns the analysis of extreme order statistics (maxima and minima).

Theorem 6.12. Extreme Value Theorem. *This result is also called the* **Fisher-Tippett-Gnedenko Theorem** *by the name of its discoverers. It states that given a random (i.i.d.) sample* $(X_1, \ldots, X_N)$*, if a convergence in distribution result of the kind*

$$\frac{X_{(N)} - b_N}{a_N} \xrightarrow{d} W$$

can be established – where $X_{(N)}$ *is the maximum order statistic while* $a_N > 0$ *and* b_N *are sequences of real constants – then:*

$$W \sim \text{GEV}\left(0, 1, \xi\right)$$

for some real ξ*; that is, the limiting distribution of the normalized maximum is some standardized type of the Generalized Extreme Value distribution.*

Proof. (Outline.) The objective of the proof is to show that, given a random variable X from which the random sample is drawn, for all the points $x \in \mathbb{X}$ in its support where the distribution $F_X\left(x\right)$ is continuous:

$$\lim_{N\to\infty} \left[F_X\left(a_N x - b_N\right)\right]^N = \exp\left(-\left(1 + \xi x\right)^{-\frac{1}{\xi}}\right)$$

where the left-hand side is the limit of the cumulative distribution of the standardized maximum, while the right-hand side is the expression of the cumulative standardized GEV distribution. By taking the the logarithm of this expression, the above is:

$$\lim_{N\to\infty} N \log F_X\left(a_N x - b_N\right) = -\left(1 + \xi x\right)^{\frac{1}{\xi}}$$

showing that $F_X\left(a_N x - b_N\right) \to 1$ as $N \to \infty$. Since $-\log\left(x\right) \approx 1 - x$ for any x is close to 1, the above expression approximates the following.

$$\lim_{N\to\infty} \frac{1}{N\left[1 - F_X\left(a_N x - b_N\right)\right]} = \frac{1}{\left(1 + \xi x\right)^{\frac{1}{\xi}}}$$

The rest of the proof is mathematically involved, and it proceeds to *i.* show that the right-hand side of the above expression on is the only admissible limit and *ii.* establish conditions under which $\xi = 0$ (Type I GEV, Gumbel), $\xi > 0$ (Type II GEV, Fréchet) and $\xi < 0$ (Type III GEV, reverse Weibull), where $\xi = 0$ is interpreted as a limit case (see Lecture 2). □

While the Extreme Value Theory is outside the scope of this discussion, it is worth to briefly comment on some implications of the Fisher-Tippett-Gnedenko Theorem.

1. First, the Theorem does not state that a standardized maximum *always* converge to a GEV distribution; it states that *if* it converges, the limiting distribution is GEV. In this respect, the Theorem differs from other results such as the Central Limit Theorem.

2. The implications of this result are not restricted to the maximum, but extend to the minimum too. By defining $Y = -X$, for every N it clearly is $Y_{(1)} = -X_{(N)}$, which helps identify the distribution of the minimum if that of the maximum is known (think for instance about the relationship between the reverse Weibull and the "traditional" Weibull distribution). This explains why the name of the theorem references "extreme values" and not just maxima.

3. As mentioned, the proof of the Theorem sets conditions that allow to identify which Type of GEV distribution is a possible limiting distribution of the maximum, by inspecting the cumulative distribution $F_X(x)$ that generates the data. These conditions are quite technical, but some of their implications are quite useful. For example, it is known that the limiting distribution of the maximum associated with a random sample drawn from the normal distribution is the Gumbel distribution.

In econometrics, the Extreme Value Theorem is invoked as the motivation behind specific assumptions made in certain models of decision-making, where the random component of choice is assumed to follow a GEV distribution. In fact, a GEV distribution is a natural choice to model the maximum value between multiple options that are considered by a decision-maker.

6.4 Central Limit Theorems

Convergence in distribution has little practical content if one does not know, or is not able to derive, the limiting distribution of some random sequence of interest. Nevertheless, in a specific but fundamental situation the limiting distribution is typically known: it is the case of generalized sample means. In fact, thanks to a set of results known as the *Central Limit Theorems*, it is possible to establish that the limiting distribution of a sample mean is a normal distribution, *regardless* of the original distribution from which the data are originated. It is because of this last point that these results are so fundamental in statistics and econometrics.

Theorem 6.13. Central Limit Theorem (Lindeberg and Lévy's). *The sample mean* $\bar{\boldsymbol{x}}_N = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i$ *associated with a* random *(i.i.d.) sample drawn from the distribution of a random vector* $\boldsymbol{x}$ *with mean and variance that are both finite:* $\mathbb{E}[\boldsymbol{x}] < \infty$ *and* $\mathbb{V}\mathrm{ar}[\boldsymbol{x}] < \infty$*, is such that the random sequence defined as a centered sample mean multiplied by* $\sqrt{N}$ *converges in distribution to a multivariate normal distribution.*

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}]\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbb{V}\mathrm{ar}[\boldsymbol{x}])$$

Proof. (Sketched.) Consider the *standardized* random vector:

$$\boldsymbol{z} = [\mathbb{V}\mathrm{ar}[\boldsymbol{x}]]^{-\frac{1}{2}}(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])$$

where the matrix $[\mathbb{V}\mathrm{ar}[\boldsymbol{x}]]^{-\frac{1}{2}}$ and its inverse $[\mathbb{V}\mathrm{ar}[\boldsymbol{x}]]^{\frac{1}{2}}$ satisfy the following.

$$[\mathbb{V}\mathrm{ar}[\boldsymbol{x}]]^{-\frac{1}{2}}\mathbb{V}\mathrm{ar}[\boldsymbol{x}][\mathbb{V}\mathrm{ar}[\boldsymbol{x}]]^{-\frac{1}{2}} = \mathbf{I}$$
$$[\mathbb{V}\mathrm{ar}[\boldsymbol{x}]]^{\frac{1}{2}}[\mathbb{V}\mathrm{ar}[\boldsymbol{x}]]^{\frac{1}{2}} = \mathbb{V}\mathrm{ar}[\boldsymbol{x}]$$

Such a matrix can always be constructed because variance-covariance matrices are positive semi-definite. The objective of the proof is to show that:

$$\bar{\bar{\boldsymbol{z}}}_N \equiv \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{z}_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I})$$

that is, the random sequence $\bar{\bar{\boldsymbol{z}}}_N$ defined above converges in distribution to a *standard* multivariate normal distribution. If this is true, the main result also follows by the linear properties of the multivariate normal distribution after recognizing the following relationship between random sequences.

$$\sqrt{N}(\bar{\boldsymbol{x}}_N - \mathbb{E}[\boldsymbol{x}]) = \sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}]\right) = [\mathbb{V}\mathrm{ar}[\boldsymbol{x}]]^{\frac{1}{2}}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{z}_i\right)$$

To show this, suppose that a moment-generating function of $\boldsymbol{z}$ exists. If so, one can express the moment generating function of $\bar{\bar{\boldsymbol{z}}}_N$, for fixed N, as:

$$\begin{aligned}
M_{\bar{\bar{\boldsymbol{z}}}_N}(\mathbf{t}) &= \mathbb{E}\left[\exp\left(\mathbf{t}^{\mathrm{T}}\bar{\bar{\boldsymbol{z}}}_N\right)\right] \\
&= \mathbb{E}\left[\exp\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{t}^{\mathrm{T}}\boldsymbol{z}_i\right)\right] \\
&= \prod_{i=1}^{N}\mathbb{E}\left[\exp\left(\frac{1}{\sqrt{N}}\mathbf{t}^{\mathrm{T}}\boldsymbol{z}\right)\right] \\
&= \left[M_{\boldsymbol{z}}\left(\frac{1}{\sqrt{N}}\mathbf{t}\right)\right]^{N}
\end{aligned}$$

by a derivation analogous to the one in the proof of the Weak Law of Large Numbers (Theorem 6.5). As in that proof, apply a Taylor expansion of the above expression around $\mathbf{t}_0 = \mathbf{0}$, but account for the second order element:

$$\begin{aligned} M_{\bar{\bar{\boldsymbol{z}}}_N}(\mathbf{t}) &= \left[1 + \frac{\mathbf{t}^{\mathrm{T}}\,\mathbb{E}[\boldsymbol{z}]}{\sqrt{N}} + \frac{\mathbf{t}^{\mathrm{T}}\,\mathbb{E}\left[\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\right]\mathbf{t}}{2N} + o\left(\frac{\mathbf{t}^{\mathrm{T}}\mathbf{t}}{2N}\right)\right]^N \\ &= \left[1 + \frac{\mathbf{t}^{\mathrm{T}}\mathbf{t}}{2N} + o\left(\frac{\mathbf{t}^{\mathrm{T}}\mathbf{t}}{2N}\right)\right]^N \end{aligned}$$

where the second line exploits the fact that $\mathbb{E}[\boldsymbol{z}] = \mathbf{0}$ and that $\mathbb{E}\left[\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\right] = \mathbf{I}$ by construction of $\boldsymbol{z}$. Clearly, taking the limit of the above expression for $N \to \infty$ gives:

$$\lim_{N\to\infty} M_{\bar{\bar{\boldsymbol{z}}}_N}(\mathbf{t}) = \exp\left(\frac{\mathbf{t}^{\mathrm{T}}\mathbf{t}}{2}\right)$$

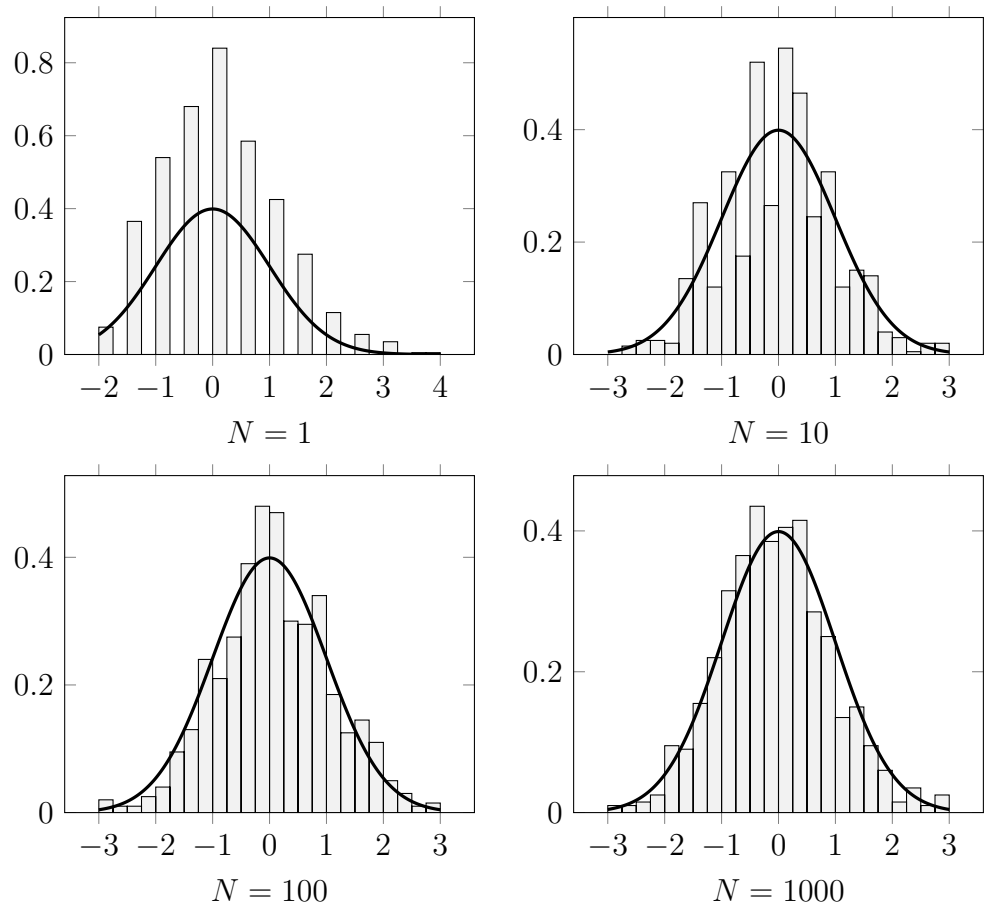
and this is nothing else but the moment-generating function of the standard multivariate normal, as it is postulated. Should $\boldsymbol{z}$ lack a moment-generating function, a similar derivation that leverages upon the characteristic function $\varphi_{\bar{\bar{\boldsymbol{z}}}_N}(\mathbf{t})$ applies instead. □

How is a Central Limit Theorem actually useful in practice? The result is to be interpreted in the sense that *for some specific value of* N, the sample mean is "approximately" normally distributed with a variance-covariance which is *decreasing in the sample size*:

$$\bar{\boldsymbol{x}}_N = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i \overset{A}{\sim} \mathcal{N}\left(\mathbb{E}[\boldsymbol{x}], \frac{1}{N}\,\mathbb{V}\mathrm{ar}[\boldsymbol{x}]\right)$$

where the notation $\overset{A}{\sim}$ indicates that the normal distribution in question, called the **asymptotic distribution**, is approximate and is valid for a fixed N, instead of being a "limiting" distribution (recall from the discussion of Example 6.6 that a limiting distribution cannot be expressed in terms of N). To illustrate, Figure 6.3 plots the empirical distribution of the *standardized* sample means obtained with the same simulation of samples drawn from the Poisson distribution as in Figure 6.1. The standardization implies that the limiting distribution is the standard normal.[5] Despite the complications entailed in the representation of a distribution via histograms, the simulation highlights how the limiting distribution is approximated increasingly better as the sample size increases.

[5]For intuition, it is as if the univariate version of the sequence $\bar{\bar{\boldsymbol{z}}}_N$, call it say $\bar{\bar{Z}}_N$, is plotted in Figure 6.1.



Note: histograms of realizations of $\sqrt{N}\left(\bar{X}_N - 4\right)/2$ obtained from multiple i.i.d. samples drawn from $X \sim \text{Pois}\,(4)$. Each histogram is obtained with 800 samples of the indicated size N. All the realizations are binned on the x-axes with bins of length 0.25. For all histograms, the y-axes measure the density of their bins. Density functions of the standard normal distribution are superimposed upon each histogram.

Figure 6.3: Simulation of the Central Limit Theorem for $X \sim \text{Pois}\,(4)$

Some more general formulations of the (multivariate) Central Limit Theorem follow next. Their proofs, however, are not presented. These versions are especially important as they extend the main result to samples whose observations are possibly *not* identically distributed (i.n.i.d.). As such, it is useful to familiarize with their statements and hypotheses since the asymptotic properties of many estimators, especially in econometrics, are based on them. As in the case of the Laws of Large Numbers, more Central Limit Theorems exist which additionally allow *dependent observations* (n.i.n.i.d. samples), but these are outside the scope of this discussion. In econometrics, these are invoked in order to motivate the use of covariance estimators such as the cluster-robust and HAC estimators.

Theorem 6.14. Central Limit Theorem (Lindeberg and Feller's). *Consider a non-random (i.n.i.d.) sample where the random vectors* $\boldsymbol{x}_i$ *that generate it have possibly heterogeneous* finite *means* $\mathbb{E}[\boldsymbol{x}_i] < \infty$*, variances* $\mathbb{V}\mathrm{ar}[\boldsymbol{x}_i] < \infty$*, and all mixed third moments are finite too. If:*

$$\lim_{N\to\infty} \left(\sum_{i=1}^{N} \mathbb{V}\mathrm{ar}[\boldsymbol{x}_i] \right)^{-1} \mathbb{V}\mathrm{ar}[\boldsymbol{x}_i] = \mathbf{0}$$

then it holds that:

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i]) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbb{V}\mathrm{ar}[\boldsymbol{x}])$$

where $\frac{1}{N}\sum_{i=1}^{N} \mathbb{V}\mathrm{ar}[\boldsymbol{x}_i] \xrightarrow{p} \mathbb{V}\mathrm{ar}[\boldsymbol{x}]$*, that is, the positive semi-definite matrix* $\mathbb{V}\mathrm{ar}[\boldsymbol{x}]$ *is the probability limit of the observations' variances.*

Theorem 6.15. Central Limit Theorem (Ljapunov's). *Consider a non-random (i.n.i.d.) sample where the random vectors* $\boldsymbol{x}_i$ *that generate it have possibly heterogeneous* finite *moments* $\mathbb{E}[\boldsymbol{x}_i] < \infty$ *and* $\mathbb{V}\mathrm{ar}[\boldsymbol{x}_i] < \infty$*. If:*

$$\lim_{N\to\infty} \left(\sum_{i=1}^{N} \mathbb{V}\mathrm{ar}[\boldsymbol{x}_i] \right)^{-\left(1+\frac{\delta}{2}\right)} \sum_{i=1}^{N} \mathbb{E}\left[|\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i]|^{2+\delta} \right] = \mathbf{0}$$

for some $\delta > 0$*, then:*

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}_i]) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbb{V}\mathrm{ar}[\boldsymbol{x}])$$

where $\mathbb{V}\mathrm{ar}[\boldsymbol{x}]$ *is the probability limit of the variances as in Theorem 6.14.*

Ljapunov's version of the Central Limit Theorem establishes that the asymptotic normality result also holds with non-identically distributed data. In econometrics, this is of particular importance since it allows observations to be drawn from different distributions with heteroscedastic disturbances. Note that with respect to the classical Central Limit Theorem by Lindeberg and Lévy, Ljapunov's version only bears the additional requirements that variances be finite and that some absolute moment of order higher than two exists but is asymptotically dominated by the variances. The latter condition appears similar to that from Markov's Law of Large Numbers and, like that, convoluted; however, in most econometric applications $\mathbb{E}[\boldsymbol{x}_i] = \mathbf{0}$ for all $i = 1, \ldots, N$, and therefore that assumption specializes to:

$$\mathbb{E}\left[|X_{ik} X_{i\ell}|^{1+\delta} \right] < \infty \tag{6.1}$$

for any two elements $k, \ell = 1, \ldots, K$ of the random vector $\boldsymbol{x}$ and for all observations i. Under the hypothesis of independent observations, the asymptotic properties of most econometric estimators are obtained by invoking Ljapunov's Central Limit Theorem, hence conditions akin to (6.1) are routinely invoked and they are referred to as the "Ljapunov conditions."

Example 6.7. Asymptotic normality of the linear regression estimator. Let us return once again to the Method of Moments estimator of the bivariate linear regression slope from example 6.3. Rewrite it as:

$$\begin{aligned}
\widehat{\beta}_{1,MM} &= \frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)Y_i}{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2} \\
&= \frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(\beta_0 + \beta_1 X_i\right)}{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2} + \frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(Y_i - \beta_0 - \beta_1 X_i\right)}{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2} \\
&= \beta_1 \frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)X_i}{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2} + \frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\varepsilon_i}{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2} \\
&= \beta_1 + \frac{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\varepsilon_i}{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}
\end{aligned}$$

where

$$\varepsilon_i \equiv Y_i - \beta_0 - \beta_1 X_i$$

is the so-called **error term** of the regression model – that is, the deviation that occurs between Y_i and the linear conditional expectation function $\mathbb{E}\left[Y_i \middle| X_i\right] = \beta_0 + \beta_1 X_i$. The error term can be interpreted as a transformed random variable defined as a linear combination of the "primitive" random variables Y_i and X_i. Note that $\mathbb{E}\left[\varepsilon_i\right] = 0$ by the hypotheses on β_0.

Recall that in the bivariate linear regression model, the Law of Iterated Expectations implies $\mathbb{E}\left[X_i \varepsilon_i\right] = 0$. This observation provides another avenue for showing consistency of the MM estimator of the regression slope. In fact, by the Continuous Mapping Theorem:

$$\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\varepsilon_i \xrightarrow{p} \underbrace{\mathbb{E}\left[X_i \varepsilon_i\right]}_{=0} - \mathbb{E}\left[X_i\right]\underbrace{\mathbb{E}\left[\varepsilon_i\right]}_{=0} = 0$$

implying $\widehat{\beta}_{1,MM} \xrightarrow{p} \beta_1$. Furthermore, since the expression on the left-hand side is a sample mean, under the proper assumptions about the sample an applicable Central Limit Theorem implies the following.

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\varepsilon_i \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left[\varepsilon_i^2\left(X_i - \mathbb{E}\left[X_i\right]\right)^2\right]\right) \tag{6.2}$$

In (6.2) the limiting variance takes the stated form because $\bar{X} \xrightarrow{p} \mathbb{E}[X_i]$ at the probability limit. The limiting variance obtains as:

$$\begin{aligned}\mathbb{V}\text{ar}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(X_i - \mathbb{E}[X_i]\right)\varepsilon_i\right] &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{V}\text{ar}\left[\left(X_i - \mathbb{E}[X_i]\right)\varepsilon_i\right] \\ &= \mathbb{E}\left[\varepsilon_i^2\left(X_i - \mathbb{E}[X_i]\right)^2\right]\end{aligned}$$

while in the even more specialized case where the squared deviations of X_i and ε_i from their respective means are mutually independent, it is:

$$\mathbb{E}\left[\varepsilon_i^2\left(X_i - \mathbb{E}[X_i]\right)^2\right] = \mathbb{E}\left[\varepsilon_i^2\right]\mathbb{E}\left[\left(X_i - \mathbb{E}[X_i]\right)^2\right] = \sigma_\varepsilon^2 \cdot \mathbb{V}\text{ar}[X_i]$$

where $\sigma_\varepsilon^2 \equiv \mathbb{E}[\varepsilon_i^2]$. This latter case is the one where the conditional variance function of ε_i given X_i is actually a constant – a scenario commonly defined *homoscedasticity* (as opposed to *heteroscedasticity*, the general case).

The expression in (6.2), the above decomposition of the MM estimator of the bivariate linear regression slope, the Cramér-Wold device, as well as the following implication of the Continuous Mapping Theorem:

$$\left[\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2\right]^{-1} \xrightarrow{p} \left[\mathbb{V}\text{ar}[X_i]\right]^{-1}$$

all imply that the limiting distribution of the MM estimator is:

$$\sqrt{N}\left(\widehat{\beta}_{1,MM} - \beta_1\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}\left[\varepsilon_i^2\left(X_i - \mathbb{E}[X_i]\right)^2\right]}{\left(\mathbb{V}\text{ar}[X_i]\right)^2}\right) \tag{6.3}$$

and for some given N, its asymptotic distribution is as follows.

$$\widehat{\beta}_{1,MM} \overset{A}{\sim} \mathcal{N}\left(\beta_1, \frac{1}{N}\frac{\mathbb{E}\left[\varepsilon_i^2\left(X_i - \mathbb{E}[X_i]\right)^2\right]}{\left(\mathbb{V}\text{ar}[X_i]\right)^2}\right) \tag{6.4}$$

In the more specialized "homoscedastic" case, the limiting distribution is:

$$\sqrt{N}\left(\widehat{\beta}_{1,MM} - \beta_1\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2}{\mathbb{V}\text{ar}[X_i]}\right) \tag{6.5}$$

while the asymptotic distribution is derived consequently.

$$\widehat{\beta}_{1,MM} \overset{A}{\sim} \mathcal{N}\left(\beta_1, \frac{1}{N}\frac{\sigma_\varepsilon^2}{\mathbb{V}\text{ar}[X_i]}\right) \tag{6.6}$$

A proper econometric treatment of the linear regression model would discuss the multivariate generalization of these expressions, while additionally introducing the appropriate *estimators* for the unknown variances (or variance-covariances) of these distributions – that is, estimators of the asymptotic variances in (6.4) and (6.6). ■

The next result is is instrumental for the analysis and derivation of the asymptotic properties of many estimators.

Theorem 6.16. Delta Method. *Suppose that some random sequence of dimension K, $\boldsymbol{x}_N$, is asymptotically normal:*

$$\sqrt{N}\left(\boldsymbol{x}_N - \mathbf{c}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Upsilon}\right) \tag{6.7}$$

for some $K \times 1$ vector $\mathbf{c}$ and some $K \times K$ matrix $\boldsymbol{\Upsilon}$. In addition, consider some vector-valued function $\boldsymbol{d}\left(\boldsymbol{x}\right) : \mathbb{R}^K \to \mathbb{R}^J$. If the latter is continuously differentiable at $\mathbf{c}$ and the $J \times K$ Jacobian matrix

$$\boldsymbol{\Delta} \equiv \frac{\partial}{\partial \boldsymbol{x}^{\mathrm{T}}} \boldsymbol{d}\left(\mathbf{c}\right)$$

has full row rank J, the limiting distribution of $\boldsymbol{d}\left(\boldsymbol{x}_N\right)$ is as follows.

$$\sqrt{N}\left(\boldsymbol{d}\left(\boldsymbol{x}_N\right) - \boldsymbol{d}\left(\mathbf{c}\right)\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Delta}\boldsymbol{\Upsilon}\boldsymbol{\Delta}^{\mathrm{T}}\right) \tag{6.8}$$

Proof. From the mean value theorem:

$$\boldsymbol{d}\left(\boldsymbol{x}_N\right) = \boldsymbol{d}\left(\mathbf{c}\right) + \frac{\partial}{\partial \boldsymbol{x}^{\mathrm{T}}} \boldsymbol{d}\left(\widetilde{\boldsymbol{x}}_N\right)\left(\boldsymbol{x}_N - \mathbf{c}\right)$$

where $\widetilde{\boldsymbol{x}}_N$ is a convex combination of $\boldsymbol{x}_N$ and $\mathbf{c}$. However, as $\boldsymbol{x}_N \xrightarrow{p} \mathbf{c}$:

$$\frac{\partial}{\partial \boldsymbol{x}^{\mathrm{T}}} \boldsymbol{d}\left(\widetilde{\boldsymbol{x}}_N\right) \xrightarrow{p} \frac{\partial}{\partial \boldsymbol{x}^{\mathrm{T}}} \boldsymbol{d}\left(\mathbf{c}\right) = \boldsymbol{\Delta}$$

hence, at the probability limit:

$$\sqrt{N}\left(\boldsymbol{d}\left(\boldsymbol{x}_N\right) - \boldsymbol{d}\left(\mathbf{c}\right)\right) \xrightarrow{p} \boldsymbol{\Delta} \cdot \sqrt{N}\left(\boldsymbol{x}_N - \mathbf{c}\right)$$

which, together with (6.7), implies (6.8). □

With these results at hand, one can demonstrate that the classes of estimators introduced in Lecture 5 (Method of Moments and Maximum Likelihood estimators) achieve asymptotic normality under quite general assumptions. To facilitate the analysis, this is restricted to random (i.i.d.) samples.

Theorem 6.17. Asymptotically, Methods of Moments estimators are normally distributed. *An estimator $\widehat{\boldsymbol{\theta}}_{MM}$ defined as the solution of a set of sample moments (5.2) is asymptotically normal. If the sample is random and the moment conditions are differentiable the limiting distribution is:*

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_{MM} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{M}_0\boldsymbol{\Upsilon}_0\mathbf{M}_0^{\mathrm{T}}\right)$$

so long as the following matrices exist, are finite and nonsingular.

$$\boldsymbol{\Upsilon}_0 = \mathbb{V}\mathrm{ar}\left[\boldsymbol{m}\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right)\right] \qquad \mathbf{M}_0 \equiv \left[\mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \boldsymbol{m}\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right)\right]\right]^{-1}$$

Proof. The proof applies the same logic as the Delta Method. By the mean value theorem, the sample moment conditions are developed as:

$$\begin{aligned}\mathbf{0} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{m}\left(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}_{MM}\right) = \\ = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{m}\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right) + \left[\frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \boldsymbol{m}\left(\boldsymbol{x}_i; \widetilde{\boldsymbol{\theta}}_N\right)\right]\left(\widehat{\boldsymbol{\theta}}_{MM} - \boldsymbol{\theta}_0\right)\end{aligned}$$

where the first expression on the upper line is equal to zero by construction of all Method of Moments estimators. After multiplying both sides by $\sqrt{N}$ and some manipulation the above expression is rendered as follows.

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_{MM} - \boldsymbol{\theta}_0\right) = -\left[\frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \boldsymbol{m}\left(\boldsymbol{x}_i; \widetilde{\boldsymbol{\theta}}_N\right)\right]^{-1} \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \boldsymbol{m}\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right)$$

Note that, since this is a random sample:

1. by a suitable Central Limit Theorem:

$$-\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \boldsymbol{m}\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbb{V}\mathrm{ar}\left[\boldsymbol{m}\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right)\right]\right)$$

   since $\mathbb{E}\left[\boldsymbol{m}\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right)\right] = \mathbf{0}$ by hypothesis;

2. while by the Weak Law of Large Numbers:

$$\frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \boldsymbol{m}\left(\boldsymbol{x}_i; \widetilde{\boldsymbol{\theta}}_N\right) \xrightarrow{p} \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \boldsymbol{m}\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right)\right]$$

   since $\widetilde{\boldsymbol{\theta}}_N \xrightarrow{p} \boldsymbol{\theta}_0$ by consistency of the estimator (at the limit, $\widetilde{\boldsymbol{\theta}}_N$, $\widehat{\boldsymbol{\theta}}_{MM}$ and $\boldsymbol{\theta}_0$ all coincide).

These intermediate results are together combined via the Continuous Mapping Theorem, Slutskij's Theorem and the Cramér-Wold device so to imply the statement. Consequently, *for a fixed* N the asymptotic distribution is:

$$\widehat{\boldsymbol{\theta}}_{MM} \overset{A}{\sim} \mathcal{N}\left(\boldsymbol{\theta}_0, \frac{1}{N}\mathbf{M}_0 \boldsymbol{\Upsilon}_0 \mathbf{M}_0^{\mathrm{T}}\right)$$

which concludes the proof. □

An analogous result holds for Maximum Likelihood estimators as well, and the proof is almost identical. In this case, however, the result is especially powerful, as the asymptotic variance coincides with the Cramér-Rao bound.

Theorem 6.18. Asymptotically, Maximum Likelihood estimators are normally distributed and they attain the Cramér-Rao bound. *An estimator* $\widehat{\boldsymbol{\theta}}_{MLE}$ *defined as the maximizer of a log-likelihood function as per (5.19) is asymptotically normal. If the sample is random and some so-called* regularity conditions *hold:*

i. *the problem is well defined, i.e.* $\boldsymbol{\theta}_0$ *is the maximizer of the population expression* $\mathbb{E}\left[\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\boldsymbol{\theta}\right)\right]$ *– where* $f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\boldsymbol{\theta}\right)$ *is the probability mass or density function that generates the data;*

ii. $f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\boldsymbol{\theta}\right)$ *is three times continuously differentiable and its derivatives are bounded in absolute value;*

iii. *the support of* $\boldsymbol{x}_i$ *does not depend on* $\boldsymbol{\theta}$*, so that derivatives for* $\boldsymbol{\theta}$ *can pass at least twice through an integral defined in terms of* $f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\boldsymbol{\theta}\right)$*;*

then the limiting distribution is expressible as:

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_{MLE}-\boldsymbol{\theta}_0\right)\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\left[\boldsymbol{I}\left(\boldsymbol{\theta}_0\right)\right]^{-1}\right)$$

where $\boldsymbol{I}\left(\boldsymbol{\theta}_0\right)$ *– without the* N *subscript – is the expression for the following "single-observation" information matrix evaluated at* $\boldsymbol{\theta}_0$*.*

$$\begin{aligned}\boldsymbol{I}\left(\boldsymbol{\theta}_0\right)&\equiv\mathbb{E}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right)^{\mathrm{T}}\right]\\&=-\mathbb{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right]\end{aligned}$$

Consequently, $\widehat{\boldsymbol{\theta}}_{MLE}$ *asymptotically attains the Cramér-Rao bound.*

Proof. The proof proceeds similarly to the Method of Moments case. By the mean value theorem, the MLE First Order Conditions can be stated as:

$$\begin{aligned}\mathbf{0}=\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{MLE}\right)&=\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)+\\&+\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\widetilde{\boldsymbol{\theta}}_N\right)\right]\left(\widehat{\boldsymbol{\theta}}_{MLE}-\boldsymbol{\theta}_0\right)\end{aligned}$$

where the entire expression is zero by definition of MLE. Once again:

$$\begin{aligned}&\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_{MLE}-\boldsymbol{\theta}_0\right)=\\&\quad=-\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\widetilde{\boldsymbol{\theta}}_N\right)\right]^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\end{aligned}$$

but in this case some additional simplifications are possible, thanks to the Information Matrix Equality. In fact, under the regularity conditions:

1. a suitable Central Limit Theorem implies that:

$$-\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{MLE}\right)\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\boldsymbol{I}\left(\boldsymbol{\theta}_0\right)\right)$$

   since $\boldsymbol{\theta}_0$ maximizes $\mathbb{E}\left[\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right]$, hence $\mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right]=\mathbf{0}$;

2. while by the Weak Law of Large Numbers:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\widetilde{\boldsymbol{\theta}}_N\right)\xrightarrow{p}-\boldsymbol{I}\left(\boldsymbol{\theta}_0\right)$$

   again since $\widetilde{\boldsymbol{\theta}}_N\xrightarrow{p}\boldsymbol{\theta}_0$ by consistency of MLE as per Theorem 6.9.

Hence, here the application of the Delta Method results in a simplified expression of the limiting variance, as given in the statement of the Theorem. Collecting terms, *for some fixed* N the asymptotic distribution is:

$$\widehat{\boldsymbol{\theta}}_{MLE}\overset{A}{\sim}\mathcal{N}\left(\boldsymbol{\theta}_0,\left[\boldsymbol{I}_N\left(\boldsymbol{\theta}_0\right)\right]^{-1}\right)$$

where $\boldsymbol{I}_N\left(\boldsymbol{\theta}_0\right)$ is the grand (sample) information matrix for some fixed N. Since the MLE is asymptotically consistent, at the probability limit its bias is zero, hence the estimator attains the Cramér-Rao bound. □

This result is celebrated, since it motivates the reputation of Maximum Likelihood as the method for constructing estimators with the best statistical properties. Yet one should be careful at overusing Maximum Likelihood on the expectation that the resulting estimators are asymptotically efficient. In fact, Maximum Likelihood is very sensitive to the assumptions about the distributions that generate the sample, and it can fail utterly (i.e. produce inconsistent estimates) if the assumptions are incorrect. On the other hand, the Method of Moments is generally more robust. This creates some tension between efficiency and robustness when choosing between estimators.

It is worth to conclude this section by summarizing the advantages of conducting statistical analysis in large samples such that asymptotic results apply. In these settings, standard estimators are known to be normally distributed; this facilitates statistical inference immensely since it is generally very difficult to derive their *exact* distributions in small samples. While the variances of these estimators are usually unknown quantities, they can be easily *consistently estimated* via their sample analogues; for example, under normal sampling it is $\frac{N-1}{N}S^2\xrightarrow{p}\sigma^2$. Another example follows suit.

Example 6.8. Asymptotically testing for the regression slope β_1 of the bivariate regression model. Let us return one more time to the bivariate regression model. Recall the two-sided hypotheses from Example 5.18 in the previous lecture:

$$\mathrm{H}_0: \ \beta_1 = C \qquad\qquad \mathrm{H}_1: \ \beta_1 \neq C$$

and similarly the analogous one-sided ones. In applied regression analysis, the most common test is the one about $C = 0$, also called *significance test* of the regression, since it is effectively a test about whether the explanatory variable X_i affects the mean of Y_i in a conditional sense.

As discussed briefly in Lecture 5, with small samples this test is problematic, since it requires to make specific assumptions about the conditional distribution of Y_i given X_i. In an asymptotic environment, however, it is possible to rely on the implications of the Central Limit Theorem discussed in Example 6.7. Thus, by Observation 6.1 the t-statistic defined as:

$$\mathrm{t}_N = \sqrt{N}\,\frac{\widehat{\beta}_{1,MM} - C}{S_{\beta_1}} \xrightarrow{d} \mathcal{N}(0,1)$$

asymptotically follows – under the null hypothesis – the standard normal distribution. Above, S_{β_1} is the *sample* standard deviation of the estimator; the corresponding sample variance $S^2_{\beta_1}$ is calculated as the sample analogue of the relevant *limiting* variance. In the general heteroscedastic case, it is:

$$S^2_{\beta_1} = N\,\frac{\sum_{i=1}^N \left(Y_i - \widehat{\beta}_{0,MM} - \widehat{\beta}_{1,MM}X_i\right)^2 \left(X_i - \bar{X}\right)^2}{\left[\sum_{i=1}^N \left(X_i - \bar{X}\right)^2\right]^2}$$

while in the more restricted homoscedastic case $S^2_{\beta_1}$ is as follows.

$$S^2_{\beta_1} = \frac{\sum_{i=1}^N \left(Y_i - \widehat{\beta}_{0,MM} - \widehat{\beta}_{1,MM}X_i\right)^2}{\sum_{i=1}^N \left(X_i - \bar{X}\right)^2}$$

The quantity $S_{\beta_1}/\sqrt{N}$ is called the **standard error** of the estimate $\widehat{\beta}_{1,MM}$. With these results at hand, it is possible to conduct tests of hypotheses and construct confidence intervals under the familiar framework of the normal distribution. For example, the confidence interval of β_1 would be as follows.

$$\beta_1 \in \left[\widehat{\beta}_{1,MM} - z^*_{\alpha/2}\frac{S_{\beta_1}}{\sqrt{N}}, \widehat{\beta}_{1,MM} + z^*_{\alpha/2}\frac{S_{\beta_1}}{\sqrt{N}}\right]$$

This example concludes the analysis of the bivariate linear regression model. All concepts and ideas related to it which were developed in various examples extend easily to the multivariate version of the model. ■

As the last example has shown, the specific formulae for the estimation of the asymptotic variance are typically context- and assumption-dependent. In random samples, however, it is easy to establish expressions with a more general validity. Consider Method of Moments estimators first; in the i.i.d. framework, a general expression for a consistent estimator of their asymptotic variance is given by $N^{-1}\widehat{\mathbf{M}}_N\widehat{\boldsymbol{\Upsilon}}_N\widehat{\mathbf{M}}_N^{\mathrm{T}}$, where

$$\widehat{\mathbf{M}}_N \equiv \left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\boldsymbol{m}\left(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{MM}\right)\right]^{-1} \xrightarrow{p} \mathbf{M}_0$$

is a consistent estimator of $\mathbf{M}_0$ (by some applicable Law of Large Numbers and the Continuous Mapping Theorem), while

$$\widehat{\boldsymbol{\Upsilon}}_N \equiv \frac{1}{N}\sum_{i=1}^{N}\left[\boldsymbol{m}\left(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{MM}\right)\right]\left[\boldsymbol{m}\left(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{MM}\right)\right]^{\mathrm{T}} \xrightarrow{p} \boldsymbol{\Upsilon}_0$$

is also a consistent estimator of the variance of the zero moment conditions by some applicable Law of Large Numbers, since in a random sample the following holds.[6]

$$\boldsymbol{\Upsilon}_0 = \mathbb{V}\mathrm{ar}\left[\boldsymbol{m}\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right] = \mathbb{E}\left[\left(\boldsymbol{m}\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right)\left(\boldsymbol{m}\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right)^{\mathrm{T}}\right]$$

These estimating matrices are not only based on sample analogues of their population counterparts (the object of estimation), something which is indicated with the subscript N instead of 0. In addition, they are also evaluated at the *estimated* parameters $\boldsymbol{\theta}$, which is symbolized by the wide "hat" used to denote them. In the Maximum Likelihood case, the information matrix equality offers two alternative routes for estimating the asymptotic variance. The first option is based on the Hessian of the mass or density function:

$$\widehat{\mathbf{H}}_N \equiv -\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{MLE}\right) \xrightarrow{p} \boldsymbol{I}\left(\boldsymbol{\theta}_0\right)$$

while the second option exploits the "squared" score.

$$\widehat{\mathbf{J}}_N \equiv \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{MLE}\right)\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{MLE}\right)\right)^{\mathrm{T}} \xrightarrow{p} \boldsymbol{I}\left(\boldsymbol{\theta}_0\right)$$

Both matrices $\widehat{\mathbf{H}}_N$ and $\widehat{\mathbf{J}}_N$ are evaluated at the MLE solution and, in random samples, let consistently estimate the information matrix; in practice the choice of a specific option is usually based on convenience. It is important to familiarize with this type of results and the associated notation, as they are typical of the standard treatment of econometric theory.

[6] Under fairly general assumptions, it is $\widehat{\boldsymbol{\Upsilon}}_N \xrightarrow{p} \boldsymbol{\Upsilon}_0$ also with i.n.i.d. observations.

Part II

Econometric Theory

Lecture 7

The Linear Regression Model

This lecture introduces a workhorse model of statistics and econometrics: the linear regression model. To this end, the lecture develops the method of Least Squares as the algebraic solution to some linear prediction problem, and subsequently discusses its properties and relationships to the (possibly linear) population regression function. The treatment and the terminology adopted are purposefully typical of the econometric approach.

7.1 Linear Socio-economic Relationships

Suppose that a researcher is studying an economic or social phenomenon, and postulates the existence of a **linear** relationship between a **dependent** variable Y_i; K independent or **explanatory** variables $\boldsymbol{x} = (X_{1i}, \ldots, X_{Ki})$; and finally an unobserved "disturbance" or **error term** ε_i.

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + \varepsilon_i \tag{7.1}$$

The researcher has access to a sample $\{(y_i, \mathbf{x}_i)\}_{i=1}^N$ of size $N > K$, where y_i is the **realization** of Y_i for the i-th observation, while the vector $\mathbf{x}_i$:

$$\mathbf{x}_i = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{Ki} \end{bmatrix}$$

has length K and collects all the **realizations** of the explanatory variables in $\boldsymbol{x}$ for the i-th observation. One can write (7.1) *in terms of realizations*:

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i \tag{7.2}$$

where $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 & \beta_2 & \ldots & \beta_K \end{bmatrix}^{\mathrm{T}}$ is the **parameter vector** of the model.

According to typical econometric terminology, (7.1) can be given a so-called "structural" interpretation in the sense that the *endogenous* variable Y_i is assumed to depend linearly upon some K independent, *exogenous* variables $(X_{1i}, \ldots, X_{Ki})$ because of *a priori* knowledge of the setting under analysis or theoretical reasoning. The linear relationship is augmented with the inclusion of the unobserved error ε_i*, which on the one hand represents all the other, gene*rally unknown factors that also determine Y_i, and on the other hand it deprives the linear relationship of any deterministic content. In fact, social phenomena are not determined according to fixed rules, thus any exact relationship is bound to be rejected by the empirical observation. Observe that no statement has been made yet about the joint probability distribution that determines $(\mathbf{x}_i, \varepsilon_i)$ – and thus $(y_i, \mathbf{x}_i)$.

Typically, researchers introduce a **constant term** as an independent parameter into the specification of linear relationships such as (7.1):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_{(K-1)} X_{(K-1)i} + \varepsilon_i \tag{7.3}$$

observe that in addition to the constant parameter β_0, the model has $K-1$ independent variables X_{ki}, so that in total the model still has K parameters. Here, the model can be still written as in (7.2), while vector $\mathbf{x}_i$ becomes:

$$\mathbf{x}_i = \begin{bmatrix} 1 \\ x_{1i} \\ \vdots \\ x_{(K-1)i} \end{bmatrix}$$

in practice a "constant" variable X_{0i}, normalized for convenience as $x_{i0} = 1$ for each observation i, is included in the model. This offers the advantage of letting researchers confidently assume that $\mathbb{E}[\varepsilon_i] = 0$: intuitively, if the average influence of the unobserved factors on Y_i were different from zero, the extent of such an effect would be conceptually equivalent to a higher constant coefficient β_0. In other words, introducing constant terms allows researchers to disregard all those unobserved or uninteresting factors that affect the *unconditional average* of the dependent variable Y_i.

According to a traditional view, the econometric analysis of linear relationships such as (7.1) or (7.3) has the objective of giving empirical content to parameters like $\boldsymbol{\beta}$, so that economists can relate the abstract relationships featured in their theories to actual quantities governing variables that can be observed in the real world. This would enable economists to make statements about causal mechanisms, ascertain the effect of economic policies, analyze counterfactuals, provide forecasts, and more. Such an intellectual tradition can be traced back to the most archaeological econometric models.

Example 7.1. The Keynesian consumption function. The diffusion of Keynes' General Theory in the middle of the twentieth century and the birth of macroeconomics as a distinct subdiscipline are associated with the diffusion of new theories about macroeconomic relationships. Among those, perhaps the most exemplary case is the "Keynesian" *consumption function*:

$$C_i = \mathrm{c}_0 + \mathrm{c}_1 Y_i + \varepsilon_i \tag{7.4}$$

where C_i represents aggregate *consumption*, Y_i aggregate *disposable income*, ε_i is a disturbance term and $(\mathrm{c}_0, \mathrm{c}_1)$ are the two parameters of interest; c_1, in particular, measures the dependence of C_i on Y_i and it bears the name of **marginal propensity of consumption**, a parameter that played an important role in the early macroeconomic theories following the Keynesian legacy. To evaluate (7.4) with real world data one should have access to a sample of economic regions with plausibly the same marginal propensity to consume, or instead to multiple observations on the same region or country. Sequences of observations about the same unit of analysis tracked over time are called **time series** and feature preminently in macroeconometrics.

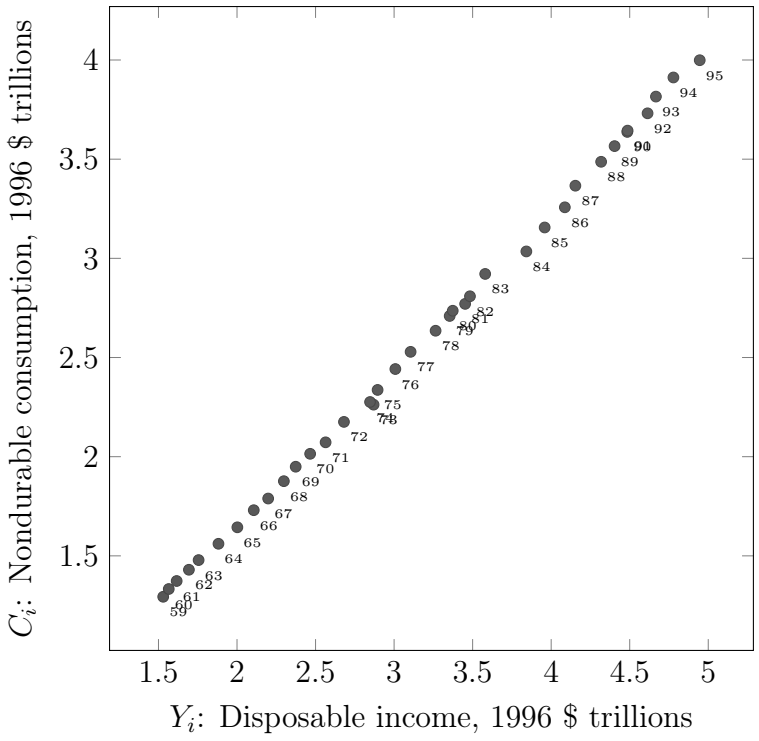


Figure 7.1: Disposable Income and Consumption, 1959-1995 US data

Figure 7.1 depicts an example of the association between two time series of Y and C based on actual macroeconomic data (both series are normalized to 1992 prices). The relationship between the two variables in question appears to be robustly linear, so that at first sight a structural relationship like (7.4) seems justified. Applied macroeconometric research has demonstrated that, in fact, the strong linear association between the levels of income and consumption which is typically observed in the data is *spurious*, in the sense that the variation of both variables, while certainly reciprocally related, is influenced by parallel trends that influence both income and consumption. This finding has led to the development of more sophisticated linear and non-linear models for the analysis of macroeconomic time series. ■

Example 7.2. Human capital and wages. A much celebrated theory in labor economics postulates that wages, being a function of the individual (marginal) productivity, are a function of those factors that makes workers more productive. Collectively, these factors fall under the name of **human capital**; while nowadays this is a common expression, the idea of extending to individuals a concept analogous to that of physical capital (Walsh, 1935; Becker, 1962) was initially quite an original theoretical contribution. The seminal framework for the empirical analysis of human capital is originally due to Mincer (1958), who introduced the following relationship between the wages W_i of workers, their experience in the workplace X_i, their education S_i, their ability α_i, and finally to some "residual" factors ϵ_i that influence their labor market outcomes, say sheer luck in landing a good job.

$$W_i = \exp\left(\beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 S_i\right) \exp\left(\alpha_i + \epsilon_i\right) \tag{7.5}$$

While this relationship is certainly non-linear, it can be easily transformed so that it becomes *linear in the parameters* $\boldsymbol{\beta}_W = \{\beta_0, \beta_1, \beta_2, \beta_3\}$, a relationship that is best known as the **Mincer equation**.

$$\log W_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 S_i + \alpha_i + \epsilon_i \tag{7.6}$$

The functional form of the Mincer equation was originally motivated on empirical observation, and it still is nowadays the workhorse model for the analysis of the **returns to education**, which in the model are subsumed by the parameter β_3 associated with the education variable S_i. The reason is that, by keeping the analogy with physical capital, human capital too is something that can be enhanced by investment – in this case, by acquiring more education. The latter can also be modeled, for example as:

$$S_i = \gamma_0 + \gamma_1 Z_i + \phi_1 X_i + \phi_2 X_i^2 + \psi_0 \alpha_i + \eta_i \tag{7.7}$$

that is as some function of S_i, which in this case is linear in the parameters like γ_1, ϕ_1 *et cetera*, and that includes a squared term for experience X_i, some generic factors Z_i that affect the individual choice in education, ability α_i, as well as other unobserved factors η_i. Empirical labor economists typically augment their analyses about returns to schooling with the inclusion of a specification of the education model (7.7), with the aim of addressing a typical econometric problem dubbed "endogeneity" which ultimately stems from the inability of econometricians to observe individual ability α_i; later lectures elaborate on this topic and elucidate the advantages of specifying a linear-in-the-parameters function for education S_i.

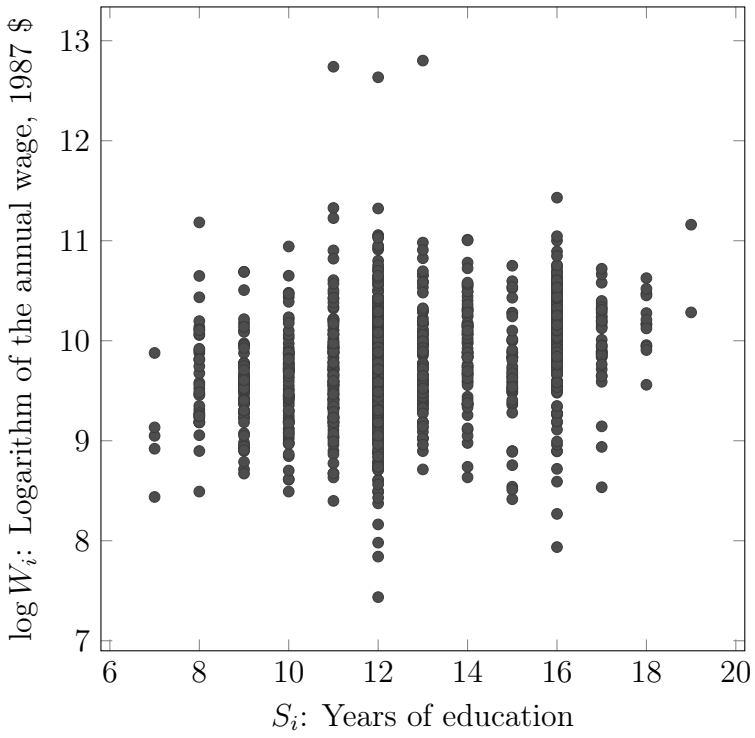


Figure 7.2: (Log) wage and education in 1987, excerpt of a survey

It is instructive to display in graphical form the relationship between (log) wages and education. Figure 7.2 is obtained from an excerpt of a longitudinal survey of workers, by isolating observations from one specific year. Using some more technical terminology, a single **cross section** of individuals was isolated from a larger **longitudinal** or **panel** dataset.[1] Certainly,

[1]This is based on the "Keane" panel, originally spanning the years 1981 to 1987, available online at `http://fmwww.bc.edu/ec-p/data/wooldridge/datasets.list.html`.

panel or longitudinal data are typically more informative and useful for the purposes of microeconometric analysis (while also being typically costlier to gather and often not readily available); however, cross-sectional data – by virtue of being simpler – have pedagogical value in selected settings.

The scatter plot displayed in Figure 7.2 represents the raw relationship between (log) wages and the attained level of education, while ignoring other variables (such as say experience and ability) which also might bear effects on earnings. Unlike the association between consumption and income from Figure 7.1, that the relationship between log-wages and education is linear is not immediately clear through the visual representation of the data. However, repeated analyses have shown that this relationship is "more linear" than the one between the *level* of wages and education (an observation that helps motivate the popularity of the Mincer equation). In addition, observe that the independent education variable S_i takes values upon a discrete set, which is a typical feature of many microeconometric datasets. ■

Both 7.1 and 7.2 are valid examples of economic structural relationships framed through linear equations. However, as motivated at length over the course of this lecture, the linear model is a powerful tool whose properties are quite useful even when studying relationships that are not strictly structural or, under certain conditions, that are structural but not necessarily linear. The analysis of the linear model often benefits from a mathematical representation based on **compact matrix notation**. Define:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} ; \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \mathbf{x}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_N^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{21} & \dots & x_{K1} \\ x_{12} & x_{22} & \dots & x_{K2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1N} & x_{2N} & \dots & x_{KN} \end{bmatrix} ; \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

that is, $\mathbf{y}$, $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ are obtained by vertically stacking over all observations, respectively, the realization y_i of the dependent variable, the transpose of the vector $\mathbf{x}_i$, and the error term ε_i. If the model features a constant term, the first column of $\mathbf{X}$ is the ι_N vector whose entries are all equal to 1. With compact matrix notation, (7.1) can be conveniently written as follows.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{7.8}$$

Econometric models are often written in terms of realizations, not in terms of abstract random variables, vector or matrices (note that the distinction does not apply to the error terms, which cannot be observed by definition). Because of this convention, in what follows the notation adopted to describe specific models alternates between the two cases, depending on purpose and convenience (that is, which notation best clarifies a certain concept).

7.2 Optimal Linear Prediction

Having specified a linear economic relationship, the objective of the econometrician is that of assigning a value to the parameters $\boldsymbol{\beta}$ that "make sense" on the basis of the real world observations of $(\mathbf{x}_i, y_i)$. For the moment I shall not speak of "estimating" parameters, since this term involves dealing with distributional assumptions and statistical inference. In what follows, I discuss how the Least Squares solution – on which the Ordinary Least Squares (OLS) estimator for linear regression is based – can be derived as the sample analog of the solution of a **population prediction problem** which is restricted to linear prediction functions. This analysis allows to appreciate certain properties of the Least Squares solution that are often invoked to motivate linear regression analysis.

Suppose that the researcher aims at specifying a **prediction function** $\widehat{y}_i = m_y(\mathbf{x}_i)$, meant as the "best guess" of some *unknown* value of $Y_i = y_i$ based on the observation of a vector of independent factors $\mathbf{x}_i$. Clearly, the farther away the prediction $\widehat{y}_i$ is from the actual realization of Y_i, the worse it is for the researchers. This implies the existence of some **loss function**:

$$L(e_i) = L(y_i - \widehat{y}_i) \tag{7.9}$$

with $e_i \equiv y_i - \widehat{y}_i$ and where $L(e_i)$ has the properties that it is increasing in $|y_i - \widehat{y}_i|$ and that $L(0) = 0$. If the researcher aims at specifying a predictor that is consistent across different realizations of $\mathbf{x}_i$, a sensible criterion is to choose the function $m(\mathbf{x}_i)$ that minimizes the **expected loss**:

$$\mathbb{E}\left[L\left(Y_i - \widehat{Y}_i\right)\right] = \mathbb{E}\left[L\left(Y_i - m_y(\boldsymbol{x}_i)\right)\right] \tag{7.10}$$

where the expectation is taken on the joint support of Y_i and $(X_{1i}, \ldots, X_{Ki})$.

This still leaves open the question about the choice of the working loss function $L(e_i)$. In general, this choice may depend on the context; here, the analysis is focused on the **quadratic loss** $L(e_i) = e_i^2$. The quadratic loss is appealing, since deviations of the prediction from the "true" realization of Y_i are disproportionately more "harmful" the higher they are. The expected quadratic loss is the so-called **mean squared error** of prediction.

$$\text{MSE} = \mathbb{E}\left[\left(Y_i - m_y(\boldsymbol{x}_i)\right)^2\right] \tag{7.11}$$

Alternative loss criteria exist. For example, the **absolute loss** $L(e_i) = |e_i|$ differs from the quadratic loss in that it does not disproportionately punish large mistakes. For some $p \in (0, 1)$, the **quantile loss**

$$L(e_i) = p\,|e_i| \cdot \mathbb{1}[e_i \geq 0] + (1 - p)\,|e_i| \cdot \mathbb{1}[e_i < 0]$$

is asymmetric: for prediction errors of the same absolute size $|e_i|$, it punishes *underprediction* ($e_i > 0$) more than *overprediction* ($e_i < 0$) if $p > 0.5$ – and vice versa; the asymmetry increases the farther p departs from 0.5. Observe that the absolute loss is a special case of the quantile loss, for $p = 0.5$.

The remainder of this analysis focuses, as anticipated, on the quadratic loss. The Mean Squared Error of prediction is associated with a well-known statistical result.

Theorem 7.1. CEF as Optimal Predictor under Quadratic Loss. *If* $\mathbb{V}\mathrm{ar}\left[Y_i \middle| \boldsymbol{x}_i\right] < \infty$*, the predictor* $m_y\left(\boldsymbol{x}_i\right)$ *that minimizes the Mean Squared Error is the Conditional Expectation Function (CEF):* $m_y\left(\boldsymbol{x}_i\right) = \mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right]$.

Proof. By the standard decomposition of the MSE:

$$\begin{aligned}
\mathbb{E}\left[\left(Y_i - m_y\left(\boldsymbol{x}_i\right)\right)^2\right] &= \mathbb{E}\left[\left(Y_i - \mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] + \mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] - m_y\left(\boldsymbol{x}_i\right)\right)^2\right] \\
&= \mathbb{E}\left[\left(Y_i - \mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right]\right)^2\right] + \mathbb{E}\left[\left(\mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] - m_y\left(\boldsymbol{x}_i\right)\right)^2\right] \\
&\quad + 2\,\mathbb{E}\left[\left(Y_i - \mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right]\right)\left(\mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] - m_y\left(\boldsymbol{x}_i\right)\right)\right] \\
&= \mathbb{E}\left[\left(Y_i - \mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right]\right)^2\right] + \mathbb{E}\left[\left(\mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] - m_y\left(\boldsymbol{x}_i\right)\right)^2\right] \\
&= \mathbb{V}\mathrm{ar}\left[Y_i \middle| \boldsymbol{x}_i\right] + \mathbb{E}\left[\left(\mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] - m_y\left(\boldsymbol{x}_i\right)\right)^2\right]
\end{aligned}$$

which is minimized if $\mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] = m_y\left(\boldsymbol{x}_i\right)$ so long as $\mathbb{V}\mathrm{ar}\left[Y_i \middle| \boldsymbol{x}_i\right] < \infty$. Note that the last term in the third line vanishes since:

$$\begin{aligned}
&\mathbb{E}\left[\left(Y_i - \mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right]\right)\left(\mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] - m_y\left(\boldsymbol{x}_i\right)\right)\right] = \\
&\qquad = \mathbb{E}\left[\left(\mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] - m_y\left(\boldsymbol{x}_i\right)\right) \cdot \underbrace{\mathbb{E}\left[\left(Y_i - \mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right]\right) \middle| \boldsymbol{x}_i\right]}_{=0}\right] = 0
\end{aligned}$$

an observation that carefully exploits the Law of Iterated Expectations. □

The fact that the conditional expectation function – the *population* conditional average of Y_i – is the best predictor may appear intuitive; however, this results does not generally hold for all loss functions! In fact, it is limited to the quadratic loss. Under different loss functions the optimal predictor is different: for example, the one associated with the absolute loss $L\left(e_i\right) = |e_i|$ is the conditional **median** of Y_i given $\boldsymbol{x}_i$; with the quantile loss, the optimal predictor is the p-th conditional **quantile** of Y_i given $\boldsymbol{x}_i$. Nevertheless, the main result should be reminiscent of the simpler observation that the mean $\mathbb{E}\left[X\right]$ of a random variable X is the latter's best "guess" (predictor) under a quadratic loss criterion (Lecture 1). Theorem 7.1 generalizes that finding.

With this result at hand, return to the researcher's prediction problem assuming while maintaining choice of a quadratic loss function. Given that the relationship (7.1) under analysis is hypothesized linear, it is intuitively interesting to examine the consequences of restricting the analysis to an **optimal linear predictor**. In other words, let $m_y(\boldsymbol{x}_i) = p_y(\boldsymbol{x}_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*$,[2] where

$$\boldsymbol{\beta}^* \in \underset{\boldsymbol{\beta}\in\mathbb{R}^K}{\arg\min}\, \mathbb{E}\left[\left(Y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2\right] \tag{7.12}$$

that is, $\boldsymbol{\beta}^*$ is one specific coefficient vector which, among all predictors that are **linear** in $\boldsymbol{x}_i$, minimizes the Mean Squared Error (observe that $\boldsymbol{\beta}^*$ needs not be unique). The implications are summarized with the next result.

Theorem 7.2. Optimal Linear Predictor as best approximation to the CEF. *Consider* any *vector* $\boldsymbol{\beta}^*$ *as defined in (7.12). If* $\mathbb{V}\mathrm{ar}[Y_i|\,\boldsymbol{x}_i] < \infty$, *then:*

$$\boldsymbol{\beta}^* \in \underset{\boldsymbol{\beta}\in\mathbb{R}^K}{\arg\min}\, \mathbb{E}\left[\left(\mathbb{E}[Y_i|\,\boldsymbol{x}_i] - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2\right] \tag{7.13}$$

that is, any optimal linear predictor of Y_i *is also an optimal linear predictor of the CEF,* $\mathbb{E}[Y_i|\,\boldsymbol{x}_i]$, *in the MSE sense.*

Proof. The demonstration is analogous to that of Theorem 7.1:

$$\begin{aligned}
\mathbb{E}\left[\left(Y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2\right] &= \mathbb{E}\left[\left(Y_i - \mathbb{E}[Y_i|\,\boldsymbol{x}_i] + \mathbb{E}[Y_i|\,\boldsymbol{x}_i] - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2\right] \\
&= \mathbb{E}\left[(Y_i - \mathbb{E}[Y_i|\,\boldsymbol{x}_i])^2\right] + \mathbb{E}\left[\left(\mathbb{E}[Y_i|\,\boldsymbol{x}_i] - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2\right] \\
&\quad + 2\,\mathbb{E}\left[(Y_i - \mathbb{E}[Y_i|\,\boldsymbol{x}_i])\left(\mathbb{E}[Y_i|\,\boldsymbol{x}_i] - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right] \\
&= \mathbb{E}\left[(Y_i - \mathbb{E}[Y_i|\,\boldsymbol{x}_i])^2\right] + \mathbb{E}\left[\left(\mathbb{E}[Y_i|\,\boldsymbol{x}_i] - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2\right] \\
&= \mathbb{V}\mathrm{ar}[Y_i|\,\boldsymbol{x}_i] + \mathbb{E}\left[\left(\mathbb{E}[Y_i|\,\boldsymbol{x}_i] - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2\right]
\end{aligned}$$

where again, the cross-term in the third line disappears by a proper application of the Law of Iterated Expectations. Therefore, the two minimizers in (7.12) and (7.13) are identical so long as $\mathbb{V}\mathrm{ar}[Y_i|\,\boldsymbol{x}_i]$ is a finite constant. □

The interpretation of Theorem 7.2 is that even if the CEF is unknown, choosing an optimal linear predictor results in the **best approximation** to the *true* CEF that can be attained with a **linear** function, where "best approximation" shall be translated as "the minimal expected squared loss."

[2]The asterisk is related to an alternative notation for the optimal linear predictor, which sometimes is written as $\mathbb{E}^*[Y_i|\,\boldsymbol{x}_i] = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*$ with $\boldsymbol{\beta}^*$ defined as in (7.12). Instead, the notation $p_y(\boldsymbol{x}_i)$ specifies that the prediction function $m_y(\boldsymbol{x}_i)$ is *linear*.

This observation has had a profound impact on motivating the use of linear regression analysis in contexts where the true form of the statistical dependence between a dependent variable Y_i and a set of independent variables $(X_{1i}, \ldots, X_{Ki})$ is unknown. This important interpretation is revisited later at the end of this lecture.

With the knowledge about the relationship between optimal predictors, optimal linear predictors and the CEF at hand, it is convenient to express the solution of the optimal linear prediction problem. One can rewrite the First Order Conditions of the problem in (7.12) as:[3]

$$\mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right] \boldsymbol{\beta}^* = \mathbb{E}\left[\boldsymbol{x}_i Y_i\right] \tag{7.14}$$

therefore a **unique solution** exists if matrix $\mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right]$ is nonsingular, and it reads as:

$$\boldsymbol{\beta}^* = \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right]^{-1} \mathbb{E}\left[\boldsymbol{x}_i Y_i\right] \tag{7.15}$$

implying that the **optimal linear predictor** is as follows.

$$p_y\left(\boldsymbol{x}_i\right) = \boldsymbol{x}_i^{\mathrm{T}} \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right]^{-1} \mathbb{E}\left[\boldsymbol{x}_i Y_i\right] \tag{7.16}$$

This expression is also called the (population) **linear projection** of Y_i given $\boldsymbol{x}_i$, for reasons that are to appear clearer after examining the algebraic and geometric properties of its sample analog: the Least Squares solution.

Example 7.3. Linear approximation of a particular quadratic CEF. It is useful to illustrate this concept by developing and visualizing an example. Suppose that some variable Y_i of interest for prediction only depends on a single explanatory variable X_i. In addition, suppose that the *true* CEF of Y_i given X_i is quadratic, and in particular it is as follows.

$$\mathbb{E}\left[Y_i \middle| X_i\right] = X_i - \frac{1}{10} X_i^2$$

Quadratic relationships similar to the above are easy to identify in social sciences. For example, the Mincer equation from Example 7.2 is quadratic in experience X_i, and its second degree parameter is usually evaluated small and negative in empirical studies. In this example the above CEF does not include a term of degree zero, but this only so for convenience.

[3] Note that (7.14) is equivalent to the First Order Condition of the problem in (7.13), that of finding the MSE-best linear approximation to the CEF. The latter FOC is

$$\mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right] \boldsymbol{\beta}^* = \mathbb{E}\left[\boldsymbol{x}_i \cdot \mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right]\right]$$

while $\mathbb{E}\left[\boldsymbol{x}_i \cdot \mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right]\right] = \mathbb{E}\left[\boldsymbol{x}_i Y_i\right]$ follows from the Law of Iterated Expectations.

Presume that the researcher who aims at predicting Y_i using X_i is unaware of the true form of the CEF. Conscious of the result from Theorem 7.2, the researcher sets out to establish an optimal linear predictor which incorporates a constant term, as follows:

$$p_y\left(X_i\right) = \beta_0^* + \beta_1^* X_i$$

such that:

$$\left(\beta_0^*, \beta_1^*\right) \in \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\arg\min}\ \mathbb{E}\left[\left(Y_i - \beta_0 - \beta_1 X_i\right)^2\right]$$

as in (7.12). In order to find the coefficients of the optimal linear predictor, it is necessary to optimize the above MSE. The First Order Conditions are:

$$\begin{aligned}\mathbb{E}\left[Y_i - \beta_0^* - \beta_1^* X_i\right] &= 0\\ \mathbb{E}\left[X_i\left(Y_i - \beta_0^* - \beta_1^* X_i\right)\right] &= 0\end{aligned}$$

which are identical to the two equations (3.8)-(3.9) that determine the coefficients of the bivariate linear regression model from Example 3.11! Therefore, the solution can be expressed as follows, after some manipulation.

$$\begin{aligned}\beta_0^* &= \mathbb{E}\left[Y_i\right] - \frac{\mathbb{E}\left[X_i Y_i\right] - \mathbb{E}\left[X_i\right]\mathbb{E}\left[Y_i\right]}{\mathbb{E}\left[X_i^2\right] - \left(\mathbb{E}\left[X_i\right]\right)^2} \cdot \mathbb{E}\left[X_i\right]\\ \beta_1^* &= \frac{\mathbb{E}\left[X_i Y_i\right] - \mathbb{E}\left[X_i\right]\mathbb{E}\left[Y_i\right]}{\mathbb{E}\left[X_i^2\right] - \left(\mathbb{E}\left[X_i\right]\right)^2}\end{aligned}$$

Observe that under the hypothesis of a quadratic CEF, the moments of the form $\mathbb{E}\left[X_i^r Y_i\right]$ – for any nonnegative integer r – can be obtained easily; in this specific case it is:

$$\begin{aligned}\mathbb{E}\left[X_i^r Y_i\right] &= \mathbb{E}\left[\mathbb{E}\left[X_i^r Y_i \middle| X_i\right]\right]\\ &= \mathbb{E}\left[X_i^r \cdot \mathbb{E}\left[Y_i \middle| X_i\right]\right]\\ &= \mathbb{E}\left[X_i^r\left(X_i - \frac{1}{10}X_i^2\right)\right]\\ &= \mathbb{E}\left[X_i^{r+1}\right] - \frac{1}{10}\mathbb{E}\left[X_i^{r+2}\right]\end{aligned}$$

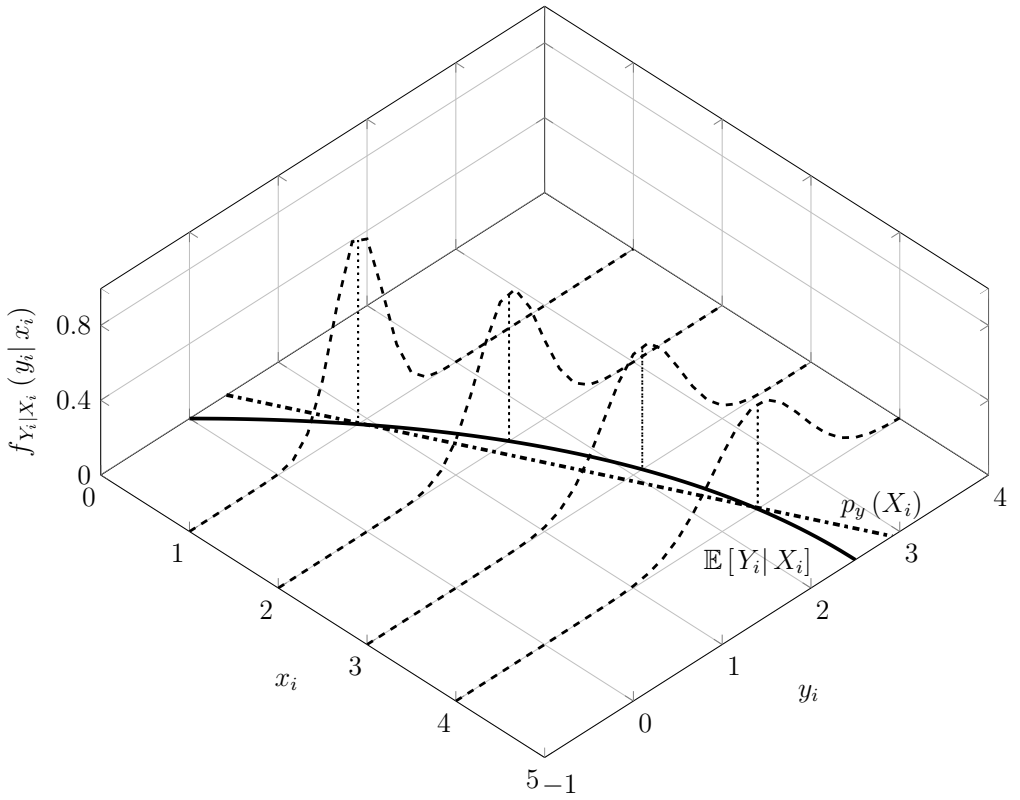
which is yet another application of the Law of Iterated Expectations. Therefore, in this specific example the two coefficients β_0^* and β_1^* are ultimately functions of some uncentered moments of X_i; as a consequence, in order to calculate the optimal linear predictor one must know the distribution of X_i (or make assumptions about it). To simplify, suppose that $X_i \sim \mathcal{U}\left[0, 5\right]$; in this case it is easy to see that:

$$\mathbb{E}\left[X_i^r\right] = \frac{5^r}{r+1}$$

for any nonnegative integer r. As one can autonomously verify, the combination of all these hypotheses implies the following optimal linear predictor.

$$p_y(X_i) = \frac{5}{12} + \frac{1}{2}X_i$$

Note that one would obtain a different optimal linear predictor if X_i were to follow a different distribution, including say a uniform distribution with a different support!



Note: the continuous curve is the CEF: $\mathbb{E}[Y_i|X_i] = X_i - X_i^2/10$; the dash-dotted line is the optimal linear predictor $p_y(X_i) = 5/12 + X_i/2$. The conditional distribution $Y_i|X_i$ is normal, with parameters that vary as a function of X_i. Selected density functions of $Y_i|X_i$ are displayed for $x_i = \{1, 2, 3, 4\}$.

Figure 7.3: The optimal linear predictor approximating a quadratic CEF

The result is illustrated graphically in Figure 7.3 above, where the continuous curve represents the quadratic CEF while the dash-dotted (straight) line is the optimal linear predictor, which is at the same time the best linear approximation of the quadratic CEF. To help visualize the random nature of the relationship between Y_i and X_i, the conditional distribution of the former given the latter is displayed as normal, but the analysis developed in this example does not depend on this specific coincidence. ■

7.3 Analysis of Least Squares

The analysis of optimal predictors under specific loss functions is admittedly quite abstract – it is grounded in statistical decision theory. However, it is useful to motivate the **Least Squares** criterion and the associated statistical estimators on sound theoretical bases. By the **analogy principle**, in fact, one can establish an appropriate *sample* version of the *population* optimal linear predictor problem under a quadratic loss function. The Least Squares problem applied to a sample $\{(y_i, \mathbf{x}_i)\}_{i=1}^N$ is:

$$\mathbf{b} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^K}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} \right)^2 \tag{7.17}$$

or equivalently, using compact matrix notation, as follows.

$$\mathbf{b} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^K}{\arg\min} \frac{1}{N} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right)^{\mathrm{T}} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right) \tag{7.18}$$

In a linear framework, the Least Squares problem is about finding a vector of coefficients $\mathbf{b}$ that minimizes the sum of the quadratic deviations between the dependent variable y_i and the corresponding linear combination of the independent variables $\mathbf{x}_i$ of each observation in the sample. Note that the N^{-1} factor is redundant towards the determination of the solution.

The K First Order Conditions of the problem (7.17), also called **normal equations**, are expressed below.

$$-\frac{2}{N} \sum_{i=1}^{N} \mathbf{x}_i \left(y_i - \mathbf{x}_i^{\mathrm{T}} \mathbf{b} \right) = \mathbf{0} \tag{7.19}$$

In analogy with the population optimal linear predictor, a unique solution $\mathbf{b}$ exists if the $K \times K$ matrix $\sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}$ is **invertible**, and such a solution reads as follows.

$$\mathbf{b} = \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{x}_i y_i \right) \tag{7.20}$$

The Least Squares solution $\mathbf{b}$ is perhaps more elegantly expressed by using compact matrix notation: in this case, the K normal equations would read as:

$$-\frac{2}{N} \mathbf{X}^{\mathrm{T}} \left(\mathbf{y} - \mathbf{X}\mathbf{b} \right) = \mathbf{0} \tag{7.21}$$

while the solution is written as follows.

$$\mathbf{b} = \left(\mathbf{X}^{\mathrm{T}} \mathbf{X} \right)^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{y} \tag{7.22}$$

A careful reader will have noted that the derivation of the Least Squares solution bears many analogies with Method of Moments estimation. For the moment, however, it is better to abstract from any statistical assumptions that might lead to make statements about estimation. A more immediately useful exercise is to rather familiarize with both the analytic vector notation (based on scalars like y_i and vectors like $\mathbf{x}_i$) and compact matrix notation. In fact, both are useful in their own right: while the former provides more visual information about certain computational details, the latter is better suited to synthetically express some more convoluted formulae. For a start, one should understand that the two $K \times K$ matrices $\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}$ and $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ are in fact the same thing!

This rest of this section illustrates some central algebraic and geometric properties that are typical of the Least Squares solution, and it culminates with a fundamental result known as the Frisch-Waugh-Lovell Theorem. All these properties aid the interpretation of Least Squares in practical applications. Before proceeding, some more definitions are in order. Let

$$\widehat{y}_i \equiv \mathbf{x}_i^{\mathrm{T}} \mathbf{b} \tag{7.23}$$

be the **fitted value** for the i-th observation, that is the value of the dependent variable that corresponds to $\mathbf{x}_i$ in the hyperplane implied by the Least Squares solution. Clearly, since the observations of y_i incorporate random, unobserved factors, they do not generally coincide with $\widehat{y}_i$. For each observation in the sample, the difference between the actual observation y_i of the dependent variable and the associated fitted value $\widehat{y}_i$ is called the **residual**:

$$\begin{aligned} e_i &\equiv y_i - \widehat{y}_i \\ &= y_i - \mathbf{x}_i^{\mathrm{T}} \mathbf{b} \end{aligned} \tag{7.24}$$

note that the Least Squares problem can be equivalently expressed as that of minimizing the sum of the squared residuals (hence its name).

One can vertically stack both sample fitted values and residuals so to adapt them to the convenient use of compact matrix notation.

$$\widehat{\mathbf{y}} = \begin{bmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ \vdots \\ \widehat{y}_N \end{bmatrix} \qquad\qquad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}$$

The vector of fitted values can be expressed compactly as:

$$\begin{aligned} \widehat{\mathbf{y}} &= \mathbf{X}\mathbf{b} \\ &= \mathbf{X} \left(\mathbf{X}^{\mathrm{T}} \mathbf{X} \right)^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{y} \\ &= \mathbf{P}_{\mathbf{X}} \mathbf{y} \end{aligned}$$

where:

$$\mathbf{P_X} \equiv \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} \tag{7.25}$$

is called the **projection matrix**, which if pre-multiplied to $\mathbf{y}$ results in the vector of fitted values $\widehat{\mathbf{y}}$. Furthermore:

$$\begin{aligned}\mathbf{e} &= \mathbf{y} - \widehat{\mathbf{y}} \\ &= \mathbf{y} - \mathbf{X}\mathbf{b} \\ &= (\mathbf{I} - \mathbf{P_X})\,\mathbf{y} \\ &= \mathbf{M_X}\mathbf{y}\end{aligned}$$

where:

$$\mathbf{M_X} \equiv \mathbf{I} - \mathbf{P_X} = \mathbf{I} - \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} \tag{7.26}$$

is the so-called **residual maker** matrix. Pre-multiplying $\mathbf{y}$ by the residual maker matrix clearly results in the vector of residuals $\mathbf{e}$.

The projection and residual maker matrices have important properties. They are both **symmetric**:

$$\begin{aligned}\mathbf{P_X} &= \mathbf{P_X^{\mathrm{T}}} \\ \mathbf{M_X} &= \mathbf{M_X^{\mathrm{T}}}\end{aligned}$$

idempotent:[4]

$$\begin{aligned}\mathbf{P_X}\mathbf{P_X} &= \mathbf{P_X} \\ \mathbf{M_X}\mathbf{M_X} &= \mathbf{M_X}\end{aligned}$$

and they are **orthogonal** to one another.

$$\mathbf{P_X}\mathbf{M_X} = \mathbf{M_X}\mathbf{P_X} = \mathbf{0}$$

In addition, it is easy to see that:

$$\begin{aligned}\mathbf{P_X}\mathbf{X} &= \mathbf{X} \\ \mathbf{M_X}\mathbf{X} &= \mathbf{0}\end{aligned}$$

with a straightforward interpretation: if one projects the columns of $\mathbf{X}$ onto themselves, the projection is identical to $\mathbf{X}$ and the residuals, consequently, are zero. Finally, observe that:

$$\begin{aligned}\mathbf{y} &= (\mathbf{I} + \mathbf{P_X} - \mathbf{P_X})\,\mathbf{y} \\ &= \mathbf{P_X}\mathbf{y} + \mathbf{M_X}\mathbf{y} \\ &= \widehat{\mathbf{y}} + \mathbf{e}\end{aligned}$$

[4]Note that $\mathbf{P_X}\mathbf{P_X} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} = \mathbf{P_X}$ and $\mathbf{M_X}\mathbf{M_X} = (\mathbf{I} - \mathbf{P_X})(\mathbf{I} - \mathbf{P_X}) = \mathbf{I} - 2\mathbf{P_X} + \mathbf{P_X}\mathbf{P_X} = \mathbf{I} - \mathbf{P_X} = \mathbf{M_X}$. The other results follow easily from these observations.

and:

$$\begin{aligned}\widehat{\mathbf{y}}^{\mathrm{T}}\mathbf{e} &= \mathbf{y}^{\mathrm{T}}\mathbf{P}_{\mathbf{X}}\mathbf{M}_{\mathbf{X}}\mathbf{y} \\ = \mathbf{e}^{\mathrm{T}}\widehat{\mathbf{y}} &= \mathbf{y}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}}\mathbf{P}_{\mathbf{X}}\mathbf{y} \\ &= 0\end{aligned}$$

that is, the decomposition of the vector $\mathbf{y}$ between the fitted values $\widehat{\mathbf{y}}$ and the residuals $\mathbf{e}$ is such that these two components are **orthogonal** to one another. This fact relates to the all-important **geometric interpretation** of the Least Squares solution $\mathbf{b}$, seen as the vector which, through the linear combination $\widehat{\mathbf{y}} = \mathbf{P}_{\mathbf{X}}\mathbf{y} = \mathbf{X}\mathbf{b}$, results in the **geometrical projection** of $\mathbf{y}$ onto the column space of $\mathbf{X}$.[5] In fact, inspecting the normal equations (7.19) or (7.21) reveals how the residual vector $\mathbf{e}$ is by construction orthogonal to the space $\mathcal{S}(\mathbf{X})$ spanned by the columns of $\mathbf{X}$ (the K explanatory variables). This is graphically represented in Figure 7.4 for the case of $K = 2$.

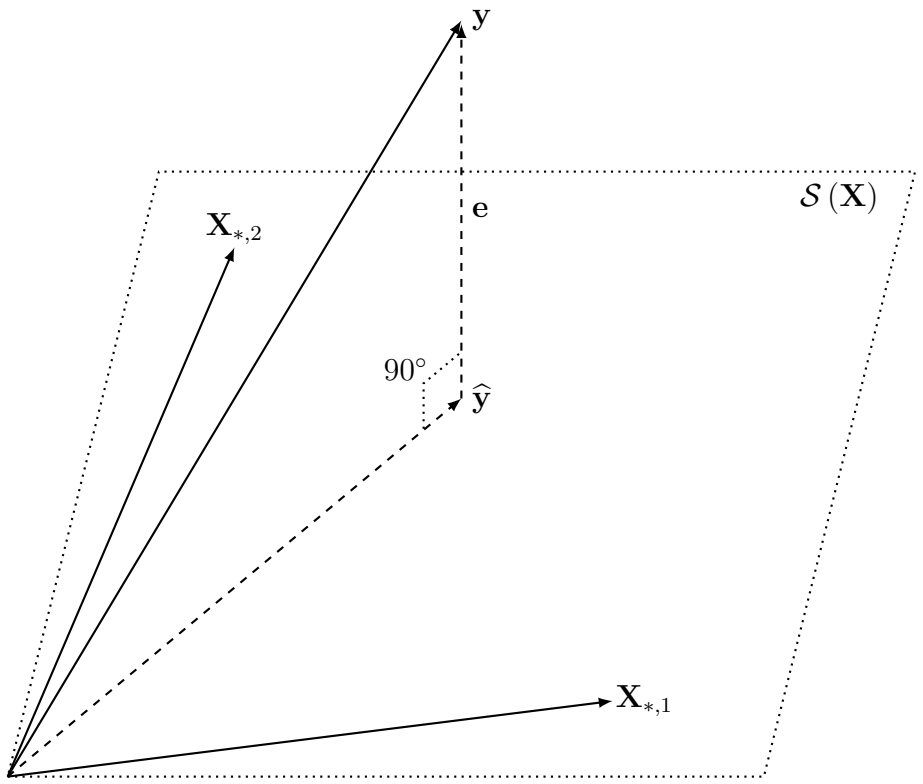


Figure 7.4: The geometric interpretation of the Least Squares solution

[5]Hence names such as "projection matrix" and "linear projection."

In order to appreciate some properties of the Least Squares solution that are especially relevant in settings with multiple explanatory variables ($K > 1$), it is useful to split these into smaller subsets. For example, one could rewrite the linear relationship (7.8) as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \tag{7.27}$$

where $\mathbf{X} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix}$ and $\boldsymbol{\beta}^{\mathrm{T}} = \begin{bmatrix}\boldsymbol{\beta}_1^{\mathrm{T}} & \boldsymbol{\beta}_2^{\mathrm{T}}\end{bmatrix}$. This amounts to "partition" the coefficient vector $\boldsymbol{\beta}$ in two smaller subvectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ respectively of length K_1 and K_2 (with $K_1 + K_2 = K$), each pertaining to a corresponding subset of explanatory variables. It is interesting to examine how the partitioned components of the Least Squares solution $\mathbf{b}$ compare to one another. To this end, rewrite the normal equations in (7.21) as:

$$\begin{bmatrix}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1 & \mathbf{X}_1^{\mathrm{T}}\mathbf{X}_2 \\ \mathbf{X}_2^{\mathrm{T}}\mathbf{X}_1 & \mathbf{X}_2^{\mathrm{T}}\mathbf{X}_2\end{bmatrix}\begin{bmatrix}\mathbf{b}_1 \\ \mathbf{b}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1^{\mathrm{T}}\mathbf{y} \\ \mathbf{X}_2^{\mathrm{T}}\mathbf{y}\end{bmatrix} \tag{7.28}$$

where $\mathbf{b}^{\mathrm{T}} = \begin{bmatrix}\mathbf{b}_1^{\mathrm{T}} & \mathbf{b}_2^{\mathrm{T}}\end{bmatrix}$. A fundamental result then follows.

Theorem 7.3. Frisch-Waugh-Lovell Theorem. *The solution for* $\mathbf{b}_2$ *can be written as:*

$$\mathbf{b}_2 = \left(\mathbf{X}_2^{*\mathrm{T}}\mathbf{X}_2^{*}\right)^{-1}\mathbf{X}_2^{*\mathrm{T}}\mathbf{y} \tag{7.29}$$

where:

$$\mathbf{X}_2^{*} \equiv \mathbf{M}_{\mathbf{X}1}\mathbf{X}_2$$

and $\mathbf{M}_{\mathbf{X}1}$ *is the residual maker matrix of* $\mathbf{X}_1$*.*

$$\mathbf{M}_{\mathbf{X}1} \equiv \mathbf{I} - \mathbf{X}_1\left(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathrm{T}}$$

Furthermore, a symmetrical result is obtained for $\mathbf{b}_1$*.*

Proof. By the algebra of partitioned matrices, one can write $\mathbf{b}_1$ as a function of $\mathbf{b}_2$ as:

$$\mathbf{b}_1 = \left(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{y} - \left(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_2\mathbf{b}_2$$

plugging the above in the lower block of K_2 rows in (7.28) gives:

$$\mathbf{X}_2^{\mathrm{T}}\mathbf{X}_1\left(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{y} - \mathbf{X}_2^{\mathrm{T}}\mathbf{X}_1\left(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_2\mathbf{b}_2 + \mathbf{X}_2^{\mathrm{T}}\mathbf{X}_2\mathbf{b}_2 = \mathbf{X}_2^{\mathrm{T}}\mathbf{y}$$

with solution:

$$\begin{aligned}\mathbf{b}_2 &= \left[\mathbf{X}_2^{\mathrm{T}}\left(\mathbf{I} - \mathbf{X}_1\left(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathrm{T}}\right)\mathbf{X}_2\right]^{-1}\left[\mathbf{X}_2\left(\mathbf{I} - \mathbf{X}_1\left(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathrm{T}}\right)\mathbf{y}\right] \\ &= \left(\mathbf{X}_2^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{X}_2\right)^{-1}\mathbf{X}_2^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{y}\end{aligned}$$

which is equivalent to (7.29) since $\mathbf{M}_{\mathbf{X}1}$ is symmetric and idempotent. The result for $\mathbf{b}_1$ is symmetrical. □

While this theorem might, at a first glance, look like a bunch of trivial if nasty-looking algebraic formulas, it does deliver quite a fundamental insight: any component ($\mathbf{b}_2$) of the least squares solution is *algebraically equivalent* to another least squares solution, which follows from a transformed model where the explanatory variables in question ($\mathbf{X}_2$) are substituted with the corresponding residuals ($\mathbf{X}_2^* = \mathbf{M}_{\mathbf{X}1}\mathbf{X}_2$) that are obtained from projecting them on the other explanatory variables ($\mathbf{X}_1$). Recall our earlier observation that the least squares projection returns a vector of fitted values and a vector of residuals that are reciprocally orthogonal. What the Frisch-Waugh-Lovell Theorem means in this framework is that in a linear model with multiple explanatory variables, each coefficient b_k obtained via Least Squares can be interpreted as the overall "contribution" of X_{ki} to Y_i, *after the contributions of the other* $K-1$ *explanatory variables to* Y_i *has been netted out* or, using more technical terminology, *partialled out*.

This property explains by a great deal the immense popularity of statistical estimators based on the Least Squares principle in econometric analysis. In fact, it allows researchers to interpret the estimated coefficients associated with a single socio-economic variable of interest by pretending that all other variables included in the model are taken "as given," corresponding with the typical *ceteris paribus* type of scientific thought experiments. Note that the theorem does *not* motivate the exclusion of relevant explanatory variables from the analysis, except for cases when they can be confidently assumed to be statistically unrelated to the variables of interest (say, $\mathbf{X}_2$). Notice, in fact, that only if $\mathbf{X}_1$ and $\mathbf{X}_2$ are orthogonal ($\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_2 = \mathbf{0}$) it holds that $\mathbf{X}_2^* = \mathbf{M}_{\mathbf{X}1}\mathbf{X}_2 = \mathbf{X}_2$, meaning that the least squares coefficients associated with the $\mathbf{X}_2$ explanatory variables are identical whether one includes the remaining factors $\mathbf{X}_1$ in the model or not.

The properties of partitioned Least Squares appear particularly powerful when considering any partitioned model with $K_1 = K-1$ and $K_2 = 1$. In this case, let $\mathbf{X}_2 = \mathbf{s}$; the K-th Least Squares coefficient is as follows.

$$b_K = \frac{\mathbf{s}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{y}}{\mathbf{s}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{s}} \tag{7.30}$$

This quantity is related to a statistical object called the **partial correlation coefficient** ρ_{YS}^* between variables Y_i and $X_{Ki} = S_i$:

$$\rho_{YS}^* = \frac{\mathbf{s}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{y}}{\sqrt{\mathbf{s}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{s}}\sqrt{\mathbf{y}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{y}}} \tag{7.31}$$

which is nothing else but the correlation coefficient of the *residuals* of both Y_i and S_i, obtained by projecting these on the other $K-1$ explanatory

variables (including a constant term). This quantity can be shown to be the sample counterpart of the *partial correlation*, a variation of the population correlation expressed in terms of conditional moments:

$$\mathbb{C}\text{orr}\left[Y_i, S_i \middle| \boldsymbol{x}_1\right] = \frac{\mathbb{C}\text{ov}\left[Y_i, S_i \middle| \boldsymbol{x}_1\right]}{\sqrt{\mathbb{V}\text{ar}\left[Y_i \middle| \boldsymbol{x}_1\right]}\sqrt{\mathbb{V}\text{ar}\left[S_i \middle| \boldsymbol{x}_1\right]}} \tag{7.32}$$

where $\boldsymbol{x}_1 = \left(X_{1i}, \ldots, X_{(K-1)i}\right)$. Intuitively, partial correlation coefficients measure the correlation between two variables once the dependence of both (in the linear projection sense) from other variables has been removed. The algebraic relationship between (7.30) and (7.31) appears evident, which explains why Least Squares coefficients are often attributed an interpretation in terms of partial correlation (although the two are not identical).[6]

Another oft-invoked application of the Frisch-Waugh-Lovell Theorem is *demeaning*, that is the operation of subtracting the respective mean from both the explanatory and dependent variables. Suppose that $\mathbf{X}_1 = \iota$ is the "constant term" of the model (an N-sized vector of ones), so that $\mathbf{X}_2$ entails $K-1$ columns like in model (7.3). In such a case, the residual maker matrix reads as:

$$\mathbf{D} \equiv \mathbf{M}_{\mathbf{X}1} = \mathbf{I} - \iota\left(\iota^{\mathrm{T}}\iota\right)^{-1}\iota^{\mathrm{T}} = \mathbf{I} - \frac{1}{N}\iota\iota^{\mathrm{T}} \tag{7.33}$$

hence for any vector $\mathbf{a}$ of length N, it is:

$$\mathbf{D}\mathbf{a} = \mathbf{a} - \bar{a}\iota$$

where $\bar{a} \equiv \frac{1}{N}\sum_{i=1}^{N} a_i$. Thus $\mathbf{b}_2$ can be equivalently obtained as the solution of a Least Squares problem applied to the *demeaned model*:

$$y_i - \bar{y} = \beta_1\left(x_{i1} - \bar{x}_1\right) + \cdots + \beta_{(K-1)}\left(x_{i(K-1)} - \bar{x}_{(K-1)}\right) + \varepsilon_i \tag{7.34}$$

where $\bar{y} \equiv \frac{1}{N}\sum_{i=1}^{N} y_i$ and $\bar{x}_k \equiv \frac{1}{N}\sum_{i=1}^{N} x_{ik}$ for all $k = 1, \ldots, K-1$. This fact is exploited in linear models for panel data, which routinely include a separate intercept (a *fixed effect*) for each panel unit in the sample.

[6] This observation resonates well with the previous discussion of the bivariate regression model, the relationship between the correlation coefficient between any two random variables Y_i and X_i and their corresponding regression slope, and the MM/OLS estimator for the latter (examples 3.11 and 5.4). In fact, in the sample the *observed* correlation coefficient between Y_i and X_i is defined as:

$$\rho_{YS} = \frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}} = \frac{\mathbf{x}^{\mathrm{T}}\mathbf{D}\mathbf{y}}{\sqrt{\mathbf{x}^{\mathrm{T}}\mathbf{D}\mathbf{x}}\sqrt{\mathbf{y}^{\mathrm{T}}\mathbf{D}\mathbf{y}}}$$

where $\bar{x} \equiv \frac{1}{N}\sum_{i=1}^{N} x_i$, $\bar{y} \equiv \frac{1}{N}\sum_{i=1}^{N} y_i$, while $\mathbf{D}$ is defined as in (7.33); the analogies of the above expression with both (5.16) and (7.31) are obvious.

7.4 Evaluation of Least Squares

As mentioned, in actual applications of Least Squares researchers typically include a constant term into their linear specifications. Two motivations for this choice have already been introduced: first, the inclusion of a constant term allows to confidently assume that $\mathbb{E}[\varepsilon_i] = 0$, which greatly simplifies the statistical modeling of econometric estimators; second, it lets interpret the calculated coefficients in terms of partial correlations. Some additional important implications are apparent from the inspection of the first normal equation in (7.19); in particular:

1. all the residuals sum up to zero: $\sum_{i=1}^{N} e_i = 0$, and so...
2. ...the mean of the fitted values $\widehat{y}_i$ coincides with that of dependent variable y_i: $\frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i^{\mathrm{T}}\mathbf{b} = \frac{1}{N}\sum_{i=1}^{N} \widehat{y}_i = \bar{y}$, and...
3. ...the point $(\bar{y}, \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_K)$ lies on the $p(\mathbf{x}_i) = \mathbf{x}_i^{\mathrm{T}}\mathbf{b}$ hyperplane.

Thanks to the above properties, the inclusion of a constant term into the model allows to employ a common criterion for the evaluation of the Least Squares' *goodness of fit*, defined as the extent by which the linear combination $\widehat{y}_i = \mathbf{x}_i^{\mathrm{T}}\mathbf{b}$ *explains*, in a statistical sense, the variation of the dependent variable y_i. This criterion is called **coefficient of determination** R^2 and is defined as:

$$\mathrm{R}^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = \frac{\sum_{i=1}^{N}(\widehat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \in [0, 1] \tag{7.35}$$

where the term in the numerator relates to the variance of the fitted values (note that $\frac{1}{N}\sum_{i=1}^{N} \widehat{y}_i = \bar{y}$ because of the inclusion of a constant term):

$$\mathrm{ESS} \equiv \sum_{i=1}^{N}(\widehat{y}_i - \bar{y})^2 = \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X}\mathbf{b}$$

the above is called **Explained Sum of Squares** (ESS). The expression in the denominator, on the other hand, corresponds with the overall *empirical* variance of Y_i (that is, the sample variance of the observations y_i):

$$\mathrm{TSS} \equiv \sum_{i=1}^{N}(y_i - \bar{y})^2 = \mathbf{y}^{\mathrm{T}}\mathbf{D}\mathbf{y}$$

and is called instead **Total Sum of Squares** (TSS). The difference between these two quantities is called **Residual Sum of Squares** (RSS).

$$\mathrm{RSS} \equiv \mathrm{TSS} - \mathrm{ESS} = \sum_{i=1}^{N} e_i^2 = \mathbf{e}^{\mathrm{T}}\mathbf{e}$$

To see that the RSS equals the sum of the squared residuals, observe first that with the inclusion of a constant term into the model the mean of the residuals themselves is zero, hence $\mathbf{De} = \mathbf{e}$ and

$$\begin{aligned}\underbrace{\mathbf{y}^{\mathrm{T}}\mathbf{D}\mathbf{y}}_{=\mathrm{TSS}} &= (\mathbf{X}\mathbf{b}+\mathbf{e})^{\mathrm{T}}\,\mathbf{D}\,(\mathbf{X}\mathbf{b}+\mathbf{e}) \\ &= \underbrace{\mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X}\mathbf{b}}_{=\mathrm{ESS}} + \underbrace{\mathbf{e}^{\mathrm{T}}\mathbf{e}}_{=\mathrm{RSS}}\end{aligned}$$

follows from the fact that $\widehat{\mathbf{y}} = \mathbf{X}\mathbf{b}$ and $\mathbf{e}$ are orthogonal by construction.[7] Consequently, the coefficient of determination R^2 can also be written as:

$$\mathrm{R}^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\sum_{i=1}^{N} e_i^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2} \in [0,1]$$

intuitively, this coefficient is close to 1 if the projection explains the overwhelming majority of the variation in Y_i, while it is close to 0 in the opposite case where the explanatory variables relate to only a small portion of it. To better appreciate this, observe that

$$\begin{aligned}\mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X}\mathbf{b} &= \widehat{\mathbf{y}}^{\mathrm{T}}\mathbf{D}\widehat{\mathbf{y}} \\ &= \widehat{\mathbf{y}}^{\mathrm{T}}\mathbf{D}\,(\mathbf{y}+\mathbf{e}) \\ &= \widehat{\mathbf{y}}^{\mathrm{T}}\mathbf{D}\mathbf{y}\end{aligned}$$

and therefore:

$$\mathrm{R}^2 = \frac{\widehat{\mathbf{y}}^{\mathrm{T}}\mathbf{D}\widehat{\mathbf{y}}}{\mathbf{y}^{\mathrm{T}}\mathbf{D}\mathbf{y}} = \frac{\widehat{\mathbf{y}}^{\mathrm{T}}\mathbf{D}\mathbf{y}}{\mathbf{y}^{\mathrm{T}}\mathbf{D}\mathbf{y}} \cdot \underbrace{\frac{\widehat{\mathbf{y}}^{\mathrm{T}}\mathbf{D}\mathbf{y}}{\widehat{\mathbf{y}}^{\mathrm{T}}\mathbf{D}\widehat{\mathbf{y}}}}_{=1} = \frac{\widehat{\mathbf{y}}^{\mathrm{T}}\mathbf{D}\mathbf{y}}{\mathbf{y}^{\mathrm{T}}\mathbf{D}\mathbf{y}} =$$

$$= \frac{\left[\sum_{i=1}^{N}(y_i-\bar{y})(\widehat{y}_i-\bar{y})\right]^2}{\left[\sum_{i=1}^{N}(\widehat{y}_i-\bar{y})^2\right]\left[\sum_{i=1}^{N}(y_i-\bar{y})^2\right]}$$

that is, the R^2 coefficient is equal to the square of the correlation coefficient between y_i and the fitted values $\widehat{y}_i$ (hence its name).

[7] Another way to see this is:

$$\sum_{i=1}^{N}(y_i-\bar{y})^2 = \sum_{i=1}^{N}(y_i-\widehat{y}_i+\widehat{y}_i-\bar{y})^2 = \sum_{i=1}^{N}e_i^2 + \sum_{i=1}^{N}(\widehat{y}_i-\bar{y})^2$$

noting that $\sum_{i=1}^{N}(y_i-\widehat{y}_i)(\widehat{y}_i-\bar{y}) = \sum_{i=1}^{N}(y_i-\widehat{y}_i)\,\widehat{y}_i = \sum_{k=1}^{K} b_k \sum_{i=1}^{N} x_{ik}(y_i-\widehat{y}_i) = 0$ follows from the normal equations.

A couple of warnings about the interpretation of R^2 in practical applications are in order. First, this coefficient is not an absolute measure of a projection's overall "quality." In fact, the size of the variances of both the dependent and explanatory variables – as well as the statistical relationship between those – are specific to every particular empirical setting. Second, one must be careful even when comparing the R^2 from different projections applied to the same setting or even dataset. The reason is that as it is easy to see, this coefficient increases *mechanically* with the inclusion of each additional explanatory variable into the model.[8] A measure that takes the last observation into account is the **adjusted** R^2 coefficient, written as:

$$\begin{aligned} \overline{R}^2 &= 1 - \frac{\sum_{i=1}^{N} e_i^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \frac{N-1}{N-K} \\ &= 1 - \left(1 - R^2\right) \frac{N-1}{N-K} \end{aligned} \tag{7.36}$$

clearly, this variation of the coefficient of determination includes a term for the total number of explanatory variables included into the model; if the contribution of one of these variables towards the explained sum of squares is negligible, the adjusted R^2 might decrease or even turn negative. Finally, it should go without saying that all the observations and interpretation of the R^2 coefficient are only valid if the linear model features a constant term. The R^2 coefficient can be computed by computer packages even for models lacking a constant term; however, since the calculated residuals would not sum up to zero, the resulting coefficient cannot be compared to that from a model featuring a constant (again, the calculated R^2 can be negative).

Some of the results and intuitions developed with the support of linear algebra can be better appreciated by putting them in practical context. It is thus helpful to review the initial examples of linear economic relationships.

Example 7.4. The Keynesian consumption function, revisited. A simple linear fit of the relationship from example (7.1) is displayed in Figure 7.5. The slope of the line, which is meant to evaluate the parameter for the marginal propensity of consumption c_1, is calculated to be about $c_1 \simeq 0.80$. In this particular case, the R^2 coefficient is, staggeringly, virtually equal to 1! However, nowadays macroeconomists give little weight a result like this.

[8] It can be shown that this increase is a function of the partial correlation ρ^*_{YS} between the dependent variable Y and some given newly added variable S:

$$R_1^2 = R_0^2 + \left(1 - R_0^2\right) \left(\rho^*_{YS}\right)^2$$

where R_0^2 and R_1^2 are the two coefficients of determination calculated respectively prior to and posterior to the inclusion of S.

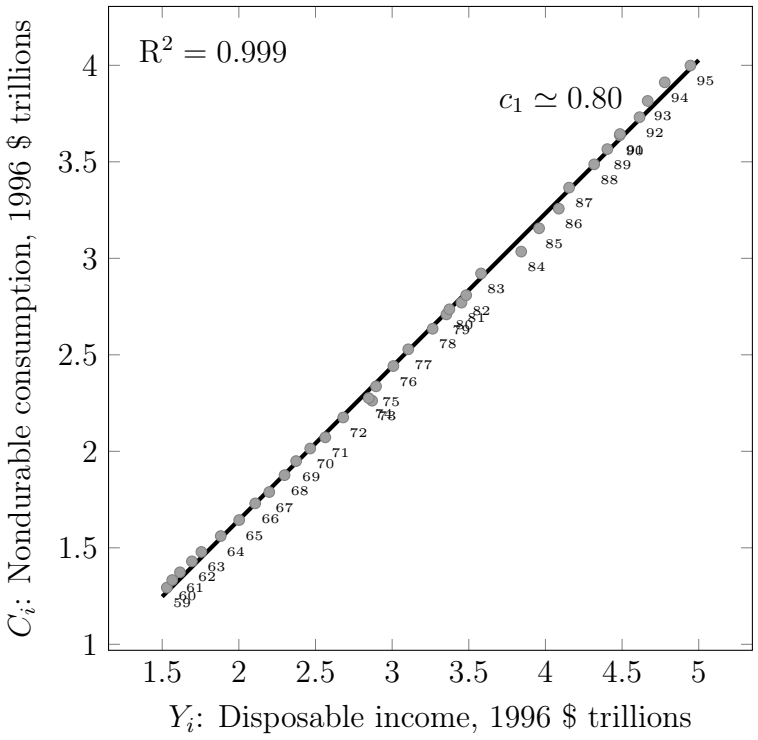


Figure 7.5: Fitted Keynesian Consumption Function

In fact, econometric research has shown that strong linear fits of this sort are standard properties of macroeconomic time series, which intuitively are due to the *co-movement* of variables because of some other factors that are possibly unaccounted by the model. Therefore, the estimate $c_1 \simeq 0.80$ can hardly be interpreted as the average increase in aggregate consumption that follows from the increase of a country's GDP. ■

Example 7.5. Human capital and wages, revisited. A simple linear fit of the relationship between the logarithmic wage and the education variable from example 7.2 returns a slope coefficient $b \simeq 0.064$ and an R^2 coefficient equal to 0.076 (see Figure 7.6, top panel). This does not imply that such a relationship is meaningless, quite the contrary! By enriching the model with a squared polynomial for the experience variable as in the proper Mincer equation (7.6), one would obtain a *higher* slope coefficient, up to $b' \simeq 0.105$; while the R^2 coefficient increases too, as expected, up to about 0.176. Note that here K is small relative to the sample size (the selected cross section of the original dataset has size $N = 1,241$), hence the *adjusted* R^2 is virtually identical to the standard R^2 in both calculations.

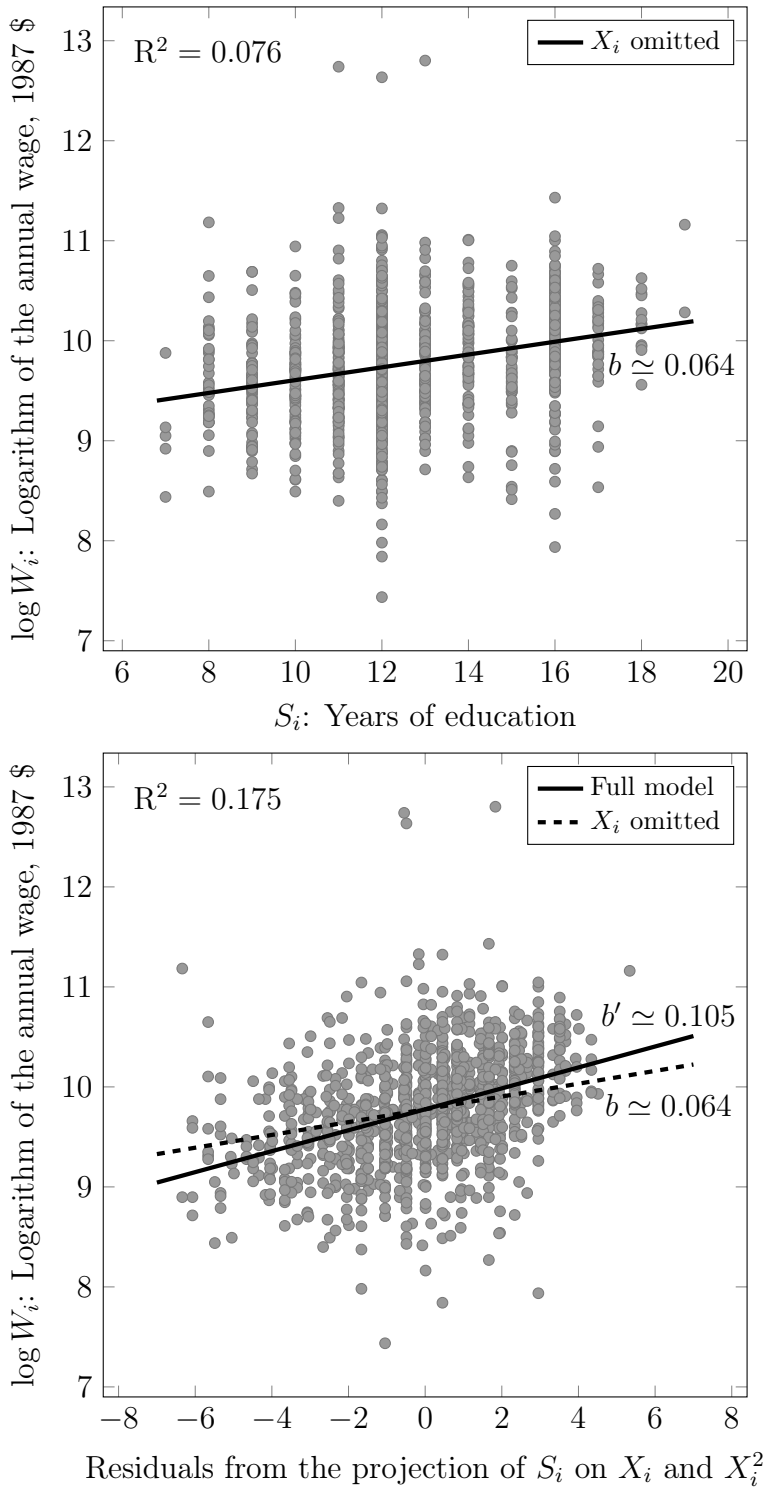


Figure 7.6: Fitted Mincer Equation, two versions

How are these changes in the output of Least Squares to be interpreted? First, one can visualize the Frisch-Waugh-Lovell theorem at work through the bottom panel of Figure 7.6, which represents the linear fit between log wages and the residuals obtained from projecting education on experience and its square (plus a constant vector ι); by *partialing out* the contribution of the polynomial for experience on logarithmic wages, the slope attributed to the education variable is to be interpreted in terms of partial correlation. The socio-economic intuition that explains the increase in the evaluated returns to schooling is that, without appropriately incorporating experience into the model, the coefficient for education is dragged down by the mechanical negative correlation between education and experience (by studying for longer, one enters the labor market later) and experience itself positively impacts upon labor market outcomes like wages. Second, the two measured values for R^2 suggest that the variation of individual wages is explained by several factors, of which both education and experience are two prominent ones; however, a large portion of this variation is due to other, idiosyncratic characteristics of individuals (like their ability, their personal connections, or simply their luck in life) that are unaccounted by the model. ■

Through these two examples it is possible to draw one final observation about the R^2 coefficient: while certainly useful to evaluate "goodness of fit," it can be a poor criterion for identifying socio-economic explanations about phenomena of interest. Instead, the issue of carefully selecting explanatory variables appears to be one of more immediate importance.

7.5 Least Squares and Linear Regression

Throughout this introduction to Least Squares, the magic word "regression" has never been used. In fact, the analysis of Least Squares has been largely treated within a merely algebraic framework, albeit initially motivated on some prediction problem. With the development of this framework at hand, it is more convenient to examine the implications of augmenting the postulated linear relationship with some **distributional assumptions**. Suppose that the CEF that generates the data is effectively linear:

$$\mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 \tag{7.37}$$

where $\boldsymbol{\beta}_0$ represents the supposed "true" vector of parameters from which the data are generated. A linear model like (7.1), when enriched with this hypothesis about the joint distribution of $(\boldsymbol{x}_i, Y_i)$, is called a **linear regression model**, $\boldsymbol{x}_i$ are called the **regressors**, and Y_i is called the **regressand**.

An implication of (7.37) is that if $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, the expectation of the error terms in (7.1), *conditional* on the explanatory variables $\boldsymbol{x}_i$, is zero:

$$\begin{aligned}\mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{x}_i\right] &= \mathbb{E}\left[Y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 \middle| \boldsymbol{x}_i\right] \\ &= \mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 \\ &= 0\end{aligned}$$

which, by the Law of Iterated Expectations, implies:

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{x}_i\varepsilon_i\right] &= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}\left[\boldsymbol{x}_i\varepsilon_i \middle| \boldsymbol{x}_i\right]\right] \\ &= \mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{x}_i \cdot \mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{x}_i\right]\right] \\ &= \mathbf{0}\end{aligned}$$

(the opposite is not true, that is, $\mathbb{E}\left[\boldsymbol{x}_i\varepsilon_i\right] = \mathbf{0}$ does *not* imply $\mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{x}_i\right] = 0$). By the standard properties of probability limits it can be shown that:[9]

$$\mathbf{b}_N = \left(\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i y_i \xrightarrow{p} \mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right]^{-1} \mathbb{E}\left[\boldsymbol{x}_i Y_i\right] = \boldsymbol{\beta}^*$$

so long as matrices $\frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}$ and $\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right]$ have full rank. This means that:

$$\mathbf{x}_i^{\mathrm{T}}\mathbf{b}_N \xrightarrow{p} \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* = p_y\left(\boldsymbol{x}_i = \mathbf{x}_i\right)$$

that is, the Least Squares projection converges in probability to the *optimal linear predictor* (7.12); this should not be too surprising, since such a result generally holds for sample analogs of population moments. What is relevant here is that under hypothesis (7.37), the optimal linear predictor *coincides with the* (*linear*) CEF for any given realization $\boldsymbol{x}_i = \mathbf{x}_i$:

$$\begin{aligned}p_y\left(\boldsymbol{x}_i = \mathbf{x}_i\right) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* &= \mathbf{x}_i^{\mathrm{T}}\,\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right]^{-1}\mathbb{E}\left[\boldsymbol{x}_i Y_i\right] \\ &= \mathbf{x}_i^{\mathrm{T}}\,\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right]^{-1}\mathbb{E}\left[\boldsymbol{x}_i\,\mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right]\right] \\ &= \mathbf{x}_i^{\mathrm{T}}\,\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right]^{-1}\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right]\boldsymbol{\beta}_0 \\ &= \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 \\ &= \mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i = \mathbf{x}_i\right]\end{aligned}$$

where the second line once again exploits the Law of Iterated Expectations. In light of Theorems 7.1 and 7.2, this result should not be too surprising.

[9] Compare with the analysis and the proof of consistency conducted for the bivariate case in Example 6.3, Lecture 6. Also note that the sequence $\mathbf{b}_N$, whose probability limit is taken, is defined here in terms of *realizations* $\mathbf{x}_i$ and y_i, This notation is conventional in the analysis of econometric estimators.

The implication of both observations is that if the CEF is linear, the corresponding Least Squares solution coincides asymptotically with the "true" parameters of the regression model.

$$\mathbf{b}_N \xrightarrow{p} \boldsymbol{\beta}_0 \tag{7.38}$$

This property motivates the use of Least Squares as a statistical or **econometric estimator** of the linear regression model, an estimator that takes the name of **Ordinary Least Squares** (**OLS**), where 'ordinary' is meant to distinguish the baseline estimator from its variations or extensions. Result (7.38) is re-framed later as the *consistency* property of the OLS estimator. In what follows, it is given an array of motivations for the use of the linear regression model in practical contexts.

Linear Regression, indeed

If the researcher can confidently assume that the CEF of the relationship under analysis is indeed linear, by (7.38) the Least Squares estimator for the linear regression model is the most natural choice. However, the researcher must be careful at correctly *specifying* the linear model, and include all the variables that might be correlated with the relevant explanatory variables of interest. In fact, if the true CEF is linear but unlike (7.37) it reads:

$$\mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i, \boldsymbol{s}_i\right] = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + \boldsymbol{s}_i^{\mathrm{T}}\boldsymbol{\delta}_0 \tag{7.39}$$

that is, it is only linear conditional on *some additional variables* $\mathbf{s}_i$ unaccounted by the researchers and that enter the CEF with associated "true" parameters $\boldsymbol{\delta}_0 \neq \mathbf{0}$, then the probability limit of Least Squares reads:

$$\begin{aligned}
\mathbf{b}_N &\xrightarrow{p} \mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right]^{-1}\mathbb{E}\left[\boldsymbol{x}_i Y_i\right] \\
&= \mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right]^{-1}\mathbb{E}\left[\boldsymbol{x}_i\,\mathbb{E}\left[Y_i\middle|\boldsymbol{x}_i,\boldsymbol{s}_i\right]\right] \\
&= \mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right]^{-1}\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + \boldsymbol{x}_i\boldsymbol{s}_i^{\mathrm{T}}\boldsymbol{\delta}_0\right] \\
&= \boldsymbol{\beta}_0 + \mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right]^{-1}\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{s}_i^{\mathrm{T}}\right]\boldsymbol{\delta}_0
\end{aligned}$$

and it coincides with $\boldsymbol{\beta}_0$ only if $\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{s}_i^{\mathrm{T}}\right] = 0$, that is, variables $\boldsymbol{x}_i$ and $\boldsymbol{s}_i$ are uncorrelated.[10] The intuition is better developed when $\boldsymbol{s}_i = S_i$ is a single variable (with associated parameter δ_0 in the CEF):

$$\mathbf{b}_N \xrightarrow{p} \boldsymbol{\beta}_0 + \delta_0\,\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right]^{-1}\mathbb{E}\left[\boldsymbol{x}_i S_i\right] \tag{7.40}$$

yielding the (in)famous formula for the **omitted variable bias**.

[10] Once again, the second line exploits the Law of Iterated Expectations, where the outer expectation in $\mathbb{E}\left[\boldsymbol{x}_i\,\mathbb{E}\left[Y_i\middle|\boldsymbol{x}_i,\boldsymbol{s}_i\right]\right]$ is taken with respect to both $\boldsymbol{x}_i$ and $\boldsymbol{s}_i$.

Expression (7.40) indicates that the omission of a relevant explanatory variable affects the probability limit of the calculated Least Squares coefficients, which are increased by a term that equals δ_0 multiplied by the population projection of S_i on $\boldsymbol{x}_i$. For instance, should experience be omitted from the Mincer Equation's estimated regression, this "bias term" would be negative, since while the contribution of experience to log wages is likely positive, its correlation with education is mechanically negative. Ability, on the other hand, is a difficult variable to observe, and its omission from the estimated Mincer Equation is likely to affect the education coefficient with a positive asymptotic bias, as more skilled individuals are likely to attain more education and also earn more money.

Non-Linear Models and Regression

Many relationships of interest between socio-economic variables are likely to be non-linear. However, the linear regression framework is flexible enough to accommodate many of these. In fact, so long a non-linear model can be **transformed** to be *linear in the parameters*, it can be treated econometrically like a linear model. This is better illustrated with some examples.

Example 7.6. A *log-lin* model. The model of human capital discussed in example 7.2 is non-linear in the variables, but it is easily shown that it is linear in the parameters. Taking logarithms on both sides of (7.5) returns the so-called Mincer equation (7.6), which is rewritten here for $\varepsilon_i \equiv \alpha_i + \epsilon_i$:

$$\log W_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 S_i + \varepsilon_i \tag{7.41}$$

which is a famous case of a **log-lin** model: that is, logarithmic on the left-hand side, linear on the right-hand side. In a log-lin model, the regression coefficients associated with the regressors entering linearly in the equation measure the **semi-elasticity** of the dependent variable relative to the regressor in question: specifically, the relative increase in the former following from a unitary increase in the latter. For example, the estimated coefficient for education $b' = 0.105$ from Figure 7.6 indicates that wages are expected to increase by about 10.5% if an individual acquires one additional year of education. It is quite easy to see that OLS would estimate the $\boldsymbol{\beta}$ parameters of the model consistently, so long as $\mathbb{E}[\varepsilon_i | X_i, S_i] = 0$ holds – an assumption that can be quite problematic in the analysis of the returns to education.

Example 7.7. A *log-log* model. A common model used by economists for describing the production process is the *Cobb-Douglas* production function:

$$Y_i = A_i K_i^{\beta_K} L_i^{\beta_L} \tag{7.42}$$

where Y_i is total output, K_i is the capital input, L_i is the labor input, A_i is total factor productivity, and the unit of analysis can range from plants to countries. Taking logarithms on both sides results in an adequately linear-in-the-parameters equation:

$$\log Y_i = \alpha + \beta_K \log K_i + \beta_L \log L_i + \omega_i \tag{7.43}$$

where $\log A_i = \alpha + \omega_i$: here ω_i is an unobserved *productivity shock*, with $\mathbb{E}[\omega_i] = 0$. Any transformed equation of this kind is called a **log-log** model (that is, logarithmic on both sides), and coefficients such as β_K and β_L are interpreted as **elasticities**, that is the ratio between the relative variations of the dependent variable and the associated regressor that is implied by the model. For example, $\beta_L = 0.6$ means that a 1% increase in the labor input induces, on average, a 0.6% increase in total output. Note that OLS would estimate the semi-elasticity parameters consistently under $\mathbb{E}[\omega_i | K_i, L_i] = 0$, which is another problematic assumption in econometric analysis.

In other cases, however, an adequate transformation is not possible.

Example 7.8. Distance Decay. Suppose one is interested in studying how *knowledge spillovers* generated by universities affect the productivity of local firms, and is modeling the phenomenon by augmenting the linearized Cobb-Douglas function (7.43) as follows:

$$\log Y_i = \alpha + \beta_K \log K_i + \beta_L \log L_i + \delta \exp(-\lambda D_i) U_i + \omega_i \tag{7.44}$$

where U_i is the size of some local university, while D_i is its distance from firm i. Here, parameter δ measures the semi-elasticity of firm productivity with respect to the university's size, weighted by the *distance decay* factor $\exp(-\lambda D_i)$ as parametrized by λ. In this model, the two parameters δ and λ enter non-linearly in the equation, and irreducibly so.

The linear regression model and Ordinary Least Squares cannot be applied to a structural equation such as (7.44); however, they can be adapted for this purpose. The **Non Linear Least Squares** (**NLLS**) estimator is often employed in order to address these types of problems, and is introduced in subsequent lectures. Other times, instead, a researcher is simply uncertain about the exact shape of the structural relationship and corresponding CEF that generate the data. In those circumstances, using a linear regression as the baseline model is most often a sensible idea: in fact, the Least Squares projection is known to converge in probability to the population projection, which by Theorem 7.2 is known to be the best approximation to the true CEF in the MSE sense. While this should not be used as an excuse to give up on the search for the right specification, starting the analysis of complex economic relationships with a good approximation may be a good start.

Groups and Dummies

The linear regression model is especially useful for handling *grouped* data, that is, samples that are drawn from populations which can be partitioned into identifiable sub-populations. If the distribution of the dependent variable Y_i is expected to be heterogeneous across such sub-populations, it is convenient to account for this heterogeneity in regression analysis; typically, this is done through devices known as **dummy variables**. The latter are simply explanatory variables that equal one ($D_i = 1$) if the i-th observation belongs to a specific group of interest, and zero ($D_i = 0$) otherwise.

To illustrate how dummy variables operate, consider the simple bivariate model

$$Y_i = \pi_0 + \pi_1 D_i + \eta_i \tag{7.45}$$

where Y_i is some outcome of interest, D_i is a dummy variable that identifies some group of interest (e.g. females, blacks, foreigners, young people) and η_i is an error term. The vector of Least Squares coefficients (p_0, p_1) which is obtained through a sample $\{y_i, d_i\}_{i=1}^N$ is calculated as:

$$\begin{aligned}
\begin{bmatrix} p_0 \\ p_1 \end{bmatrix} &= \begin{bmatrix} N & N_D \\ N_D & N_D \end{bmatrix}^{-1} \begin{bmatrix} N\bar{y} \\ N_D\bar{y}_D \end{bmatrix} \\
&= \frac{1}{N - N_D} \begin{bmatrix} N\bar{y} - N_D\bar{y}_D \\ -N\bar{y} + N\bar{y}_D \end{bmatrix} \\
&= \frac{1}{N - N_D} \begin{bmatrix} N\bar{y} - N_D\bar{y}_D \\ (N - N_D)\,\bar{y}_D - N\bar{y} + N_D\bar{y}_D \end{bmatrix} \\
&= \begin{bmatrix} \bar{y}_{\setminus D} \\ \bar{y}_D - \bar{y}_{\setminus D} \end{bmatrix}
\end{aligned}$$

where N_D is the number of observations with $d_i = 1$, $\bar{y}$ is the grand average of y_i in the sample, $\bar{y}_D \equiv N_D^{-1} \sum_{i=1}^N y_i d_i$ is the average of y_i in the "dummy" group with $d_i = 1$, whereas

$$\bar{y}_{\setminus D} \equiv \frac{1}{N - N_D} \sum_{i=1}^{N} y_i\,(1 - d_i) = \frac{N\bar{y} - N_D\bar{y}_D}{N - N_D}$$

is the average of y_i in the complementary group with $d_i = 0$. By the Laws of Large Numbers and the properties of linear projections it follows that:

$$\begin{bmatrix} p_0 \\ p_1 \end{bmatrix} = \begin{bmatrix} \bar{y}_{\setminus D} \\ \bar{y}_D - \bar{y}_{\setminus D} \end{bmatrix} \xrightarrow{p} \begin{bmatrix} \mathbb{E}\left[Y_i \middle| D_i = 0\right] \\ \mathbb{E}\left[Y_i \middle| D_i = 1\right] - \mathbb{E}\left[Y_i \middle| D_i = 0\right] \end{bmatrix} = \begin{bmatrix} \pi_0 \\ \pi_1 \end{bmatrix}$$

endowing the two regression parameters π_0 and π_1 with a clear interpretation in terms of group-specific population averages.

A similar result can be obtained through the following – perhaps simpler – alternative model, with two dummy variables and no constant term:

$$Y_i = \pi'_1 D_i + \pi'_2 (1 - D_i) + \eta'_i \tag{7.46}$$

and it is even easier to show that:

$$\begin{bmatrix} p'_1 \\ p'_2 \end{bmatrix} = \begin{bmatrix} \bar{y}_{\setminus D} \\ \bar{y}_D \end{bmatrix} \xrightarrow{p} \begin{bmatrix} \mathbb{E}[Y_i | D_i = 1] \\ \mathbb{E}[Y_i | D_i = 0] \end{bmatrix} = \begin{bmatrix} \pi'_1 \\ \pi'_2 \end{bmatrix}$$

with an even more straightforward interpretation. Observe, however, that it is **impossible** to run a model like (7.46) with the addition of a constant term:

$$Y_i = \pi''_0 + \pi''_1 D_i + \pi''_2 (1 - D_i) + \eta''_i \quad (?)$$

because no unique vector of Least Squares coefficient is possibly computed. The reason is that the columns of the regressors matrix $\mathbf{X}$ are by construction linearly dependent, implying that the 3×3 matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is singular. This problem, which can be generalized to higher dimensions, is popularly known as the "dummy variable trap."

Another important observation about the two dummy variable models (7.45) and (7.45) is that no statistical assumption was necessary in order to attribute them an interpretation as group means (both in the sample and, asymptotically, in the population). This can be generalized: suppose that a population is partitioned between K non-overlapping groups, which may also represent the intersections of more aggregate divisions; for example, a partition that intersects two binary groups like "gender" and "age" is the set $\mathbb{G} = \{male\&young, female\&young, male\&old, female\&old\}$, with $K = 4$. If a dummy variable X_{ki} is associated to each group for $k = 1, \ldots, K$, thus taking care that the dummy variable trap is avoided, it holds that:

$$\mathbb{E}[Y_i | X_{1i}, \ldots, X_{Ki}] = \mathbb{E}[Y_i | \boldsymbol{x}_i] = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\pi}_0 \tag{7.47}$$

for every dependent variable Y_i, and the parameters $\boldsymbol{\pi}_0$ are interpreted as the K group-specific means of Y_i. Such a model is called a **fully saturated regression**; it finds its primary use is in those statistical exercises that go by the name of "Analysis of Variance" (ANOVA) and whose objective is to examine group differences in selected populations.

Dummy variables are routinely used in econometrics in order to account for group heterogeneity. A dummy variable D_i can be added autonomously to a regression model (so that the associated parameter is interpreted as a group-specific shifter of the constant term) and also "interacted" with some regressor of interest X_{ik}, that is added as an additional regressor taking the form $X_{ik} D_i$ (whose parameter would thus be interpreted as group-specific variation of the contribution of X_{ik} to the CEF). An example follows.

Example 7.9. Human capital and wages – blacks and whites. Let us return again to the analysis of returns to schooling, examining how they may differ between the two major racial groups in the US: blacks and whites. Figure 7.7 below helps visualize the racial split in the cross-sectional excerpt from Examples 7.2 and 7.5: blacks, who constitute about 33% of the sample, appear more prevalent amongst the lower income brackets.

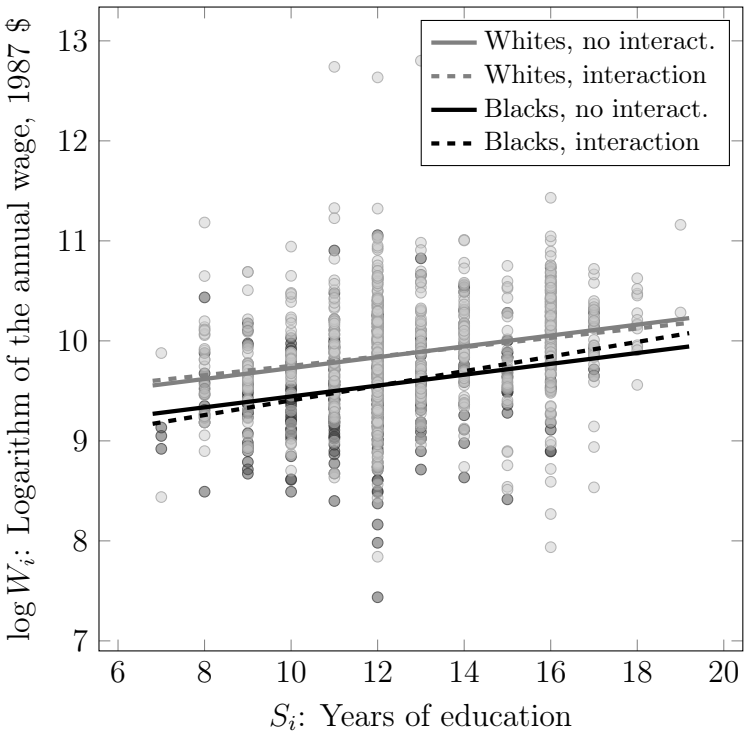


Figure 7.7: (Log) wage and education in 1987; node colors match race

Consider the simple regression model:

$$\log W_i = \beta_0 + \pi_0 D_i + \beta_1 S_i + \varepsilon_i \tag{7.48}$$

where experience is ignored for simplicity. The estimates are represented as the two parallel, solid lines from Figure 7.5; the parameter π_0 is evaluated as $p_0 \simeq -0.283$ and represents the average difference in log-wages between the two groups across all education levels. Adding an "interaction term:"

$$\log W_i = \beta_0 + \pi_0 D_i + \beta_1 S_i + \pi_1 D_i S_i + \varepsilon_i \tag{7.49}$$

is equivalent to allowing for one regression line per group. The associated estimates are displayed through the dashed lines in Figure 7.5, suggesting that the earning gap tends to close at higher levels of education. ■

Regression and the CEF Derivative

The view and use of Linear Regression as an "approximation" of some unknown CEF is also motivated by another property of Least Squares: that they provide a good approximation to the **average derivative** of the CEF. This is especially important in light of the relationship between the CEF and the *causal effects* that are discussed later in Lecture 9. This property was observed first by Yitzhaki (1996) in a bivariate setting; it was expanded later by Angrist and Krueger (1999) in a multivariate model where the CEF between of some dependent variable Y_i and some regressor of interest S_i with a *continuous* support $\mathbb{X}_S$ is unknown, but S_i is expected to depend linearly upon a set of other explanatory variables $\boldsymbol{x}_i$ according to a linear CEF.

$$\mathbb{E}\left[S_i \middle| \boldsymbol{x}_i\right] = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\pi}_0 \tag{7.50}$$

Relationship (7.50) above is satisfied, for example, in a fully saturated environment where $\boldsymbol{x}_i$ represents a complete dummy variable partition of the population. Denote the following **derivative** of the CEF as $\mu'_{Y|S,\boldsymbol{x}}(s_i; \boldsymbol{x}_i)$:

$$\mu'_{Y|S,\boldsymbol{x}}(s_i; \boldsymbol{x}_i) \equiv \frac{\partial}{\partial S_i} \mathbb{E}\left[Y_i \middle| S_i; \boldsymbol{x}_i\right]\Big|_{S_i = s_i} \tag{7.51}$$

which here measures the expected marginal increase of Y_i for observations with $S_i = s_i$, as a function of $\boldsymbol{x}_i$. If S_i has a discrete support, an analogous definition in terms of discrete variation applies.

Observe that in a Least Squares fit of Y_i on $\boldsymbol{x}_i$ and S_i:

$$Y_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \delta_0 S_i + \varepsilon_i \tag{7.52}$$

the coefficient associated with education is the following variation of a partial correlation coefficient.

$$\widehat{\delta}_{OLS} = \frac{\mathbf{s}^{\mathrm{T}} \mathbf{M}_{\mathbf{X}} \mathbf{y}}{\mathbf{s}^{\mathrm{T}} \mathbf{M}_{\mathbf{X}} \mathbf{s}} \tag{7.53}$$

Furthermore, if the CEF of S_i conditional on $\boldsymbol{x}_i$ is linear, it is possible to show through some algebraic manipulation that the population projection coefficient associated with education S_i – the probability limit of (7.53) – is the following ratio of conditional moments.

$$\widehat{\delta}_{OLS} \xrightarrow{p} \delta^* = \frac{\mathbb{C}\mathrm{ov}\left[Y_i, S_i \middle| \boldsymbol{x}_i\right]}{\mathbb{V}\mathrm{ar}\left[S_i \middle| \boldsymbol{x}_i\right]} \tag{7.54}$$

By noting that $\mathbb{E}\left[S_i - \mathbb{E}\left[S_i \middle| \boldsymbol{x}_i\right]\right] = 0$, the above simplifies as follows.

$$\delta^* = \frac{\mathbb{E}\left[Y_i \left(S_i - \mathbb{E}\left[S_i \middle| \boldsymbol{x}_i\right]\right)\right]}{\mathbb{E}\left[S_i \left(S_i - \mathbb{E}\left[S_i \middle| \boldsymbol{x}_i\right]\right)\right]} \tag{7.55}$$

The property in question is that (7.55) is also equal to:

$$\delta^* = \frac{\mathbb{E}_{\boldsymbol{x}}\left[\int_{\mathbb{X}_S} \mu'_{Y|S,\boldsymbol{x}}\left(s_i;\boldsymbol{x}_i\right)\phi\left(s_i;\boldsymbol{x}_i\right)ds_i\right]}{\mathbb{E}_{\boldsymbol{x}}\left[\int_{\mathbb{X}_S} \phi\left(s_i;\boldsymbol{x}_i\right)ds_i\right]} \tag{7.56}$$

hence it corresponds to the **derivative** $\mu'_{Y|S,\boldsymbol{x}}\left(s_i;\boldsymbol{x}_i\right)$ of the CEF, **averaged** over the support of $\boldsymbol{x}_i$, after having been **weighted** through the support of S_i by the following term, which depends on s_i and varies with $\boldsymbol{x}_i$.

$$\begin{aligned}\phi\left(s_i;\boldsymbol{x}_i\right) \equiv \left\{\mathbb{E}\left[S_i \middle| S_i \geq s_i, \boldsymbol{x}_i\right] - \mathbb{E}\left[S_i \middle| S_i < s_i, \mathbf{x}_i\right]\right\} \times \\ \times \left\{\mathbb{P}\left(S_i \geq s_i \middle| \boldsymbol{x}_i\right)\left[1 - \mathbb{P}\left(S_i \geq s_i \middle| \boldsymbol{x}_i\right)\right]\right\}\end{aligned}$$

The term $\phi\left(s_i,\boldsymbol{x}_i\right)$ is hard to interpret, but intuitively it takes larger values around the median of S_i – as an inspection of the formula would suggest. The original derivation of (7.56) given by Angrist and Krueger was initially applied to a discrete S_i (say years of education); with some manipulation of integrals this can be proved for a continuous S_i too.[11]

[11] The proof proceeds as follows. After having defined $s^* \equiv \limsup \mathbb{X}_S$, develop:

$$\begin{aligned}\mathbb{E}\left[Y_i\left(S_i - \mathbb{E}\left[S_i \middle| \boldsymbol{x}_i\right]\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[Y_i \middle| S_i, \boldsymbol{x}_i\right]\left(S_i - \mathbb{E}\left[S_i \middle| \boldsymbol{x}_i\right]\right)\right] \\ &= \mathbb{E}\left[\left[\int_{\mathbb{X}_S} \mu'_{Y|S,\boldsymbol{x}}\left(s_i\right)ds_i\right]\left(S_i - \mathbb{E}\left[S_i \middle| \boldsymbol{x}_i\right]\right)\right] \\ &= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_S\left[\int_{\mathbb{X}_S} \mu'_{Y|S,\boldsymbol{x}}\left(s_i\right)\left(s_i - \mathbb{E}\left[S_i \middle| \boldsymbol{x}_i\right]\right)ds_i \middle| \boldsymbol{x}_i\right]\right] \\ &= \mathbb{E}_{\boldsymbol{x}}\left[\int_{\mathbb{X}_S}\int_{\mathbb{X}_S} \mu'_{Y|S,\boldsymbol{x}}\left(s_i\right)\left(s_i - \mathbb{E}\left[S_i \middle| \boldsymbol{x}_i\right]\right)f_{S|\boldsymbol{x}}\left(s_i \middle| \boldsymbol{x}_i\right)ds_i ds_i\right] \\ &= \mathbb{E}_{\boldsymbol{x}}\left[\int_{\mathbb{X}_S} \mu'_{Y|S,\boldsymbol{x}}\left(s_i\right)\left[\int_{s_i}^{s^*}\left(s_i - \mathbb{E}\left[S_i \middle| \boldsymbol{x}_i\right]\right)f_{S|\boldsymbol{x}}\left(s_i \middle| \boldsymbol{x}_i\right)ds_i\right]ds_i\right]\end{aligned}$$

where the first and third lines above are consequent to the Law of Iterated Expectations, the second line follows from the Fundamental Theorem of Calculus, and the rest obtains with some manipulation. In addition, by standard properties of conditional moments:

$$\mathbb{E}\left[S_i \middle| \boldsymbol{x}_i\right] = \mathbb{E}\left[S_i \middle| S_i \geq s_i, \boldsymbol{x}_i\right]\mathbb{P}\left(S_i \geq s_i \middle| \boldsymbol{x}_i\right) + \mathbb{E}\left[S_i \middle| S_i < s_i, \boldsymbol{x}_i\right]\left[1 - \mathbb{P}\left(S_i \geq s_i \middle| \boldsymbol{x}_i\right)\right]$$

and by repeated substitution one can verify that:

$$\begin{aligned}\int_{s_i}^{s^*}\left(s_i - \mathbb{E}\left[S_i \middle| \boldsymbol{x}_i\right]\right)f_{S|\boldsymbol{x}}\left(s_i \middle| \boldsymbol{x}_i\right)ds_i &= \left\{\mathbb{E}\left[S_i \middle| S_i \geq s_i, \boldsymbol{x}_i\right] - \mathbb{E}\left[S_i \middle| \boldsymbol{x}_i\right]\right\}\mathbb{P}\left(S_i \geq s_i \middle| \boldsymbol{x}_i\right) \\ &= \phi\left(s_i;\boldsymbol{x}_i\right)\end{aligned}$$

showing the numerator of (7.56) is as stated above; a similar, simpler analysis also applies to the denominator of (7.56).

Example 7.10. Human capital and wages: average extra return to schooling. The usefulness of this interpretation rests on the understanding of the $\phi(s_i; \boldsymbol{x}_i)$ weights, and is consequently context-specific. The top panel from Figure 7.8 below reports the calculation of these weights for different years of schooling S_i, using the same data about education and wages from previous examples; it shows how the weights are largest for those values of S_i between 12 and 16, that is between high school and college graduation.[12] In this context, the weighting scheme attributes more weight to those values of S_i that are arguably most consequential for current education policy.

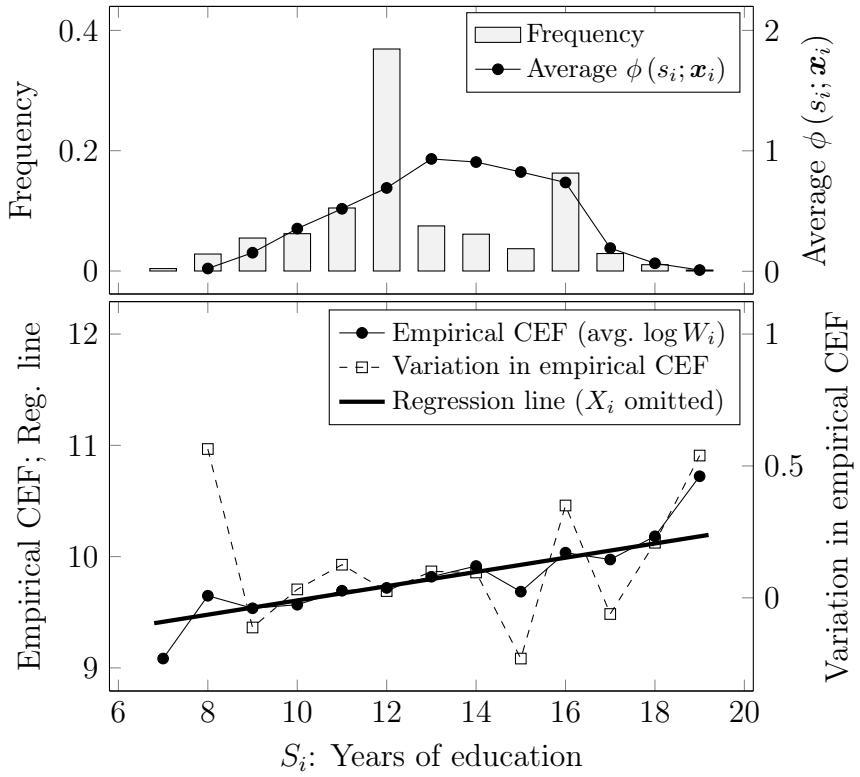


Figure 7.8: Regression as the average derivative: returns to schooling

Here, $\mu'_{\log W|S,\boldsymbol{x}}(s_i; \boldsymbol{x}_i) \equiv \mathbb{E}[\log W_i | S_i = s_i; \boldsymbol{x}_i] - \mathbb{E}[\log W_i | S_i = s_i - 1; \boldsymbol{x}_i]$; is a *discrete* variation of the CEF which expresses the expected log-wage for one extra year of education. The associated estimates are represented in the bottom panel of Figure 7.8, along with the *empirical* (non-parametric) CEF and the regression line. The average CEF variation, weighted by $\phi(s_i; \boldsymbol{x}_i)$, is about 0.064, just as the regression slope which was found earlier. ■

[12] All calculations replicate those from Angrist and Krueger (1999) with different data; the results are similar. All averages in both panels are computed conditionally on S_i.

Lecture 8

Least Squares Estimation

Having developed some general statistical and practical motivations for the use of Least Squares, this lecture examines the statistical properties of the OLS estimator, which are instrumental to statistical estimation and inference. While both small and large sample properties are analyzed, the latter are discussed first as the standard choice for use in empirical research, data permitting. Finally, the lecture develops the implications of departures from the assumption on independence between observations, along with the options available for performing reliable inference under those conditions.

8.1 Large Sample Properties

The starting point is the analysis of the **large sample** properties of OLS, which rely on asymptotic results and thus – as their name suggests – a sample size N that tends to infinity. These properties are dependent upon a number of specific **statistical assumptions**, which are sequentially introduced next by adapting the original assumptions from the treatment by White (1980). The motivation and implications of all such assumptions is discussed at length.

Assumption 1. Linearity. The data are generated by a linear model with "true" parameter vector $\boldsymbol{\beta}_0$.

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \varepsilon_i \tag{8.1}$$

This assumption may seem obvious, but is necessary to rule out all kinds of *specification errors*, such as mistaking the functional form that relates y_i and $\mathbf{x}_i$ – in which case a linear regression model estimated via OLS may not be the best econometric choice. In addition, it is conceptually important to specify that there exists a "true" parameter vector $\boldsymbol{\beta}_0$ of interest.

To denote the OLS estimator, the notation $\widehat{\boldsymbol{\beta}}_{OLS}$ will be used throughout this lecture and beyond. While the algebraic expression of the estimator is identical to that of the Least Squares solution $\mathbf{b}$ in (7.20) and (7.22), the above notation is preferred when the intention is to highlight that OLS is being used as a proper statistic and econometric estimator. Under Assumption 1, the OLS estimator can be decomposed as:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{OLS} &= \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i \left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \varepsilon_i\right) \\ &= \boldsymbol{\beta}_0 + \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \varepsilon_i\end{aligned} \tag{8.2}$$

or equivalently, in compact matrix notation, as follows.

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{OLS} &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}} \left(\mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}\right) \\ &= \boldsymbol{\beta}_0 + \left(\frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \frac{1}{N}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}\end{aligned} \tag{8.3}$$

This decomposition turns out to be very useful throughout this analysis.

Assumption 2. Independently but not identically distributed data. The observations in the sample $\{(y_i, \mathbf{x}_i)\}_{i=1}^{N}$ are *independently*, but *not* necessarily *identically*, distributed (i.n.i.d.).

This assumption characterizes the data sample. Note how the conditions imposed on it are less restrictive than those from most baseline statistical results, which usually require identically and independently distributed (i.i.d.) observations. By letting observations to be not identically distributed, they are not only allowed to have different absolute or conditional moments, but also to be drawn from different distributions. The independence assumption remains problematic in many practical contexts, and econometric solutions for scenarios when it likely fails are discussed in the last part of this lecture.

Assumption 3. Moments and realizations of the regressors. The regressors random vector $\boldsymbol{x}_i$ has a finite second moment, and for some $\delta > 0$:

$$\mathbb{E}\left[\left|X_{ik}X_{i\ell}\right|^{1+\delta}\right] < \infty \tag{8.4}$$

for $k, \ell = 1, \ldots, K$ and $i = 1, \ldots, N$. In addition, its realizations $\mathbf{x}_i$ are such that, for any two $K \times 1$ vectors $\boldsymbol{\beta}'$ and $\boldsymbol{\beta}''$:

$$\mathbf{X}\boldsymbol{\beta}' = \mathbf{X}\boldsymbol{\beta}'' \;\text{ iff }\; \boldsymbol{\beta}' = \boldsymbol{\beta}'' \tag{8.5}$$

thus, $\mathbf{X}$ has full column rank and $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ ($\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}$) is nonsingular.

This assumption specifies the nature of the regressors $\mathbf{X}$ used for estimation, and is composed of two parts. The first part is about its stochastic properties. In fact, it implicitly allows the regressors to be actually stochastic – which is not to be taken for granted, since in the classical treatment of the linear regression model the regressors are assumed to be "fixed" (that is identical in repeated samples, such as when regression is used to evaluate some kind of experimental variation); in those classical treatments the only random component of the model is the error term ε_i. Stochastic regressors are assumed to have finite second (mixed) moments while conforming to condition (8.4). All this implies that the following probability limit:

$$\frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{X} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} \xrightarrow{p} \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right] \equiv \mathbf{K}_0 \tag{8.6}$$

is a $K \times K$ matrix, written as $\mathbf{K}_0$, which is of full rank K and as such invertible (observe incidentally that when observations are *identically distributed* this matrix takes a simpler expression: $\mathbf{K}_0 = \mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right]$). The second part of the assumption states that also the actual realizations of the regressors must satisfy an analogous invertibility condition. Recall that this condition is necessary for the Least Squares solution to be unique; it rules out issues such as the dummy variable trap.[1]

Assumption 4. Exogeneity. Conditional on the regressors $\boldsymbol{x}_i$, the error term ε_i has mean zero (with typical terminology, it is **mean-independent** of the regressors $\boldsymbol{x}_i$).

$$\mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{x}_i\right] = 0 \tag{8.7}$$

This is the all-important assumption, against which one's estimates are evaluated, since it is the crucial one for obtaining **consistency** of the OLS estimator. As it was already observed in the previous lecture, (8.7) amounts to assume that the CEF is indeed linear in $\boldsymbol{x}_i$, and it implies $\mathbb{E}\left[\boldsymbol{x}_i\varepsilon_i\right] = \mathbf{0}$, hence:

$$\frac{1}{N}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\varepsilon_i \xrightarrow{p} \mathbf{0} \tag{8.8}$$

if the conditions for the application of an appropriate Law of Large Numbers are satisfied by the other assumptions, and so the residual element that adds to $\boldsymbol{\beta}_0$ on the right-hand side of (8.2) and (8.3) vanishes asymptotically. The intuition is that since $\boldsymbol{x}_i$ does not provide information on ε_i, any variation in Y_i associated with a variation in $\boldsymbol{x}_i$ must be due, on average, to $\boldsymbol{x}_i$ alone.

[1] As Lecture 9 clarifies, this is a standard *identification condition* specific to the OLS estimator.

Like much of the econometric terminology, the name "exogeneity" for this assumption originates with the analysis of Simultaneous Equations Models, (SEMs, see Lectures 9 and 10) although a more appropriate name is the longer (and hence less popular) **mean independence** of the error term. The motivation for the shorter name is best understood later in the context of the analysis of SEMs. A discussion of those frequent scenarios where this condition might fail are reviewed in later lectures.

Assumption 5. Heteroscedastic, Independent Errors. The variance of the error term ε_i conditional on $\mathbf{x}_i$ is left unrestricted (*heteroscedasticity*). Since observations are independent, the conditional covariance between two error terms from two different observations $i, j = 1, \ldots, N$ is zero.

$$\mathbb{E}\left[\varepsilon_i^2 \middle| \boldsymbol{x}_i\right] = \sigma^2\left(\boldsymbol{x}_i\right) \equiv \sigma_i^2 \tag{8.9}$$

$$\mathbb{E}\left[\varepsilon_i \varepsilon_j \middle| \boldsymbol{x}_i, \boldsymbol{x}_j\right] = 0 \tag{8.10}$$

In addition, for some $\delta > 0$ the following holds for all $i = 1, \ldots, N$.

$$\mathbb{E}\left[\left|\varepsilon_i^2\right|^{1+\delta}\right] < \infty \tag{8.11}$$

The above is written in compact matrix notation as follows.

$$\boldsymbol{\Sigma} \equiv \mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}} \middle| \boldsymbol{X}\right] = \begin{bmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_N^2 \end{bmatrix} \tag{8.12}$$

This assumption specifies the second moments of the error term ε_i. First, the conditional variances are allowed to vary on the support of $\boldsymbol{x}_i$, a notion that is named **heteroscedasticity** in econometrics. Heteroscedasticity is a natural property of most empirical settings of interest in economics and other social sciences more generally. For example, an inspection of Figure 7.2 from the previous lecture suggests that log-wages are differentially dispersed for individuals with different level of education, in a way that *might not be explained solely by the regressors* such as education, and hence must be due to some inherent variation of the other "residual" factors (the error term). The circumstance where $\sigma^2\left(\boldsymbol{x}_i\right) = \sigma_0^2$ is independent of $\boldsymbol{x}_i$, and thus identical for all observations $i = 1, \ldots, N$:

$$\boldsymbol{\Sigma} = \mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}} \middle| \boldsymbol{X}\right] = \mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\right] = \sigma_0^2 \mathbf{I} \tag{8.13}$$

is called **homoscedasticity** and must be seen as an exceptional case. In the classical analysis of the linear regression model, instead, homoscedasticity is traditionally considered a working assumption.

Assumption 5 entails some additional conditions. Actually, that the conditional cross-observation covariance of the errors is zero is not technically part of Assumption 5, since this property follows directly from Assumption 2 (independent observations); yet it is useful to state it here for a better understanding of (8.12). The usefulness of property (8.11) is clarified below.

Assumption 6. Moments of $\boldsymbol{x}_i\varepsilon_i$. For $i = 1, \ldots, N$, matrix $\mathbb{E}\left[\varepsilon_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right]$ exists, is finite, semi-definite positive and it has full rank K. Furthermore, for some $\delta > 0$, for $i = 1, \ldots, N$, and for $k, \ell = 1, \ldots, K$, the following Ljapunov condition holds.

$$\mathbb{E}\left[\left|\varepsilon_i^2 X_{ik} X_{i\ell}\right|^{1+\delta}\right] < \infty \tag{8.14}$$

This assumption allows to establish the asymptotic normality of OLS. To see this, hereinafter denote the following limiting variance with $\boldsymbol{\Xi}_0$.

$$\boldsymbol{\Xi}_0 \equiv \lim_{N\to\infty} \mathbb{V}\mathrm{ar}\left[\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \boldsymbol{x}_i \varepsilon_i\right] \tag{8.15}$$

Clearly, under Assumption 2 (*independent observations*) matrix $\boldsymbol{\Xi}_0$ assumes a more straightforward expression:

$$\begin{aligned} \boldsymbol{\Xi}_0 = \lim_{N\to\infty} \mathbb{V}\mathrm{ar}\left[\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \boldsymbol{x}_i \varepsilon_i\right] &= \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{V}\mathrm{ar}\left[\boldsymbol{x}_i \varepsilon_i\right] \\ &= \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\varepsilon_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right] \end{aligned} \tag{8.16}$$

which is semi-definite positive and has full rank by Assumption 6. Note that if the observations were also *identically distributed*, then $\boldsymbol{\Xi}_0 = \mathbb{E}\left[\varepsilon_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right]$. An implication of (8.16) is that by some Law of Large Numbers:

$$\frac{1}{N} \sum_{i=1}^{N} \varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \xrightarrow{p} \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\varepsilon_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right] = \boldsymbol{\Xi}_0 \tag{8.17}$$

and by the Ljapunov condition (8.14), the following Central Limit Theorem result holds too.

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{x}_i \varepsilon_i \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Xi}_0\right) \tag{8.18}$$

Finally, notice that in the special case of homoscedasticity, the variance of the error term is independent of the regressors, hence:

$$\mathbb{E}\left[\varepsilon_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right] = \mathbb{E}\left[\varepsilon_i^2\right] \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right] = \sigma_0^2 \, \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right]$$

for $i = 1, \ldots, N$; this implies, by (8.6) and (8.17), that $\boldsymbol{\Xi}_0 = \sigma_0^2 \mathbf{K}_0$.

Having discussed all the six White's Assumptions at length, proving the large sample properties of the OLS estimator is straightforward.

Theorem 8.1. The Large Sample properties of the OLS Estimator. *Under Assumptions 1-6 the OLS estimator is consistent, that is:*

$$\widehat{\boldsymbol{\beta}}_{OLS} \xrightarrow{p} \boldsymbol{\beta}_0 \tag{8.19}$$

and asymptotically normal, that is:

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{K}_0^{-1}\boldsymbol{\Xi}_0\mathbf{K}_0^{-1}\right) \tag{8.20}$$

hence its asymptotic distribution is, for a given N, as follows.

$$\widehat{\boldsymbol{\beta}}_{OLS} \overset{A}{\sim} \mathcal{N}\left(\boldsymbol{\beta}_0, \frac{1}{N}\mathbf{K}_0^{-1}\boldsymbol{\Xi}_0\mathbf{K}_0^{-1}\right) \tag{8.21}$$

Proof. The consistency result (8.19) was in a way already proved in the previous lecture by exploiting the properties of the linear projection when the CEF is linear; under Assumptions 1-6 it can be alternatively seen by applying the probability limit (8.8) to the decomposition of the OLS estimator in (8.2). Regarding asymptotic normality, it follows from "rephrasing" the Central Limit Theorem result in (8.18) in terms of the random sequence:

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0\right) = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{x}_i\varepsilon_i$$

which, by Slutskij's Theorem and the Cramér-Wold Device, gives – under Assumptions 1-6 – results (8.20) and (8.21).[2] □

This results constitutes the motivation for the use of the OLS estimator, thanks to its consistency property, and for performing statistical tests (on its estimated parameters) that are based on the normal distribution and the associated test statistics. Just like all asymptotic results that follow from the Central Limit Theorem, (8.20) was derived *regardless of the underlying*

[2]This conclusion is but a special case of some more general results (which themselves extend Theorems 6.8 and 6.17) about the asymptotic behavior of Method of Moments estimators for possibly non i.i.d. data. As it shall be expanded later, the OLS estimator is in fact alternatively seen as the Method of Moments estimator based on the following moment conditions.

$$\mathbb{E}\left[\boldsymbol{x}_i\varepsilon_i\right] = \mathbb{E}\left[\boldsymbol{x}_i\left(Y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right)\right] = \mathbf{0}$$

Clearly, the sample analogues of the above moment conditions are the K normal equations (7.19) which solve the Least Squares problem.

distribution that generates the sample, a result whose importance cannot be stressed enough. In the classical regression model, conversely, the analogous result is obtained under the assumption of of *normally distributed errors*, which is very restrictive since – for example – it disqualifies the use of the linear regression model in those settings where the error terms are known to follow other distributions *by construction* (for example, when the dependent variable is discrete).

A practical problem with employing the asymptotic variance in (8.21) for estimation and testing purposes is that it contains the generally unknown quantities $\mathbf{K}_0$ and $\boldsymbol{\Xi}_0$. However, the Laws of Large Numbers suggest a way to address the issue: to *estimate* these unknown expressions with the quantities that are known to asymptotically converge to them as per (8.6) and (8.17). This results in the following *estimator* of the asymptotic variance:

$$\widehat{\mathbb{A}\mathrm{var}}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right) = \left[\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right]^{-1}\left[\sum_{i=1}^{N}\left(y_i-\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{OLS}\right)^2\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right]\left[\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right]^{-1} \tag{8.22}$$

a formula called **heteroscedasticity-consistent**, **Huber-Eicker-White**, or simply **"robust"** estimator of the OLS asymptotic variance.[3] Note that in the meat of this "sandwich expression," the squared error terms ε_i^2 are substituted with the N squared residuals that are calculated with the actual OLS estimates; this is *not* an exercise about estimating the residuals, but an approach for consistent estimation of the limiting matrix $\boldsymbol{\Xi}_0$. Since this estimation problem has fixed dimension K (the number of estimated OLS parameters), it does indeed converge in probability to the desired result as N goes to infinity – a result due to Eicker (1967).

While the "robust" formula should be the **preferred** default option in empirical research, some additional insights can be gained by assuming that the errors are actually homoscedastic. If the variance of the error terms is independent of the regressors, $\boldsymbol{\Xi}_0 = \sigma_0^2\mathbf{K}_0$ holds and the asymptotic variance simplifies as:

$$\widehat{\boldsymbol{\beta}}_{OLS} \overset{A}{\sim} \mathcal{N}\left(\boldsymbol{\beta}_0, \frac{\sigma_0^2}{N}\mathbf{K}_0^{-1}\right) \tag{8.23}$$

which is consistently estimated by:

$$\widehat{\mathbb{A}\mathrm{var}}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i-\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{OLS}\right)^2\left[\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right]^{-1} \tag{8.24}$$

[3] That name arose in popularity due to the `robust` option of many STATA commands. Observe that this option computes (8.22) and its extensions by applying a multiplicative degrees of freedom correction $\frac{N}{N-K}$, although there is no theoretical basis for this.

therefore, the variance of the OLS estimator is *inversely proportional to the variance of the regressors*. Intuitively, if the linear dependence of Y_i on $\boldsymbol{x}_i$ is measured over a larger empirical support for the independent variables $\boldsymbol{x}_i$, it would appear more credible – this intuition extends naturally to the heteroscedasticity-consistent "robust" formula (8.22). It is useful to observe how the formulas for the limiting (or asymptotic) variance of the OLS estimator differ, between the heteroscedastic and the homoscedastic case, in a way that resembles the parallel formulas for the bivariate regression model, which are examined in Example 6.7.

The estimation of the variance of the OLS estimates allows to performs statistical inference about the linear regression model. Consider the simple case where hypotheses of interest are specific to one parameter, as follows.[4]

$$\mathrm{H}_0 : \ \beta_{k0} = \mathrm{c}_k \qquad\qquad \mathrm{H}_1 : \ \beta_{k0} \neq \mathrm{c}_k$$

In this case, the statistic of interest is the following **t-statistic**:[5]

$$\mathrm{t}_{\mathrm{H}_0} = \frac{\widehat{\beta}_{k,OLS} - \mathrm{c}_k}{\sqrt{\widehat{\mathbb{A}\mathrm{var}}\left(\widehat{\beta}_{k,OLS}\right)}} \xrightarrow{d} \mathcal{N}(0,1) \tag{8.25}$$

where the expression in the denominator is the square root of the kk-th entry of the estimated asymptotic variance of the OLS estimates, also called the **standard error** of the k-th estimated parameter (standard errors are typically reported, along the estimated coefficients, in the output of regressions performed by the main statistical computer packages).

After having estimated the whole variance-covariance matrix of the OLS estimates, it is possible to test hypotheses that involve multiple parameters. Consider, for example, the following $L \geq 0$ *linear* hypotheses:

$$\mathrm{H}_0 : \ \mathbf{R}\boldsymbol{\beta}_0 = \mathbf{c} \qquad\qquad \mathrm{H}_1 : \ \mathbf{R}\boldsymbol{\beta}_0 \neq \mathbf{c}$$

where $\mathbf{R}$ is a $L \times K$ matrix of full row rank L, while $\mathbf{c}$ is a $L \times 1$ vector. This setup affords great flexibility for recombining the K parameters into a set of multiple linear hypotheses. It is easy to verify that under the null hypothesis, the following asymptotic result holds.

$$\sqrt{N}\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_{OLS} - \mathbf{c}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{R}\left[\mathbb{V}\mathrm{ar}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right)\right]\mathbf{R}^{\mathrm{T}}\right)$$

[4] In most practical applications it is $\mathrm{c}_k = 0$: these are tests about the significance of a particular regressor which is included in the model.

[5] This denomination is traditional and is derived from the classical linear regression model with normally distributed errors; technically, this is in fact a z-statistic.

Such L hypotheses are typically *simultaneously* tested through the so-called **Wald statistic**:

$$\mathrm{W}_{\mathrm{H}_0} = \left(\mathbf{R}\widehat{\boldsymbol{\beta}}_{OLS} - \mathbf{c}\right)^{\mathrm{T}} \left[\mathbf{R}\widehat{\mathbb{A}\mathrm{var}}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right)\mathbf{R}^{\mathrm{T}}\right]^{-1} \left(\mathbf{R}\widehat{\boldsymbol{\beta}}_{OLS} - \mathbf{c}\right) \xrightarrow{d} \chi^2_L \tag{8.26}$$

which is nothing else but a quite particular case of a Hotelling's t-squared statistic. Therefore, by Observation 6.2 the Wald statistic asymptotically follows a chi-squared distribution with L degrees of freedom. A variation of the Wald statistic can be adapted for testing multiple *nonlinear* hypotheses; however, nonlinear hypotheses are treated later as part of the more general discussion of tests in the M-Estimation framework (lecture 11), of which the linear regression model is a particular case.

8.2 Small Sample Properties

Multiple references to a "classical" or "traditional" linear regression model have been made throughout this discussion. This model is characterized by *fixed* (non-stochastic) regressors, as well as *spherical* (homoscedastic) and *normally distributed* errors; the model's statistical properties are evaluated in terms of *exact moments* (expectation and variance) of the OLS estimates. Thus, it allows to perform statistical inference even in *small samples*, that is when asymptotic properties do not apply. In the early days of econometrics this was especially important, since the availability of large economic datasets was limited and much of applied work revolved around the analysis of macroeconomic time series, cross-country regressions or other contexts typically characterized by small sample sizes N.

For both pedagogical and practical reasons (sometimes, you do need to work with small samples) it is worth to examine the exact moments of the OLS estimators under the assumptions made so far – stochastic regressors, heteroscedasticity *et cetera* – are all maintained. It is quite convenient to perform this analysis by making use of compact matrix notation. First, it is easy to see that the OLS estimator is **unbiased**; with compact notation:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\boldsymbol{\beta}}_{OLS}\right] &= \boldsymbol{\beta}_0 + \mathbb{E}\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon}\right] \\
&= \boldsymbol{\beta}_0 + \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon}\middle|\,\boldsymbol{X}\right]\right] \\
&= \boldsymbol{\beta}_0 + \mathbb{E}_{\boldsymbol{X}}\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\underbrace{\mathbb{E}\left[\boldsymbol{\varepsilon}\middle|\,\boldsymbol{X}\right]}_{=\mathbf{0}}\right] \\
&= \boldsymbol{\beta}_0
\end{aligned} \tag{8.27}$$

that is, in expectation the OLS estimator returns the true value $\boldsymbol{\beta}_0$. Note how the exogeneity (mean independence) assumption is instrumental for obtaining this result – just like in the case of consistency – and that using the Law of Iterated Expectations allows to sidestep the fact that regressors are stochastic. The **conditional variance** of the OLS estimator, given a *specific realization* of the regressors $\mathbf{X}$, is calculated instead as follows.

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\widehat{\boldsymbol{\beta}}_{OLS}\middle|\,\mathbf{X}\right] &= \mathbb{E}\left[\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\middle|\,\mathbf{X}\right] \\ &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\,\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\,\mathbf{X}\right]\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \\ &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\end{aligned} \tag{8.28}$$

In small samples it is more convenient to work with the conditional variance of the OLS estimator; the unconditional variance can be obtained by taking the corresponding expectation over the random matrix $\boldsymbol{X}$.

These results are not immediately applicable for performing statistical inference. To this end, the model needs to be augmented with some variations of the classical assumptions.

Assumption 7. Spherical Errors. The errors are homoscedastic, that is $\sigma^2\left(\boldsymbol{x}_i\right) = \mathbb{V}\mathrm{ar}\left[\varepsilon_i\middle|\,\boldsymbol{x}_i\right] = \sigma_0^2$ or equivalently $\boldsymbol{\Sigma} = \mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\,\boldsymbol{X}\right] = \sigma_0^2\mathbf{I}$.

Assumption 8. Conditionally Normal Errors. The error term follows, given a regressor matrix $\mathbf{X}$, a conditionally normal distribution.

Together, Assumptions 7 and 8 can be expressed as follows.

$$\boldsymbol{\varepsilon}\middle|\,\mathbf{X} \sim \mathcal{N}\left(\mathbf{0}, \sigma_0^2\mathbf{I}\right) \tag{8.29}$$

Note that the regressors are still maintained stochastic. By Assumption 7, (8.28) can be re-written as:

$$\mathbb{V}\mathrm{ar}\left[\widehat{\boldsymbol{\beta}}_{OLS}\middle|\,\mathbf{X}\right] = \sigma_0^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \tag{8.30}$$

a formula associated with a fundamental, celebrated result about the classical regression model.

Theorem 8.2. Gauss-Markov Theorem. *Consider the linear regression model under Assumptions 1-7. Within the class of all* linear*,* unbiased *estimators defined as:*

$$\mathbb{B} = \left\{\widetilde{\boldsymbol{\beta}} = \mathbf{B}_0\mathbf{y}: \quad \mathbb{E}\left[\mathbf{B}_0\boldsymbol{y}\middle|\,\mathbf{X}\right] = \mathbf{B}_0\mathbf{X}\boldsymbol{\beta}_0 + \mathbf{B}_0\,\mathbb{E}\left[\boldsymbol{\varepsilon}\middle|\,\mathbf{X}\right] = \boldsymbol{\beta}_0\right\}$$

the OLS estimator is the element of $\mathbb{B}$ *that yields the* minimum variance *estimate of any element of* $\boldsymbol{\beta}_0$*, as well as of all possible linear combinations* $\mathbf{l}^{\mathrm{T}}\boldsymbol{\beta}_0$ *of* $\boldsymbol{\beta}_0$*, where* $\mathbf{l}$ *is a* $K \times 1$ *vector.*

Proof. By the definition of $\mathbb{B}$ and by Assumption 4, it follows that $\mathbf{B}_0\mathbf{X} = \mathbf{I}$ for all estimators in $\mathbb{B}$. Define $\mathbf{B}_1 \equiv \mathbf{B}_0 - \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}$ and observe that:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left[\widetilde{\boldsymbol{\beta}}\middle|\mathbf{X}\right] &= \mathbf{B}_0\,\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\mathbf{X}\right]\mathbf{B}_0^{\mathrm{T}} \\
&= \sigma^2\left[\mathbf{B}_1 + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right]\left[\mathbf{B}_1 + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right]^{\mathrm{T}} \\
&= \sigma^2\left(\mathbf{B}_1\mathbf{B}_1^{\mathrm{T}}\right) + \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \\
&= \sigma^2\left(\mathbf{B}_1\mathbf{B}_1^{\mathrm{T}}\right) + \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}
\end{aligned}$$

where the third line follows from $\mathbf{B}_1\mathbf{X} = \mathbf{B}_0\mathbf{X} - \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{0}$. Thus:

$$\mathbf{l}^{\mathrm{T}}\,\mathbb{V}\mathrm{ar}\left[\widetilde{\boldsymbol{\beta}}\middle|\mathbf{X}\right]\mathbf{l} \geq \mathbf{l}^{\mathrm{T}}\,\mathbb{V}\mathrm{ar}\left[\widehat{\boldsymbol{\beta}}_{OLS}\middle|\mathbf{X}\right]\mathbf{l}$$

which proves the *conditional* (on $\mathbf{X}$) version of the theorem; the unconditional version is easily obtained by taking the expectation over the random matrix $\boldsymbol{X}$, of which $\mathbf{X}$ is a specific realization. □

This result – which, note, has not required invoking Assumption 8 yet – is the one for which the OLS estimator deserves the denomination of **Best Linear Unbiased Estimator** (**BLUE**). In this phrase, "Best" must be interpreted in the sense of **efficient**, that is of minimum variance. However, even in small samples this result is no longer valid when homoscedasticity does not hold, as it is observed later while analyzing the Generalized Least Squares model. Since in the current empirical practice homoscedasticity is seen more as an exception and researchers are advised to employ variance estimates that are robust to heteroscedasticity in large samples – such as the "robust" formula (8.22) – the Gauss-Markov Theorem has lost much of its original significance. However, it is still seldom useful as a benchmark for efficiency comparisons.

In order to obtain a distributional result that that is usable for inference purposes, observe that by Assumption 8 it would hold *exactly* that:

$$\widehat{\boldsymbol{\beta}}_{OLS}\middle|\mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\beta}_0, \sigma_0^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\right) \tag{8.31}$$

by the properties of the normal distribution, recalling that the OLS estimator is a linear function of the error terms $\boldsymbol{\varepsilon}$ as per (8.3). This result can be immediately used for inference so long as σ_0^2 is known; since it is generally unknown, one needs to estimate this parameter. Intuitively, one could use the same estimator for (8.30) that follows from the large sample properties

under homoscedasticity, which in this case would be adapted, by writing it in compact matrix notation, as follows.

$$\widehat{\widehat{\mathbb{V}\mathrm{ar}}}\left[\widehat{\boldsymbol{\beta}}_{OLS}\middle|\mathbf{X}\right]=\frac{\mathbf{e}^{\mathrm{T}}\mathbf{e}}{N}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$$

Note that $N^{-1}\mathbf{e}^{\mathrm{T}}\mathbf{e}$ is a *consistent* estimator of σ_0^2; however, it is also a *biased* one, as one can show by taking its expectation conditional on $\mathbf{X}$:

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}\middle|\mathbf{X}\right] &= \mathbb{E}\left[\boldsymbol{y}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}}\boldsymbol{y}\middle|\mathbf{X}\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}_0\right)^{\mathrm{T}}\mathbf{M}_{\mathbf{X}}\left(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}_0\right)\middle|\mathbf{X}\right] \\
&= \mathbb{E}\left[\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}}\boldsymbol{\varepsilon}\middle|\mathbf{X}\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}}\boldsymbol{\varepsilon}\right)\middle|\mathbf{X}\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left(\mathbf{M}_{\mathbf{X}}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\right)\middle|\mathbf{X}\right] \\
&= \mathrm{Tr}\left(\mathbf{M}_{\mathbf{X}}\,\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\mathbf{X}\right]\right) \\
&= \mathrm{Tr}\left(\sigma_0^2\mathbf{M}_{\mathbf{X}}\right) \\
&= \sigma_0^2\mathrm{Tr}\left(\mathbf{I}-\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right) \\
&= \sigma_0^2\mathrm{Tr}\left(\mathbf{I}\right)-\sigma_0^2\mathrm{Tr}\left(\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\right) \\
&= \sigma_0^2\left(N-K\right)
\end{aligned}$$

hence $\mathbb{E}\left[N^{-1}\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}\middle|\mathbf{X}\right]=\frac{N-K}{N}\sigma_0^2<\sigma_0^2$.[6] Thus in small samples, even under homoscedasticity, (8.30) *underestimates* the true variance, and is inappropriate if estimators shall be evaluated in terms of their exact moments. The appropriate variance-covariance estimator for this setup would instead be:

$$\widehat{\mathbb{V}\mathrm{ar}}\left[\widehat{\boldsymbol{\beta}}_{OLS}\middle|\mathbf{X}\right]=\frac{\mathbf{e}^{\mathrm{T}}\mathbf{e}}{N-K}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \tag{8.32}$$

which, with respect to (8.30), applies a multiplicative "degrees of freedom" correction $\frac{N}{N-K}$.[7] This estimator of the OLS variance-covariance is, under

[6] This derivation makes use of the properties of the trace operator, here applied to the scalar $\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}}\boldsymbol{\varepsilon}$. Note that the order of matrices that are arguments of trace operators can be changed so long as the resulting matrix is conformable; moreover:

$$\mathrm{Tr}\left(\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right)=\mathrm{Tr}\left(\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)=\mathrm{Tr}\left(\mathbf{I}_K\right)=K$$

where $\mathbf{I}_K$ is the $K\times K$ identity matrix. In addition, the sixth equality follows from the fact that the expectation conditions on $\mathbf{X}$, hence it can pass through the trace operator as well as matrix $\mathbf{M}_{\mathbf{X}}$ (the only function of $\mathbf{X}$ in the trace).

[7] This is analogous to estimating the variance a random variable using the standard sample variance S^2, without applying the rescaling factor $\frac{N}{N-1}$ (see Theorem 4.3).

homoscedasticity, both consistent *and* unbiased, and is the default formula calculated by most statistical computer packages.

In small samples, tests of hypotheses *cannot rely on asymptotic properties*. As a consequence, while t-statistics are calculated similarly as in large samples (using appropriate estimates of the OLS variance-covariance), under Assumptions 1-8 they follow a Student's T distribution with $N-K$ degrees of freedom. To better appreciate this, consider that under (8.31) and some null hypothesis $\mathrm{H}_0 : \beta_{k0} = \mathrm{c}_k$, the following "unfeasible" t-statistic, which is denoted as $\mathrm{t}^*_{\mathrm{H}_0}$, follows a standard normal distribution conditionally on $\mathbf{X}$:

$$\mathrm{t}^*_{\mathrm{H}_0} \big| \mathbf{X} = \left. \frac{\widehat{\beta}_{k,OLS} - \mathrm{c}_k}{\sqrt{\sigma_0^2 \widetilde{\mathrm{x}}_{kk}}} \right| \mathbf{X} \sim \mathcal{N}(0,1) \tag{8.33}$$

where $\widetilde{\mathrm{x}}_{kk}$ is a shorthand notation for the kk-th element of $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$. Once again, a problem arises as σ_0^2 is unknown and must be estimated; however, it is easy to see that under the normality Assumption 8, the distribution of the *standardized sum of squared residual* $\sigma_0^{-2}\mathbf{e}^{\mathrm{T}}\mathbf{e}$, conditionally on $\mathbf{X}$:

$$\left. \frac{\mathbf{e}^{\mathrm{T}}\mathbf{e}}{\sigma_0^2} \right| \mathbf{X} = \left. \frac{\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}}\boldsymbol{\varepsilon}}{\sigma_0^2} \right| \mathbf{X} \sim \chi^2_{N-K} \tag{8.34}$$

is a chi-squared distribution with degrees of freedom equal to the rank of $\mathbf{M}_{\mathbf{X}}$; this quantity equals the trace of $\mathbf{M}_{\mathbf{X}}$, that is $N-K$ as per the earlier derivation. Furthermore, one can show that (8.33) and (8.34) are independent;[8] therefore, the *actual* t-statistic $\mathrm{t}_{\mathrm{H}_0}$ which is obtained by substituting σ_0^2 with its unbiased estimate $\mathbf{e}^{\mathrm{T}}\mathbf{e}/(N-K)$ discussed above:

$$\mathrm{t}_{\mathrm{H}_0} = \sqrt{\sigma_0^2}\sqrt{N-K}\,\frac{\mathrm{t}^*_{\mathrm{H}_0}}{\sqrt{\mathbf{e}^{\mathrm{T}}\mathbf{e}}} = \sqrt{N-K}\,\frac{\widehat{\beta}_{k,OLS} - \mathrm{c}_k}{\sqrt{\mathbf{e}^{\mathrm{T}}\mathbf{e}\cdot\widetilde{\mathrm{x}}_{kk}}} \tag{8.35}$$

follows a Student's t-distribution with $N-K$ degrees of freedom, conditionally on $\mathbf{X}$ (as usual, this follows from Observation 3.2).

$$\mathrm{t}_{\mathrm{H}_0} \big| \mathbf{X} \sim \mathcal{T}_{N-K}$$

This result is usable for inference purposes; in small but sizable samples ($N > 20$), however, this is known to yield results that are not very different from approximations based on the standard normal.

[8] The vector of OLS estimates $\widehat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta}_0 + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}$ is shown to be independent of (8.34) by the following observation.

$$\sigma_0^{-2}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}} = \mathbf{0}$$

This result also applies to each individual element of $\widehat{\boldsymbol{\beta}}_{OLS}$, from which $\mathrm{t}^*_{\mathrm{H}_0}$ is constructed.

By a similar argument it is shown that multiple linear hypotheses cannot be tested with a Wald statistic in small samples: here, the analogue of (8.26)

$$\mathrm{W}^{*}_{\mathrm{H}_0} = \left(\mathbf{R}\widehat{\boldsymbol{\beta}}_{OLS} - \mathbf{c}\right)^{\mathrm{T}} \frac{\left[\mathbf{R}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{R}^{\mathrm{T}}\right]^{-1}}{\sigma_0^2}\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_{OLS} - \mathbf{c}\right) \sim \chi^2_L \quad (8.36)$$

does indeed follow an *exact* χ^2_L distribution with L degrees of freedom, but again this is an expression that depends upon the unknown parameter σ_0^2. Therefore, in small samples an ***F*-statistic** must be used instead:

$$\mathrm{F}_{\mathrm{H}_0} = \frac{N-K}{L}\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_{OLS} - \mathbf{c}\right)^{\mathrm{T}} \frac{\left[\mathbf{R}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{R}^{\mathrm{T}}\right]^{-1}}{\mathbf{e}^{\mathrm{T}}\mathbf{e}}\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_{OLS} - \mathbf{c}\right) \quad (8.37)$$

this quantity results from dividing (8.36) by (8.34) – which are independent from one another[9] – and multiplying the ratio in question by $L^{-1}(N-K)$. By Observation 3.3, this F-statistic follows *exactly* an F-distribution with degrees of freedom L and $N-K$, conditionally on $\mathbf{X}$.

$$\mathrm{F}_{\mathrm{H}_0} | \, \mathbf{X} \sim \mathcal{F}_{L,N-K}$$

A customary use of the F-statistic is in the **model *F*-test** (or simply the **model test**) corresponding to the null hypothesis $\mathrm{H}_0 : \boldsymbol{\beta}_0 = \mathbf{0}$ that all the parameters of the model (except the constant term, if present) are jointly meaningful. The F-statistic obtained from this test is typically part of the default regression output returned by statistical computer packages.[10]

Generalized Least Squares

All the results derived so far for the "traditional" model are only valid under the restrictive assumption that the errors are homoscedastic. This problem was well acknowledged even in those days when the use of linear regression

[9]The argument is similar to that from footnote 8: since $\mathrm{W}^{*}_{\mathrm{H}_0}$ is a quadratic form in the OLS estimates, its random component is proportional to the random variable:

$$W^{*}_{N} = \boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{R}^{\mathrm{T}}\left[\mathbf{R}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{R}^{\mathrm{T}}\right]^{-1}\mathbf{R}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}$$

and it is easy to see that the central coefficient matrix of this quadratic form returns $\mathbf{0}$ whether it is pre- or post-multiplied to $\mathbf{M}_{\mathbf{X}}$.

[10]The F-statistic and the model F-test are typically evaluated even in large sample environments, in which cases they are calculated through the appropriate estimates of the asymptotic variance of the OLS estimator. The F-distribution might, in fact, provide a better approximation of the true underlying probabilities.

relied for the most part on its small sample properties, which motivated the search for an adequate solution within the same framework. This resulted in the development of the **Generalized Least Squares** (GLS) model.

The intuition behind GLS is simple. Suppose that the errors are indeed heteroscedastic, but matrix $\boldsymbol{\Sigma} = \mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}} \middle| \boldsymbol{X}\right]$ is known. If one performs OLS estimation on the **generalized linear model**:

$$\widetilde{\mathbf{y}} = \widetilde{\mathbf{X}}\boldsymbol{\beta}_0 + \widetilde{\boldsymbol{\varepsilon}} \tag{8.38}$$

where:

$$\widetilde{\mathbf{y}} \equiv \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{y}; \quad \widetilde{\mathbf{X}} \equiv \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{X}; \quad \widetilde{\boldsymbol{\varepsilon}} \equiv \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\varepsilon}$$

and:

$$\boldsymbol{\Sigma}^{-\frac{1}{2}} \equiv \begin{bmatrix} \sigma_1^{-1} & 0 & \dots & 0 \\ 0 & \sigma_2^{-1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_N^{-1} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{\sigma_1^2}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{\sigma_2^2}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sqrt{\sigma_N^2}} \end{bmatrix}$$

such that $\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}} = \boldsymbol{\Sigma}$, then the **Generalized Least Squares** estimator:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{GLS} &= \left(\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{y}} \\ &= \left(\mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{y} \\ &= \boldsymbol{\beta}_0 + \left(\mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon} \end{aligned} \tag{8.39}$$

is easily seen to be *unbiased* with respect to $\boldsymbol{\beta}_0$. Moreover, since:

$$\mathbb{E}\left[\widetilde{\boldsymbol{\varepsilon}}\widetilde{\boldsymbol{\varepsilon}}^{\mathrm{T}} \middle| \boldsymbol{X}\right] = \boldsymbol{\Sigma}^{-\frac{1}{2}}\,\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}} \middle| \boldsymbol{X}\right]\boldsymbol{\Sigma}^{-\frac{1}{2}} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-\frac{1}{2}} = \mathbf{I}$$

the GLS estimator is **homoscedastic by construction**; its conditional variance is:

$$\widehat{\mathbb{V}\mathrm{ar}}\left[\widehat{\boldsymbol{\beta}}_{GLS} \middle| \mathbf{X}\right] = \left(\mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1} \tag{8.40}$$

and it is easy to show that, by an extension of the Gauss-Markov theorem, the GLS estimator is **efficient** under heteroscedasticity. In addition, under Assumptions 1-6 and Assumption 8, the GLS estimator follows an *exact* conditional normal distribution.

$$\widehat{\boldsymbol{\beta}}_{GLS} \middle| \mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\beta}_0, \left(\mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\right) \tag{8.41}$$

Obviously, in large samples one could evaluate the asymptotic properties of the GLS estimator as well: it is consistent and asymptotically normal as per (8.41), whether the error terms are conditionally normal or not.

The main problem with the GLS estimator is that $\boldsymbol{\Sigma}$ is, clearly, generally unknown, therefore this estimator is *unfeasible* in practice. A solution would be to *substitute* $\boldsymbol{\Sigma}$ with some plausible estimate of it: this approach is called **Feasible Generalized Least Squares** (FGLS) and it works as follows.

1. Assume a functional form for the dependence of the variance of the error term on the covariates $\mathbf{X}$; for example, a simple and popular choice is the *exponential conditional variance* $\sigma^2(\boldsymbol{x}_i) = \exp\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\psi}\right)$;

2. estimate the main regression model of interest via OLS, which returns an unbiased and consistent estimate of $\boldsymbol{\beta}_0$, and calculate the resulting *squared residuals* $(e_1^2, e_2^2, \ldots, e_N^2)$;

3. estimate via OLS the assumed model for the conditional variance; in the exponential case this model would be $\log e_i^2 = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\psi} + \varpi_i$, where ϖ_i is some error term with $\mathbb{E}[\varpi_i | \boldsymbol{x}_i] = 0$;

4. construct matrix $\widehat{\boldsymbol{\Sigma}}$, the estimate of $\boldsymbol{\Sigma}$, accordingly; in the exponential case it would be, for example:

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} \exp\left(\mathbf{x}_1^{\mathrm{T}}\widehat{\boldsymbol{\psi}}_{OLS}\right) & 0 & \ldots & 0 \\ 0 & \exp\left(\mathbf{x}_2^{\mathrm{T}}\widehat{\boldsymbol{\psi}}_{OLS}\right) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \exp\left(\mathbf{x}_N^{\mathrm{T}}\widehat{\boldsymbol{\psi}}_{OLS}\right) \end{bmatrix}$$

5. finally, calculate the FGLS estimator as follows.

$$\widehat{\boldsymbol{\beta}}_{FGLS} = \left(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{y} \tag{8.42}$$

   Note that by denoting as $\widehat{\sigma}_i$ the square root of the i-th element of the diagonal of $\widehat{\boldsymbol{\Sigma}}$, the above could be equivalently obtained by running OLS on the following transformed model (possibly, $x_{i1} = 1$ for all i).

$$\frac{y_i}{\widehat{\sigma}_i} = \beta_1\frac{x_{i1}}{\widehat{\sigma}_i} + \beta_2\frac{x_{i2}}{\widehat{\sigma}_i} + \cdots + \beta_K\frac{x_{iK}}{\widehat{\sigma}_i} + \frac{\varepsilon_i}{\widehat{\sigma}_i} \tag{8.43}$$

   This approach is known as **Weighted Least Squares** (WLS).

Under Assumptions 1-6 the FGLS-WLS estimator is both unbiased and consistent. Moreover, *if the conditional variance model is correctly specified*, in small samples it provides efficiency gains relative to naive OLS when the errors are heteroscedastic, while in large samples its unconditional variance converges in probability to the "theoretical" GLS conditional variance on the right-hand side of (8.40). Therefore, under these ideal conditions, inference is more reliable when using the FGLS-WLS estimator instead of OLS.

The problem with this approach is that it may fail if the conditional variance model is incorrectly specified. In this case, FGLS-WLS might be *less* efficient than "traditional" OLS, even in small samples. Consequently, a theory about *tests for heteroscedesticity* was developed, whose objective is to guide researchers in search of the right specification of the heteroscedasticity model. Nowadays, these tests and GLS altogether are seen as largely redundant, since modern econometric practice relies on large samples, asymptotic properties and "heteroscedasticity-robust" variance estimators.[11] However, learning GLS can still be useful, for both pedagogical and practical reasons. The pedagogical reason is that it is instructive to make efficiency comparisons between certain estimators (like 3SLS or linear GMM) and the GLS benchmark. The practical reason is that GLS is still used in some settings, for example in models for panel data featuring so-called "random effects."

8.3 Dependent Errors

A fundamental assumption that has been maintained throughout the discussion of both large and small sample properties of OLS is that of *independent observations*. This hypothesis allows to assume *independent errors*: $\mathbb{E}\left[\varepsilon_i\varepsilon_j|\,\boldsymbol{x}_i,\boldsymbol{x}_j\right]=0$ for any two observations i and j. In large samples, this property is especially convenient for applying the Central Limit Theorem, because it implies a convenient estimator for $\boldsymbol{\Xi}_0$ as per (8.17); in small samples, it is instrumental for establishing the Gauss-Markov efficiency bound. Yet error independence is quite as useful an assumption as it is unlikely to be tenable in a wide array of situations, which can be classified as follows.

- **Autocorrelation in Time**. Traditionally, this was the original cause for concern about dependent observations in econometrics, and is particularly relevant in time-series and macroeconometric analysis. In a time-series model where time is indexed by $t=1,\ldots,T$:

$$y_t=\mathbf{x}_t^{\mathrm{T}}\boldsymbol{\beta}_0+\varepsilon_t \tag{8.44}$$

  the unobserved "shock" α_t of today can be related to that of the past:

$$\mathbb{E}\left[\varepsilon_t\varepsilon_{t-s}|\,\boldsymbol{x}_t,\boldsymbol{x}_{t-s}\right]\neq 0 \tag{8.45}$$

  where $s\neq 0$. This circumstance is called **autocorrelation** and must be considered an inherent feature of time series data rather than an exception, since the external factors that affect different observations of different kind – from countries to stocks – change slowly over time.

[11] Remarkably, a command for direct implementation of GLS is missing from STATA.

- **Spatial Correlation**. The concept of autocorrelation can be naturally extended from time to some notion of "space." Suppose that in a standard cross-sectional model where observations are indexed by $i = 1, \ldots, N$, pairs of observations can be characterized by some measure of reciprocal "distance" $d_{ij} \geq 0$. This concept can be attributed some different interpretations, from actual distance in physical space to more abstract notions such as network distance. Spatial correlation is the scenario in which the errors of two different observations i and j are increasingly more correlated the closer the two observations are:

$$\mathbb{E}\left[\varepsilon_i \varepsilon_j \middle| \boldsymbol{x}_i, \boldsymbol{x}_j\right] = g\left(d_{ij}\right) \neq 0 \tag{8.46}$$

  where $g\left(d_{ij}\right)$ is some decreasing function of distance, possibly yielding zero for d_{ij} large enough. Intuitively, individuals, firms or cities that are closer in physical space might be subject to more similar external circumstances, and so do "friends" in a network.

- **Within-Group Correlation**. Suppose that the sample can be split by a number $C < N$ of **groups** or **clusters** indexed by $c = 1, \ldots, C$; in addition, each observation belongs to one and only one group or cluster. Hence, observations can be indexed by group as well.

$$y_{ic} = \mathbf{x}_{ic}^{\mathrm{T}} \boldsymbol{\beta}_0 + \varepsilon_{ic} \tag{8.47}$$

  Within-group correlation is the case where individual errors can be split between two components:

$$\varepsilon_{ic} = \alpha_c + \epsilon_{ic} \tag{8.48}$$

  where the **idiosyncratic shock** ϵ_{ic} is independent across any pair of observations $i, j = 1, \ldots, N$, regardless of their group:

$$\mathbb{E}\left[\epsilon_{ic} \epsilon_{jg} \middle| \boldsymbol{x}_{ic}, \boldsymbol{x}_{jg}\right] \begin{cases} = \sigma_\epsilon^2\left(\boldsymbol{x}_i\right) & \text{if } i = j \\ = 0 & \text{if } i \neq j \end{cases} \tag{8.49}$$

  while the **group** or **cluster shock** α_c correlates within groups, but not across groups.

$$\mathbb{E}\left[\alpha_c \alpha_g \middle| \boldsymbol{x}_{ic}, \boldsymbol{x}_{jg}\right] \begin{cases} \neq 0 & \text{if } c = g \\ = 0 & \text{if } c \neq g \end{cases} \tag{8.50}$$

  This setup is suited to describe similar "shocks" that affect groups of individuals (e.g. classmates, compatriots) or "categories" (firms in the same industry, cities in the same administrative unit and so on).

- **Combinations of the Above.** These scenarios can co-exist at the same time. In a panel data model indexed by panel unit $i = 1, \ldots, N$, time $t = 1, \ldots, T$ and group $c = 1, \ldots, C$ for example:

$$y_{itc} = \mathbf{x}_{itc}^{\mathrm{T}} \boldsymbol{\beta}_0 + \varepsilon_{itc} \tag{8.51}$$

  autocorrelation in time, spatial correlation, and group correlation can all be simultaneously present. In particular, in panel data "groups" may coincide with panel units ($C = N$); in this case the group shocks α_c are called *random effects* and are interpreted as those specific factors affecting the same individual unit repeatedly (albeit possibly not to the same extent in different moments).

In these circumstances, inference based on any of the variance estimators examined so far is unlikely to be accurate. However, a number of solutions exist, which may vary by context. In small samples, the GLS framework can be adapted to allow for dependent errors too. In large samples, *cluster-based* covariance estimators are especially suited for the case of within-group correlation, while the more complex cases of spatial and temporal dependence can also be addressed with appropriate *"heteroscedasticity-autocorrelation-consistent"* (HAC) covariance estimators. All of these are reviewed in turn.

Generalized Least Squares, Revisited

The GLS framework can incorporate dependent errors. The reason is that the GLS estimator (8.39) is well defined even if $\boldsymbol{\Sigma}$ is non-diagonal. Likewise, a Cholesky decomposition of $\boldsymbol{\Sigma}$ in such a way that:

$$\boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-\frac{1}{2}} = \mathbf{I}$$

is always possible since $\boldsymbol{\Sigma}$ is semi-definite positive, although $\boldsymbol{\Sigma}^{-\frac{1}{2}}$ may be non-diagonal. A classical application of GLS in the early days of econometrics was for autocorrelated time series. Suppose, for example, that in (8.44) the shock ε_t follows a **first-order autoregressive** – AR(1) – process:

$$\varepsilon_t = \rho \varepsilon_{t-1} + \xi_t$$

where $|\rho| < 1$ and ξ_t is i.i.d. (homoscedastic, uncorrelated over time). Then:

$$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho & \ldots & \rho^T \\ \rho & 1 & \ldots & \rho^{T-1} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^T & \rho^{T-1} & \ldots & 1 \end{bmatrix}$$

and OLS estimation of (8.44) is inefficient. The FGLS approach in this case is about estimating ρ and thereby the non-diagonal part of $\boldsymbol{\Sigma}$.

Similar extensions of GLS for spatial correlation and group dependence are possible, so long as a specific parametric form of the structure of error dependence is assumed. To illustrate, consider the so-called **cluster-specific random effects** (CSRE) model, where (8.50) is specified as:

$$\mathbb{E}\left[\alpha_c \alpha_g \middle| \boldsymbol{x}_{ic}, \boldsymbol{x}_{jg}\right] \begin{cases} = \sigma_\alpha^2 & \text{if } c = g \\ = 0 & \text{if } c \neq g \end{cases}$$

which amounts to assume constant within-group covariance. If, in addition, standard homoscedasticity is assumed, that is $\sigma_\epsilon^2(\boldsymbol{x}_i) = \sigma_\epsilon^2$ is equal for all observations, then $\boldsymbol{\Sigma}$ is *block-diagonal* over the C clusters:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\Sigma}_C \end{bmatrix}$$

where, given an identity matrix $\mathbf{I}$ and a unit vector ι of the same dimension as the size N_c of some cluster c:

$$\boldsymbol{\Sigma}_c = \begin{bmatrix} \sigma_\epsilon^2 + \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\epsilon^2 + \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\epsilon^2 + \sigma_\alpha^2 \end{bmatrix} = \sigma_\epsilon^2 \mathbf{I} + \sigma_\alpha^2 \iota \iota^{\mathrm{T}}$$

for $c = 1, \dots, C$. This setup is restrictive due to the assumptions of constant group covariance and homoscedasticity, but is particular easy to handle by GLS. Intuitively, a first set of OLS estimates is used to consistently estimate the within-cluster covariance σ_α^2 and thereby construct the appropriate estimate of $\boldsymbol{\Sigma}$, then FGLS estimation follows.[12] In fact, the CSRE model is still quite used in practice, especially in panel data models featuring unit-specific random effects. In general, however, clustered covariance and HAC estimators that rely on large sample asymptotic results instead of parametric variance models are considered preferable whenever applicable.

[12]Note that under dependent errors FGLS cannot be interpreted in terms of an equivalent Weighted Least Squares model, since the transformed model implied by GLS, in this like in other cases of error dependence, is about *linear combinations* of observations obtained in such a way that resulting errors $\widetilde{\boldsymbol{\varepsilon}}$ are both homoscedastic and uncorrelated. In the case of the CSRE model, for example, the transformation is quite convenient:

$$y_{ic} - \varpi_c \bar{y}_c = \left(\mathbf{x}_{ic} - \varpi_c \bar{\mathbf{x}}_c\right)^{\mathrm{T}} \boldsymbol{\beta}_0 + \left(\varepsilon_{ic} - \varpi_c \bar{\varepsilon}_c\right)$$

where $\varpi_c \equiv 1 - \sigma_\epsilon \left(\sigma_\epsilon^2 + N_c \sigma_\alpha^2\right)^{-\frac{1}{2}}$ and where $\bar{y}_c$, $\bar{\mathbf{x}}_c$ and $\bar{\varepsilon}_c$ are the *cluster-specific* sample means of y_{ic}, $\mathbf{x}_{ic}$ and ε_{ic} respectively.

Clustered Covariance Estimation

In large samples, and in presence of a high number of groups or clusters C, within-group dependence of any kind is elegantly addressed by an extension of the "robust" heteroscedasticity-consistent formula (8.22). To appreciate this, the development of some additional notation is necessary. Consider a single group or cluster c, index its observations as $i = 1, \ldots, N_c$, and stack them vertically as follows.

$$\mathbf{y}_c = \begin{bmatrix} y_{1c} \\ y_{2c} \\ \vdots \\ y_{N_c c} \end{bmatrix}; \quad \mathbf{X}_c = \begin{bmatrix} \mathbf{x}_{1c}^{\mathrm{T}} \\ \mathbf{x}_{2c}^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_{N_c c}^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} x_{11c} & x_{21c} & \ldots & x_{K1c} \\ x_{12c} & x_{22c} & \ldots & x_{K2c} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1N_c c} & x_{2N_c c} & \ldots & x_{KN_c c} \end{bmatrix}; \; \boldsymbol{\varepsilon}_c = \begin{bmatrix} \varepsilon_{1c} \\ \varepsilon_{2c} \\ \vdots \\ \varepsilon_{N_c c} \end{bmatrix}$$

In the case of model (8.47) featuring groups or clusters, the usual compact matrix notation equation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$ is therefore obtained by vertically stacking the following system of equations:

$$\mathbf{y}_c = \mathbf{X}_c \boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}_c$$

over each of the C clusters. Note that these C groups are allowed to have different sizes N_c. Thus, the OLS estimator of model (8.47) can be expressed in the following three equivalent ways.

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{OLS} &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}}\mathbf{y} \\ &= \left(\sum_{c=1}^{C} \mathbf{X}_c^{\mathrm{T}} \mathbf{X}_c \right)^{-1} \sum_{c=1}^{C} \mathbf{X}_c^{\mathrm{T}} \mathbf{y}_c \\ &= \left(\sum_{c=1}^{C} \sum_{i=1}^{N_c} \mathbf{x}_{ic} \mathbf{x}_{ic}^{\mathrm{T}} \right)^{-1} \sum_{c=1}^{C} \sum_{i=1}^{N_c} \mathbf{x}_{ic} y_{ic} \end{aligned}$$

With group dependence, standard inference is invalid because now:

$$\begin{aligned} \boldsymbol{\Xi}_0 = \lim_{N\to\infty} \mathbb{V}\mathrm{ar} \left[\frac{1}{\sqrt{N}} \sum_{c=1}^{C} \sum_{i=1}^{N_c} \boldsymbol{x}_{ic} \varepsilon_{ic} \right] &= \lim_{N\to\infty} \frac{1}{N} \sum_{c=1}^{C} \mathbb{V}\mathrm{ar} \left[\sum_{i=1}^{N_c} \boldsymbol{x}_{ic} \varepsilon_{ic} \right] \\ &= \lim_{N\to\infty} \frac{1}{N} \sum_{c=1}^{C} \sum_{i=1}^{N_c} \sum_{j=1}^{N_c} \mathbb{E}\left[\varepsilon_{ic} \boldsymbol{x}_{ic} \boldsymbol{x}_{jc}^{\mathrm{T}} \varepsilon_{jc} \right] \end{aligned} \tag{8.52}$$

but this expression cannot be reduced to (8.16); this invalidates the "meat" matrix of the heteroscedasticity-consistent estimator of the OLS variance. Intuitively, under group dependence the appropriate estimate of $\boldsymbol{\Xi}_0$ should be a sample version of the ultimate expression in the derivation (8.52) above.

Conveniently, under some appropriate modifications of Assumptions 2-6, a Central Limit Theorem result *for dependent observations* holds.

$$\frac{1}{\sqrt{N}}\sum_{c=1}^{C}\sum_{i=1}^{N_c}\mathbf{x}_{ic}\varepsilon_{ic}\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\boldsymbol{\Xi}_0\right) \tag{8.53}$$

Furthermore, a consistent estimator $\widehat{\boldsymbol{\Xi}}_{CCE}\xrightarrow{p}\boldsymbol{\Xi}_0$ for the limiting variance above obtains by letting two realizations of $\boldsymbol{x}_{ic}\varepsilon_{ic}$ and $\boldsymbol{x}_{jc}\varepsilon_{jc}$ from the same cluster c interact in the estimation:[13]

$$\begin{aligned}\widehat{\boldsymbol{\Xi}}_{CCE} &= \frac{1}{N}\sum_{c=1}^{C}\mathbf{X}_c^{\mathrm{T}}\mathbf{e}_c\mathbf{e}_c^{\mathrm{T}}\mathbf{X}_c \\ &= \frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{N_c}\sum_{j=1}^{N_c}\left(y_{ic}-\mathbf{x}_{ic}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{OLS}\right)\mathbf{x}_{ic}\mathbf{x}_{jc}^{\mathrm{T}}\left(y_{jc}-\mathbf{x}_{jc}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{OLS}\right)\end{aligned} \tag{8.54}$$

where $\mathbf{e}_c\equiv\mathbf{y}_c-\mathbf{X}_c\widehat{\boldsymbol{\beta}}_{OLS}$ substitutes for $\boldsymbol{\varepsilon}_c$ by arguments analogous to those in the standard case. Since the "bread" matrices of the limiting variance of OLS are unchanged, the estimator of the corresponding asymptotic variance can be obtained by substituting the meat matrix of (8.22) with (8.54) above. The resulting variance-covariance estimator is:

$$\widehat{\mathbb{A}\mathrm{var}}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right)=\left[\sum_{c=1}^{C}\mathbf{X}_c^{\mathrm{T}}\mathbf{X}_c\right]^{-1}\left[\sum_{c=1}^{C}\mathbf{X}_c^{\mathrm{T}}\mathbf{e}_c\mathbf{e}_c^{\mathrm{T}}\mathbf{X}_c\right]\left[\sum_{c=1}^{C}\mathbf{X}_c^{\mathrm{T}}\mathbf{X}_c\right]^{-1} \tag{8.55}$$

and is called **cluster-robust** or **clustered covariance estimator** (CCE). Observe how this result is obtained *without placing any restrictions on the within-group correlation*, unlike the CSRE-FGLS case where some parametric assumptions on the structure of error dependence are necessary.

[13] To gain further intuition, it is useful to think of $\mathbf{X}=\iota$ as a single constant vector, so that the only parameter that OLS tries to estimate is the unconditional mean of Y_i: $\beta_0=\mathbb{E}\left[Y_i\right]$, with variance $\mathbb{V}\mathrm{ar}\left[Y_i\right]$. The estimation challenge lies in the possibility that the variation of Y_i is correlated within groups. The OLS estimator here is just the sample mean $\bar{Y}$; normally, one would estimate its asymptotic variance as follows.

$$\widehat{\mathbb{V}\mathrm{ar}}\left[\bar{Y}\right]=\frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{N_c}\left(Y_{ic}-\bar{Y}\right)^2$$

Under within-group dependence, the appropriate asymptotic estimator of the variance of $\bar{Y}$ is instead a *rescaled sum* of the C *squared total within-cluster deviations*.

$$\widehat{\mathbb{V}\mathrm{ar}}\left[\bar{Y}\right]=\frac{1}{N^2}\sum_{c=1}^{C}\left[\sum_{i=1}^{N_c}\left(Y_{ic}-\bar{Y}\right)\right]^2$$

The above is identical to (8.54) under the maintained hypothesis that $\mathbf{X}=\iota$.

This result, while quite convenient, is obtained under the condition that the number of clusters C is large and grows to infinity. In general, however, the number of clusters is finite and typically not very large. This is one of the reasons that has motivated the frequent use of the CCE formula (8.55) with a multiplicative "degrees of freedom correction" $\frac{C}{C-1}\frac{N}{N-K}$ that takes into account the fact that both the number of clusters and the sample size are finite.[14] With a very low number of clusters C – usually between 20 and 50 – however, the CCE formula is employed along statistical tests for small samples (based on the Student's T and the F distributions). A paper by Bester et al. (2011) provides theoretical foundations for this practice: if C is small and fixed but N_c goes to infinity in all clusters, intuitively the C within-group averages of $\boldsymbol{x}_{ic}\varepsilon_{ic}$ are asymptotically normally distributed; in addition they show that CCE estimation works even under some weak forms of cross-cluster error dependence, so long as clusters are similar enough in their observable and unobservable characteristics, as well as in their size.

In current microeconometric practice, the majority of non-experimental studies feature some form of clustered covariance estimation. This is not in minor part due to some influential papers (Moulton, 1986; Bertrand et al., 2004) which observed that failing to account for within-group dependence can lead to seriously biased inference results.[15] In particular, **panel data** estimates are routinely clustered *at least* at the level of panel units, however it often makes sense to define clusters at an even higher level of aggregation (for example, in a panel of firms one may want to consider industry-level clusters, including all observations of firms of the same industry over all the years T). In ideal **experimental studies**, instead, it is *not* necessary to cluster standard errors: intuitively, even if the errors are correlated within groups, if $\boldsymbol{x}_{ic}$ is independent of ε_{ic}, therefore $\boldsymbol{\Xi}_0 = \sigma_0^2\mathbf{K}_0$ holds and standard estimation under homoscedasticity is asymptotically consistent.

[14] This is similar to the standard practice of estimating "robust" standard errors with a multiplicative degrees of freedom correction $\frac{N}{N-K}$, a habit which is motivated however more by customs than by either theory or data concerns.

[15] In some stylized cases, it is possible to solve for the explicit analytic expression of this bias. Consider, for example, the CSRE model with *equal group sizes* $M = N_c = N/C$ for $c = 1, \dots, C$ and *identical regressors* across clusters $\mathbf{X}_c = \mathbf{X}_g$ for $c \neq g$; in this case the asymptotic variance of the OLS estimator can be shown to simplify as:

$$\mathbb{A}\text{var}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right) = \left[1 + (M-1)\,\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}\right] \cdot \sigma_\epsilon^2 \left(\sum_{c=1}^{C} \mathbf{X}_c^{\mathrm{T}}\mathbf{X}_c\right)^{-1}$$

therefore, a standard estimate of the homoscedastic variance of the OLS estimate would be *downward biased*. The extent of the bias is expressed by the muliplicative term within brackets, which is often referred to as the *Moulton bias* from Moulton (1986).

Clustering Covariance Estimation can be generalized to *multiple group dimensions* within which errors are expected to be dependent. To put things in practical perspective, in a panel data model the errors may be correlated within the same panel unit over different time periods, and at the same time they may be correlated for multiple panel units observed in the same time period. Since clearly it is unfeasible to characterize one giant cluster of the entire panel, a possibility is to estimate the variance-covariance matrix of interest through the **two-way clustering formula**. Without any loss of generality, call $\mathbb{I}$ the first set of relevant groups (e.g. all panel units); $\mathbb{T}$ the second group (e.g. all time periods) and $\mathbb{J}$ the set of all elements defined by all possible *intersections* $\mathbb{I} \cap \mathbb{T}$ (in a panel dataset, these would be unique observations at the *i-t* level). The two-way clustering formula is:

$$\widehat{\mathbb{A}\text{var}}_{\mathbb{I},\mathbb{T}}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right) = \widehat{\mathbb{A}\text{var}}_{\mathbb{I}}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right) + \widehat{\mathbb{A}\text{var}}_{\mathbb{T}}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right) - \widehat{\mathbb{A}\text{var}}_{\mathbb{J}}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right) \quad (8.56)$$

where $\widehat{\mathbb{A}\text{var}}_{\mathbb{I}}(\cdot)$, $\widehat{\mathbb{A}\text{var}}_{\mathbb{T}}(\cdot)$ and $\widehat{\mathbb{A}\text{var}}_{\mathbb{J}}(\cdot)$ are, respectively, expressions of the general CCE formula (8.55) based on the groups defined by the sets $\mathbb{I}$, $\mathbb{T}$ and $\mathbb{J}$; clearly, in a panel setting the third expression – which enters negatively in (8.56) – is identical to the standard heteroscedasticity-robust covariance estimator of OLS as given in (8.22). Two-way clustering is actually a particular case of the more general **multi-way clustering**; see Cameron et al. (2011) for an extended discussion. In all cases of multi-way clustering, the relevant number of clusters to look at in order to gauge the goodness of the asymptotic approximation is that of the *smallest* set under consideration. For example, in panel data with $|\mathbb{I}| = N$ and $|\mathbb{T}| = T$, it usually is $T < N$ and T is very small (hardly larger than 20); hence, the previous discussion about the theory and practice of clustering with few groups applies.

HAC Estimation

In large samples, alternatives to CCE exist under specific structures of the cross-error dependence. Such estimators of the OLS variance-covariance go by the name of **heteroscedasticity-autocorrelation-consistent** (HAC) estimators, since they were originally devised for the case of autocorrelation in time. Like CCE estimators as well as all asymptotic covariance estimators of OLS more generally, HAC estimators are based on $K \times K$ matrices $\widehat{\boldsymbol{\Xi}}_{HAC}$ such that:

$$\widehat{\boldsymbol{\Xi}}_{HAC} \xrightarrow{p} \boldsymbol{\Xi}_0$$

and if a Central Limit Theorem for dependent observations can be applied:

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{K}_0^{-1}\boldsymbol{\Xi}_0\mathbf{K}_0^{-1}\right)$$

then, HAC estimation of the asymptotic variance-covariance of OLS can be performed as follows:

$$\widehat{\mathbb{A}\text{var}}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right) = N\left[\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right]^{-1}\widehat{\boldsymbol{\Xi}}_{HAC}\left[\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right]^{-1} \tag{8.57}$$

where N (or T) usually simplifies with N^{-1} (T^{-1}) in the expression of $\widehat{\boldsymbol{\Xi}}_{HAC}$.

The idea by Newey and West (1987) and Andrews and Monahan (1991) in the case of autocorrelated time series from a model such as $y_t = \mathbf{x}_t^{\mathrm{T}}\boldsymbol{\beta}_0 + \varepsilon_t$ (with $t = 1, \ldots, T$), is that an appropriate solution can be as follows:

$$\widehat{\boldsymbol{\Xi}}_{NW} = \sum_{s=-(T-1)}^{T-1}\kappa_T(s)\frac{1}{T}\sum_{t=1}^{T}e_t\mathbf{x}_t\mathbf{x}_{t+s}^{\mathrm{T}}e_{t+s} \tag{8.58}$$

where NW stands for "Newey-West," $e_t = y_t - \mathbf{x}_t^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{OLS}$ (and similarly for e_{t+s}), while $\kappa_T(s)$ is a **weighting kernel** decreasing in $|s|$, and such that $\kappa_T(s) = 0$ if $t + s < 1$ or $t + s > T$. The most popular weighting kernel is the **Bartlett kernel**, named after Bartlett (1950):

$$\kappa_{B_T}(s) = \left(1 - \frac{|s|}{B_T}\right)^{+}$$

where $2B_T$ is the **base** of the Kernel; note that $\kappa_T(0) = 1$ and it decreases uniformly for higher values of $|s|$ until $\kappa_T(|\widetilde{s}|) = 0$ for $\widetilde{s} \geq B_T$. The intuition behind this estimator is that of assigning to each observation in the series an interval of a certain length (like the base of the Bartlett kernel) within which autocorrelation can be nonzero; through the kernel, two close enough, possibly autocorrelated observations "interact" in the HAC estimator (8.58) of the variance-covariance of $\boldsymbol{x}_t\varepsilon_t$, just like observations of the same cluster interact in the CCE formula (8.55).

In the mentioned theoretical contributions, conditions are established in order for the HAC estimator to be a consistent estimator of $\boldsymbol{\Xi}_0$ and for the applicability of a Central Limit Theorem. Clearly, the true autocorrelation must be zero for observation pairs not captured by $\kappa_T(s)$; with the Bartlett kernel, this means that the base B_T must be long enough so to capture the actual extent of the autocorrelation. This creates an empirical tension, since of course a longer base implies a larger estimated variance and less precise estimates, while shortening the base entails the risk of underestimating the true variance. In addition, consistency of the HAC estimator requires that for integers $s > 0$, the kernel tends to zero sufficiently fast so that the overall estimate of the variance vanishes as T grows itself increasingly larger.

The Newey-West-Andrews-Monahan estimator can be easily extended to autocorrelated panel data. In a double indexed (i-t) model $y_{it} = \mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}_0 + \varepsilon_{it}$, (8.58) rewrites as:

$$\widehat{\boldsymbol{\Xi}}_{NW} = \sum_{s=-(T-1)}^{T-1} \kappa_T(s) \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} e_{it}\mathbf{x}_{it}\mathbf{x}_{i(t+s)}^{\mathrm{T}} e_{i(t+s)} \tag{8.59}$$

where $e_{it} = y_{it} - \mathbf{x}_{it}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{OLS}$ and so on. Clearly enough, in this case the kernel is allowed to cover the entire panel length T, since consistent HAC estimation follows from the asymptotic properties obtained as N grows larger. Observe that if $\kappa_T(s) = 1$ for all observations of the same panel unit and equals zero otherwise, (8.59) would coincide with the CCE formula when clusters are defined at the panel unit level. The HAC estimator is also easily ported to a setting featuring *spatial correlation*. Recall that in such a case, cross-error dependence decays with some measure of *distance* d_{ij} between observations i and j; in a standard (say, cross-sectional) model $y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + \varepsilon_i$, the HAC estimator is easily adapted as:

$$\widehat{\boldsymbol{\Xi}}_{HSC} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \kappa_N(d_{ij}) e_i\mathbf{x}_i\mathbf{x}_j^{\mathrm{T}} e_j \tag{8.60}$$

where HSC stands for heteroscedasticity and *spatial* correlation consistent, and the kernel is sufficiently decreasing in d_{ij}. This estimator was analyzed principally by Conley (1999) and Kelejian and Prucha (2007); the intuition is as in the time series case; the difference is that instead of a time interval, the kernel captures an "area" (possibly, in an abstract sense) around each observation. Similarly, (8.60) coincides with the CCE if the kernel captures only observations within segregated groups, and weighs them equally. In a panel data environment with both temporal and spatial correlation, (8.59) and (8.60) can be combined:

$$\widehat{\boldsymbol{\Xi}}_{HASC} = \sum_{s=-(T-1)}^{T-1} \kappa_T(s) \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} \kappa_N(d_{ij}) e_{it}\mathbf{x}_{it}\mathbf{x}_{j(t+s)}^{\mathrm{T}} e_{j(t+s)} \tag{8.61}$$

the appropriate comparison in this case would be with two-way clustering.

To conclude this discussion, both CCE and HAC estimators are flexible enough tools which can be utilized in the econometric practice in order to perform correct inference even under dependent errors. The choice between each depends on the context and the data, and is informed by the conditions that are necessary for the appropriate asymptotic results to be applicable. While CCE can be seen as a special case of HAC estimation, it is considerably more popular, mostly for reasons of easier practical implementation.

Lecture 9

Econometric Models

This lecture provides an introduction to structural models in econometrics, while contextually discussing the two fundamental concepts of *identification* and *causality*, which govern the choice of empirical models in the applied econometric practice. These notions are purposely introduced following the treatment of the single-equation linear model from previous lectures, which can thus be exploited as a useful source of examples.

9.1 Structural Models

A **structural econometric model** is a set of relationships regarding some socio-economic variables relative to some unit of observation (the latter is denoted here by i). The following treatment distinguishes between:

- some P **endogenous** variables: $\boldsymbol{y}_i = (Y_{1i}, Y_{2i}, \ldots, Y_{Pi})$;
- some Q **exogenous** variables: $\boldsymbol{z}_i = (Z_{1i}, Z_{2i}, \ldots, Z_{Qi})$;
- and some R **unobserved** variables (or factors) $\boldsymbol{\varepsilon}_i = (\varepsilon_{1i}, \varepsilon_{2i}, \ldots, \varepsilon_{Ri})$.

A structural model relates endogenous variables to themselves, to exogenous variables and to unobserved factors via P functional relationships.

$$\boldsymbol{y}_i = \boldsymbol{s}\left(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{\varepsilon}_i; \boldsymbol{\theta}\right) \tag{9.1}$$

Typically, these relationships combine as system of P equations, although inequalities are occasionally included. Relying either on economic theory or on *a priori* knowledge of the setting under analysis, econometricians specify the functions expressed through $\boldsymbol{s}(\cdot)$. The **parameters** that govern these relationships are collected in the vector $\boldsymbol{\theta}$, which here is given dimension K ($|\boldsymbol{\theta}| = K$), and whose parameter space is written as $\boldsymbol{\Theta}$.

The objective of econometric analysis is to characterize techniques for performing **statistical inference** about the **true value** $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, from a **data sample** $\{(\mathbf{y}_i, \mathbf{z}_i)\}_{i=1}^N$ made of N observations, and on the basis of the a-priori knowledge (informed by economic theory) of the structural function $\boldsymbol{s}(\cdot)$. To this end, the econometrician postulates adequate **distributional assumptions** about the vector of unobservables $\boldsymbol{\varepsilon}_i$ that allow to transform (9.1) into a statistical model whose parameters $\boldsymbol{\theta}$ can actually be estimated. Such distributional assumptions can be of different kinds: they range from simple restrictions on the first moment of the unobservables (e.g. the value of $\mathbb{E}[\boldsymbol{\varepsilon}_i]$), through features of their conditional distribution given $\boldsymbol{z}_i$ (say, the value of $\mathbb{E}[\boldsymbol{\varepsilon}_i | \boldsymbol{z}_i]$), up to the full-fledged specification of the joint probability distribution function of $(\boldsymbol{z}_i, \boldsymbol{\varepsilon}_i)$. This has implications in terms of what set of statistical and econometric techniques is available for estimation.

Leading examples of structural models are the (**linear**) **Simultaneous Equations Models** (SEMs) which, for $P = R$, generalize as follows.

$$\begin{aligned}
\gamma_{11}Y_{1i} + \gamma_{12}Y_{2i} + \ldots + \gamma_{1P}Y_{Pi} &= \phi_{11}Z_{1i} + \phi_{12}Z_{2i} + \ldots + \phi_{1Q}Z_{Qi} + \varepsilon_{1i} \\
\gamma_{21}Y_{1i} + \gamma_{22}Y_{2i} + \ldots + \gamma_{2P}Y_{Pi} &= \phi_{21}Z_{1i} + \phi_{22}Z_{2i} + \ldots + \phi_{2Q}Z_{Qi} + \varepsilon_{2i} \\
&\ldots = \ldots \\
\gamma_{P1}Y_{1i} + \gamma_{P2}Y_{2i} + \ldots + \gamma_{PP}Y_{Pi} &= \phi_{P1}Z_{1i} + \phi_{P2}Z_{2i} + \ldots + \phi_{PQ}Z_{Qi} + \varepsilon_{Pi}
\end{aligned}$$

In this model, the parameter set is given by $\boldsymbol{\theta} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_P; \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_P)$, where $\boldsymbol{\gamma}_p = (\gamma_{p1}, \ldots, \gamma_{pP})$ and $\boldsymbol{\phi}_p = (\phi_{p1}, \ldots, \phi_{pQ})$ for $p = 1, \ldots, P$; for the sake of making the model meaningful, certain parameters are typically normalized – usually $\gamma_{pp} = 1$ for $p = 1, \ldots, P$. A SEM can be conveniently written in **compact vectorial notation**:

$$\boldsymbol{\Gamma}\mathbf{y}_i = \boldsymbol{\Phi}\mathbf{z}_i + \boldsymbol{\varepsilon}_i \tag{9.2}$$

where $\boldsymbol{\Gamma}$ and $\boldsymbol{\Phi}$ are, respectively, matrices of dimension $P \times P$ and $P \times Q$, which collect the $\boldsymbol{\gamma}_p$ and $\boldsymbol{\phi}_p$ parameter vectors along their rows; while:

$$\mathbf{y}_i = \begin{bmatrix} y_{1i} \\ y_{2i} \\ \vdots \\ y_{Pi} \end{bmatrix}, \quad \mathbf{z}_i = \begin{bmatrix} z_{1i} \\ z_{2i} \\ \vdots \\ z_{Qi} \end{bmatrix}, \quad \boldsymbol{\varepsilon}_i = \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \vdots \\ \varepsilon_{Pi} \end{bmatrix}$$

collect the observation-specific realizations of $\boldsymbol{y}_i$ and $\boldsymbol{z}_i$ as well as the values of $\boldsymbol{\varepsilon}_i$. A simple example of a SEM was already given in Lecture 7, as the combinations of the two models (7.6) and (7.7) for the analysis of the returns to education. There, log-wages W_i and S_i are the endogenous variables, the

random vector $\boldsymbol{z}_i = (X_i, X_i^2, Z_i)$ collects the exogenous variables, whereas $\varepsilon_{1i} = \alpha_i + \epsilon_i$ and $\varepsilon_{2i} = \psi_0 \alpha_i + \eta_i$ are two "combined" unobserved factors. It is useful to make other examples of structural econometric models.

Example 9.1. The Klein I Model. SEMs were introduced by the famous "Cowles commission" back in the '40s, at a time when econometrics was just developed with the ambitious aim of creating a large macroeconomic model of the entire economy that would guarantee both full employment and no more repetitions of the Great Depression trauma. The underlying idea was that the system would let policymakers control the "endogenous" variables, such as the GDP, via the manipulation of "exogenous" policy variables such as say government expenditures. The legacy of this intellectual undertaking is controversial, but one of its heritages is a set of small, self-contained SEMs that are still used for illustrative purposes. A famous one is the "Klein I" model (Klein, 1950), which features three structural equations:

$$\begin{aligned} C_t &= \alpha_0 + \alpha_1 P_t + \alpha_2 P_{t-1} + \alpha_3 \left(W_t^p + W_t^g\right) + \varepsilon_{1t} \\ I_t &= \beta_0 + \beta_1 P_t + \beta_2 P_{t-1} + \beta_3 K_{t-1} + \varepsilon_{2t} \\ W_t^p &= \gamma_0 + \gamma_1 X_t + \gamma_2 X_{t-1} + \gamma_3 A_t + \varepsilon_{3t} \end{aligned}$$

and which is accompanied by the following **identities** (the first one is also usually seen as an **equilibrium condition**[1]):

$$\begin{aligned} X_t &= C_t + I_t + G_t \\ P_t &= X_t - T_t - W_t^p \\ K_t &= K_{t-1} + I_t \end{aligned}$$

where: *i.* C_t is consumption; *ii.* I_t is investment; *iii.* G_t is the government's nonwage expenditure; *iv.* X_t is the aggregate demand or GDP; *v.* T_t are the indirect business tax plus net exports; *vi.* K_t is the aggregate capital stock and K_{t-1} is its *lagged* value; *vii.* P_t is the aggregate level of profits realized in the private sector and P_{t-1} is its *lagged* value; *viii.* W_t^p are wages paid in the private sector; *ix.* W_t^g are wages paid in the business sector; and *x.* A_t is a constant time trend. A simplified version of first structural equation – the one for consumption – is the Keynesian consumption function from Examples 7.1 and 7.4. Observe that in this model, variables are denoted by the **time subscript** t, rather than the standard subscript i. The reason is that macroeconomic models such as this one are typically estimated on **time series** data. Here, time **lags** of specific variables, that are represented by a subscript such as $t-1$, introduce dynamics into the system. ■

[1]Specifically, $C_t + I_t + G_t$ represents the *aggregate demand* in the economy which, in equilibrium, must equal supply, resulting in an equilibrium level of output or GDP X_t.

Example 9.2. Entry Models. Among the fields of Economics, Industrial Organization – the one that focuses on the analysis of specific markets – is the one that makes more intensive use of non-linear structural models. The set of applications ranges from the estimation of demand functions, that of supply and production functions, the analysis of oligopolies, auctions, and more. Perhaps, the archetypical structural models of industrial organization are the **entry models** (or **entry games**), which concern with the analysis of **market structure** – the number of and the degree of competition among firms in a market – as a function of factors that relate to both demand and supply. A very minimalistic entry model is sketched next.

Consider N separate markets indexed as $i = 1, \dots, N$, each populated by an *endogenous* number F_i of identical firms. These may be, for example, geographically segregated markets for homogeneous goods or services. The average profit of a firm in market i, as a function of F_i, can be written as:

$$\pi_i(F_i) = \pi_{Vi}(F_i, \boldsymbol{z}_i, \nu_i; \boldsymbol{\theta}_M) - C_i \tag{9.3}$$

where $\pi_{Vi}(\cdot)$ is a **variable profits function**, whose arguments are F_i, the market's various *exogenous* characteristics $\boldsymbol{z}_i$ (as parameterized by $\boldsymbol{\theta}_M$), and some *unobserved factors* ν_i; instead, C_i are the market-specific **fixed costs**. With standard characterizations of the demand and supply functions, and of the modes of competition, $\pi_{Vi}(\cdot)$ is decreasing in F_i. Furthermore, economic theory predicts that, under complete information, **in equilibrium** as many firms will enter the market as the possibility to make positive profits allows. Therefore, the endogenous variable F_i relates to the exogenous variables $\boldsymbol{z}_i$ and to the unobserved factor ν_i as follows:

$$F_i \in \arg\min_{F \in \mathbb{N}} \pi_{Vi}(F, \boldsymbol{z}_i, \nu_i; \boldsymbol{\theta}_M) \quad \text{s.t.} \quad \pi_{Vi}(F_i, \boldsymbol{z}_i, \nu_i; \boldsymbol{\theta}_M) - C_i \geq 0 \tag{9.4}$$

which clearly results in a non-linear relationship with a specific "step function" shape – yet, the parameters $\boldsymbol{\theta}_M$ can be estimated with the appropriate econometric techniques under specific distributional assumptions. Clearly, this requires a functional form for $\pi_{Vi}(\cdot)$, which in turn depends on specific hypotheses. For example, if the demand function has a constant elasticity ζ and it is directly proportional to a measure of "market size" $\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\theta}_D + \nu_i$, where $\boldsymbol{\theta}_M = (\boldsymbol{\theta}_D, \zeta)$ and $\boldsymbol{z}_i$ are factors that affect demand (e.g. demographic characteristics); while firms have constant marginal costs and compete *à la* Cournot, then:

$$\pi_{Vi}(F_i, \boldsymbol{z}_i, \nu_i; \boldsymbol{\theta}_M) = \frac{\zeta\left(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\theta}_D + \nu_i\right)}{F_i^2} \tag{9.5}$$

which is similar to Berry (1992), and convenient for the sake of estimation.

The objective of estimating a model of this kind would be, for example, that of finding out what factors $\boldsymbol{z}_i$ best predict the profitability of a market. Extensions of such a stylized model might allow for heterogeneous firms, or for cost factors that vary across markets, hence extending the scope of the analysis towards supply factors that also affect profitability. Other models introduce include endogenous variables, incomplete information, and more; for an introduction to this literature, see Berry and Reiss (2007). ■

This completes the exposition of three quite different econometric models, each grounded on a specific piece of economic theory. The rest of these lectures is devoted to the analysis of methods for the **estimation** of models like these. Before proceeding to estimation, however, the careful econometricians should ask themselves questions of the following sort.

1. *Is it possible to use the results of my estimates for the sake of attributing unique values to each parameter within the set* $\boldsymbol{\theta}$*?*

2. *If so, is it possible to use these estimates in order to answer questions about the "effect" of certain variables upon the others?*

Questions like these lie at the core of econometric analysis. These relate, respectively, to the notions of **identification** and **causality** – while intertwined, these two concepts are often confused for one another, and it is thus useful to provide appropriate introductions to both. Most of the remainder of this lecture is devoted to this objective.

9.2 Model Identification

There are several informal definitions of "identification," all of them somehow expressing the notion that for a certain parameter set $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ to be *identified* in a statistical model, no other set $\boldsymbol{\theta}' \in \boldsymbol{\Theta}$ should have the same probabilistic implications in terms of "generating" a certain data sample. In econometrics, the concept of identification originates with the analysis of Simultaneous Equations Models where – as it is elaborated later – short of ex-ante assuming specific restrictions, an infinite number of parameter sets is typically equally capable of rationalizing *a posteriori* the same data. This idea can be intuitively connected to example 7.2: if education S_i and wages W_i are related in the data, is it because of some "effect" of the former on the latter (β_3) or due to the indirect effect of ability α_i? A classical, formal definition of identification is the one by Rothenberg (1971), developed in the context of a **fully parametric model** – one where the joint probability distribution generating the data is fully specified.

Definition 9.1. A **data generation process** (DGP) is the joint probability distribution $F_{\boldsymbol{\theta}}(\boldsymbol{z}_i, \boldsymbol{\varepsilon}_i)$ parametrized by $\boldsymbol{\theta}$ or, given (9.1), $G_{\boldsymbol{\theta}}(\boldsymbol{z}_i, \boldsymbol{y}_i)$.

Definition 9.2. A **family** $\mathcal{P}$ of DGPs is some given set of similar DGPs.

Definition 9.3. A **structure** $\boldsymbol{\theta}'$, is a specific restriction on $\boldsymbol{\theta}$ that uniquely determines a particular DGP $\mathcal{P}_{\boldsymbol{\theta}'}(\boldsymbol{z}_i, \boldsymbol{\varepsilon}_i) \in \mathcal{P}$.

Definition 9.4. A **statistical model** $\mathcal{M}$ the set of valid structures, which needs not to be equivalent with the family of DGPs $\mathcal{P}$. A statistical model $\mathcal{M}$ is best understood as the set of structures $\mathcal{M} \subset \mathcal{P}$ compatible with the restrictions implied in the "structural" model (9.1).

Example 9.3. Parametric Bivariate Regression. Consider a bivariate linear model analogous to the one from Examples 3.11, 6.3 and 6.7:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where the data are generated according to a well-known family $\mathcal{P}$ of DGPs, a bivariate normal distribution:

$$\begin{pmatrix} X_i \\ \varepsilon_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_x \\ \mu_\varepsilon \end{pmatrix}; \begin{pmatrix} \sigma_x^2 & \sigma_{x\varepsilon} \\ \sigma_{x\varepsilon} & \sigma_\varepsilon^2 \end{pmatrix} \right)$$

implying $Y_i \sim \mathcal{N}(\beta_0 + \beta_1\mu_x + \mu_\varepsilon, \beta_1^2\sigma_x^2 + 2\beta_1\sigma_{x\varepsilon} + \sigma_\varepsilon^2)$. To operationalize this model, one usually imposes the restriction $\mu_\varepsilon = \mathbb{E}[\varepsilon_i] = 0$. The statistical model $\mathcal{M}$ is the set of admissible structures $\boldsymbol{\theta} = (\beta_0, \beta_1, \mu_x, \sigma_x^2, \sigma_\varepsilon^2, \sigma_{x\varepsilon})$. An example of structure is the restriction $\boldsymbol{\theta}_0 = (5, 2, 0, 2, 2, 1)$. ■

Quoting Rothenberg, "the identification problem concerns the existence of a unique inverse association" from the data to the structure. That is, it is a question about the possibility of recovering one exact structure $\boldsymbol{\theta}$ when knowing the complete probability distribution of the data. Whether this is possible in theory determines what econometric techniques are available for estimation, if any. Some further definitions are in order.

Definition 9.5. Observational Equivalence. Two structures $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}''$ are *observationally equivalent* if $\mathbb{P}(\boldsymbol{y}_i, \boldsymbol{z}_i | \boldsymbol{\theta}') = \mathbb{P}(\boldsymbol{y}_i, \boldsymbol{z}_i | \boldsymbol{\theta}'')$.

Definition 9.6. Global Identification. A Structure $\boldsymbol{\theta}' \in \boldsymbol{\Theta}$ is *globally point identified* if there is no other structure $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ that is observationally equivalent to it.

Definition 9.7. Local Identification. A Structure $\boldsymbol{\theta}' \in \boldsymbol{\Theta}$ is *locally point identified* if there is no other structure in an open neighborhood of $\boldsymbol{\theta}'$ that is observationally equivalent to it.

One additional notion, that of set identification, is outside the scope of this discussion and thus left aside. For simplicity, the term "identification" henceforth denotes more generally the notion of *global point identification*.

Definition 9.8. Model Identification. An econometric model $\mathcal{M}$ is identified if all its Structures $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ are identified.

Armed with these definitions, one can provide rigorous answers to the question whether some models are "identified" or not.

Theorem 9.1. Identification of a fully parametric bivariate regression. *The statistical model $\mathcal{M}$ from Example 9.3 is not (point) identified. However, the* restricted *model given by $\mathcal{M}' = \{\boldsymbol{\theta} \in \mathcal{M} : \sigma_{x\varepsilon} = 0\}$ is instead (point) identified.*

Proof. (Sketched.) Here it is most convenient to proceed in steps. The first step is about showing how parameters (μ_x, σ_x^2) are always identified given appropriate observations of X_i. To this end, a specific rule that associates observations to parameter values is necessary; since the model in question is fully parametric, the *likelihood principle* (see Lecture 5) appears the most straightforward choice. Specifically, to any (set of) observations of X_i that are drawn from some distribution $F_{X_i}(x_i|\,\boldsymbol{\theta})$, the parameters $\boldsymbol{\theta}$ chosen to rationalize the data are those that maximize the (log-)likelihood function $\log \mathcal{L}(\boldsymbol{\theta}|\, x_1, \ldots, x_N)$. One can resort to the Implicit Function Theorem to establish that such an association exists under mild conditions. By following a similar approach one can show how $(\beta_0, \beta_1, \sigma_\varepsilon^2, \sigma_{x\varepsilon})$ are not identified under analogous conditions. The full-fledged demonstration of identification under the restriction that $\sigma_{x\varepsilon} = 0$ is however left as an exercise.

First, consider the information about X_i contained in a sample of size N, that is the collection of realizations $\{x_1, \ldots, x_N\}$ of X_i which – importantly – must *not be all identical to each other* (however, this occurrence has probability zero under the maintained hypotheses). The following log-likelihood function should resonate as familiar, as the sample is drawn from the normal distribution.

$$\log \mathcal{L}\left(\mu_x, \sigma_x^2 \middle|\, x_1, \ldots, x_N\right) = -\frac{N}{2} \log 2\pi\sigma_x^2 - \sum_{i=1}^{N} \frac{(x_i - \mu_x)^2}{2\sigma_x^2}$$

Example 5.6 outlines the First and Second conditions for a maximum of this log-likelihood function, and the consequent expressions for estimators of the two parameters, call them $\widehat{\mu}_x$ and $\widehat{\sigma}_x^2$. The question of identification here is: "can these conditions [for a maximum] characterize a *univocal* association

from the sample to the parameters, $(\widehat{\mu}_x, \widehat{\sigma}_x^2) : \mathbb{S}_x \to \mathbb{R} \times \mathbb{R}_{++}$?" To answer this question, recall again from example 5.6 that the Jacobian matrix of the score, that is the Hessian matrix of the log-likelihood function, is evaluated *at the solution* as follows.

$$\mathbf{H}\left(\widehat{\mu}_x, \widehat{\sigma}_x^2 \middle| x_1, \ldots, x_N\right) = -N \begin{bmatrix} \widehat{\sigma}_x^{-2} & 0 \\ 0 & 2\widehat{\sigma}_x^{-4} \end{bmatrix}$$

Since $\widehat{\sigma}_x^2 \neq 0$ the determinant is nonzero, hence the Hessian has full rank. Thus, by the Implicit Function Theorem it is (almost) always possible to solve for unique values of (μ_x, σ_x^2): these parameters are identified.

Second, consider the log-likelihood function of $\vartheta = (\beta_0, \beta_1, \sigma_\varepsilon^2, \sigma_{x\varepsilon})$ given the information about Y_i contained in the sample $\{(y_i, x_i)\}_{i=1}^N$ (here one can abstract from μ_x and σ_x^2 as they are shown to be identified).

$$\log \mathcal{L}\left(\vartheta \middle| y_1, \ldots, y_N, x_1, \ldots, x_N\right) = -\frac{N}{2} \log 2\pi \left(\beta_1^2 \sigma_x^2 + 2\beta_1 \sigma_{x\varepsilon} + \sigma_\varepsilon^2\right) - \\ - \sum_{i=1}^N \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\left(\beta_1^2 \sigma_x^2 + 2\beta_1 \sigma_{x\varepsilon} + \sigma_\varepsilon^2\right)}$$

The First Order Conditions now read as:

$$\frac{\partial \log \mathcal{L}\left(\widehat{\vartheta} \middle| y_1, \ldots, y_N, x_1, \ldots, x_N\right)}{\partial \beta_0} = \sum_{i=1}^N \frac{e_i}{\widehat{\sigma}_y^2} = 0$$

$$\frac{\partial \log \mathcal{L}\left(\widehat{\vartheta} \middle| y_1, \ldots, y_N, x_1, \ldots, x_N\right)}{\partial \beta_1} = -\frac{N\left(\widehat{\beta}_1 \widehat{\sigma}_x^2 + \widehat{\sigma}_{x\varepsilon}\right)}{\widehat{\sigma}_y^2} + \sum_{i=1}^N \frac{e_i x_i}{\widehat{\sigma}_y^2} + \\ + \sum_{i=1}^N \frac{e_i^2\left(\widehat{\beta}_1 \widehat{\sigma}_x^2 + \widehat{\sigma}_{x\varepsilon}\right)}{\widehat{\sigma}_y^4} = 0$$

$$\frac{\partial \log \mathcal{L}\left(\widehat{\vartheta} \middle| y_1, \ldots, y_N, x_1, \ldots, x_N\right)}{\partial \sigma_\varepsilon^2} = -\frac{N}{2\widehat{\sigma}_y^2} + \sum_{i=1}^N \frac{e_i^2}{2\widehat{\sigma}_y^4} = 0$$

$$\frac{\partial \log \mathcal{L}\left(\widehat{\vartheta} \middle| y_1, \ldots, y_N, x_1, \ldots, x_N\right)}{\partial \sigma_{x\varepsilon}} = -2\widehat{\beta}_1 \left(\frac{N}{2\widehat{\sigma}_y^2} - \sum_{i=1}^N \frac{e_i^2}{2\widehat{\sigma}_y^4}\right) = 0$$

where $\widehat{\sigma}_y^2 \equiv \widehat{\beta}_1^2 \widehat{\sigma}_x^2 + 2\widehat{\beta}_1 \widehat{\sigma}_{x\varepsilon} + \widehat{\sigma}_\varepsilon^2$ and $e_i \equiv y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$ for $i = 1, \ldots, N$. Clearly, as the two derivatives with respect to σ_ε^2 and $\sigma_{x\varepsilon}$ are linearly dependent, no Jacobian matrix of full rank can be formed out of the First Order Conditions. Therefore, parameters $\vartheta = (\beta_0, \beta_1, \sigma_\varepsilon^2, \sigma_{x\varepsilon})$ are not identified.

A useful exercise is to show, following the example above about the identification of (μ_x, σ_x^2), that the model is identified when the restriction $\sigma_{x\varepsilon} = 0$ is imposed. It is easiest to start from a simpler case, where X_i is "fixed in repeated samples" (that is, some N realizations occur with probability one) which greatly simplifies the expression of the likelihood function. □

The identification condition $\sigma_{x\varepsilon} = 0$ from Example 9.3, which states that the covariance between X_i and ε_i must be zero, is intimately connected with the so-called "exogeneity" condition of the linear regression model, which is abundantly discussed in other lectures, but it is also worth to revisit it here. This condition requires that the expectation of the error term conditional on the explanatory variables is zero (here, $\mathbb{E}[\varepsilon_i | X_i = x_i] = 0$ for all $x_i \in \mathbb{X}$) and it implies that the CEF of Y_i given X_i is linear as well:

$$\begin{aligned}\sigma_{x\varepsilon} = \mathbb{C}\text{ov}(X_i, \varepsilon_i) &= \mathbb{E}[X_i\varepsilon_i] - \mathbb{E}[X_i]\,\mathbb{E}[\varepsilon_i] \\ &= \mathbb{E}_X[\mathbb{E}[X_i\varepsilon_i | X_i]] \\ &= 0\end{aligned}$$

because $\mathbb{E}[\varepsilon_i] = 0$ and by the Law of Iterated Expectations (Example 3.11). In abstract terms, the intuition can be formulated as follows: if $\sigma_{x\varepsilon} \neq 0$, it is apparent that X_i and Y_i move together it is impossible to tell whether they do because X_i affects Y_i directly, or rather through the indirect influence of ε_i. Clearly, here something has to give, and to properly interpret the data it is necessary to place some "restriction" on the statistical model $\mathcal{M}$.

The concept of identification is not restricted to fully parametric models. Indeed, it can apply as well to **semi-parametric models**: that is, models in which only some features of the joint probability distribution of $(\boldsymbol{z}_i, \boldsymbol{\varepsilon}_i)$ is specified. A full-fledged treatment of identification in the semi-parametric case is outside the scope of this discussion, but it is still worth to illustrate the main intuition via an example.

Example 9.4. Parametric Bivariate Regression. Consider once again the bivariate linear model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, but unlike in Example 9.3, abstain from imposing fully parametric assumptions: the joint distribution of (X_i, ε_i) is left unspecified. In this case one could re-define the concept of "model" $\mathcal{M}$ as a set of structures of the kind

$$\boldsymbol{\theta} = \left(\beta_0, \beta_1, \mathcal{P}_x, \mathcal{P}_\varepsilon, \mathcal{P}_{\varepsilon|x}\right) \tag{9.6}$$

where $\mathcal{P}_x$, $\mathcal{P}_\varepsilon$ and $\mathcal{P}_{\varepsilon|x}$ are *families of probability distributions*, respectively of X_i, of ε_i, and of ε_i conditional on X_i, that are allowed by the model $\mathcal{M}$. A straightforward restriction here is that all elements of $\mathcal{P}_\varepsilon$ must conform to $\mathbb{E}[\varepsilon_i] = 0$; clearly, an unrestricted mean is indistinguishable from β_0. ■

The definitions of observational equivalence and identification are extended here to unique values of the exact parameters – like (β_0, β_1) – and families of distributions that are compatible with the data. In analogy with the fully parametric case, identification of a semi-parametric bivariate linear model is possible by imposing a restriction on the elements from family $\mathcal{P}_{\varepsilon|x}$, one with the same intuitive interpretation as in the fully parametric case. Unsurprisingly, the restriction in question is of the form $\mathbb{E}[\varepsilon_i | X_i] = 0$.

Theorem 9.2. Identification of a semi-parametric bivariate regression. *Consider the semi-parametric model $\mathcal{M}$ from Example 9.4, incorporating the restriction $\mathbb{E}[\varepsilon_i] = 0$; this model is not identified. However, the* restricted *model $\mathcal{M}' = \{\boldsymbol{\theta} \in \mathcal{M} : \mathbb{E}[\varepsilon_i | X_i] = 0\}$ is identified.*

Proof. (Outline.) This case cannot be evaluated by the likelihood principle because clearly, without fully parametric assumptions, a likelihood function cannot be specified. Instead, it is necessary to analyze the *cross moments* of the model (like covariances) that involve X_i and ε_i, and evaluate whether a unique association from the data to the parameters can be established by the analogy principle (see Lecture 5 again). In general, $\mathcal{M}$ is not identified because it allows for $\mathbb{E}[X_i \varepsilon_i] = g(X_i)$ where $g(X_i)$ is some function of X_i. Consequently, a zero moment condition of the form $\mathbb{E}[X_i \varepsilon_i - g(X_i)] \neq 0$ is uninformative, as the analyst does not generally know $g(X_i)$ – or else this knowledge can be used to impose some restriction. On the other hand, $\mathcal{M}'$ is identified by familiar arguments: moments (3.8) and (3.9) and their sample analogues can be used to establish a unique association from the data to β_0 and β_1. The identification of $\mathcal{P}_x$, $\mathcal{P}_\varepsilon$, and $\mathcal{P}_{\varepsilon|x}$ under the stated restriction, which is necessary to formally show identification of $\boldsymbol{\theta}$, is then trivial, since all non-degenerate joint distributions that allow for mean independence of the error term comply with the definition. □

In many situations, however – whether one is working with a fully parametric model or a semi-parametric one – the problems of identification that emerge do not depend on the statistical relationship between observed and unobserved variables, but rather on the very structural relationships implied by one's model. An example of this sort is the "dummy variable trap" that was briefly introduced in Lecture 7; one cannot estimate a linear model where the columns of $\mathbf{X}$ are linearly dependent, because there is no unique Least Squares solution. This is intuitively related to the formal definition of identification: in such cases, there exists an infinite number of parameter combinations which are equally capable of making sense of the data. It is for this reason that White's Assumption 3, the *identification assumption* of OLS, must be imposed in order to explicitly rule out this kind of issues.

9.3 Linear Simultaneous Equations

The implications of a structural model's functional relationships on identification are especially meaningful for those models that feature multiple endogenous variables ($P \geq 2$) such as linear SEMs. Analyzing the latter is especially useful for both their pedagogical value and their ubiquitous appearance in the applied practice, whether implicit or explicit. All methods of instrumental variable estimation like those discussed in Lecture 10 are, in fact, based on SEMs (although often implicitly). This section is devoted specifically to the analysis of identification in SEMs, which is illustrated via a classical example about partial equilibrium in a market. To begin, some definitions (which are not specific to SEMs) are in order.

Definition 9.9. The **reduced form** of a structural econometric model is its solution for $\boldsymbol{y}_i$.

$$\boldsymbol{y}_i = \boldsymbol{r}\left(\boldsymbol{z}_i, \boldsymbol{\varepsilon}_i; \boldsymbol{\theta}\right) \tag{9.7}$$

Definition 9.10. A **separable structural model** is one that possesses a reduced form representation like (9.7).

For example, SEMs are separable if the parameter matrix $\mathbf{\Gamma}$ from expression (9.2) is invertible; in such a case, the reduced form can be written as:

$$\begin{aligned} \mathbf{y}_i &= \mathbf{\Gamma}^{-1}\left(\mathbf{\Phi}\mathbf{z}_i + \boldsymbol{\varepsilon}_i\right) \\ &= \mathbf{\Pi}\mathbf{z}_i + \boldsymbol{\eta}_i \end{aligned} \tag{9.8}$$

where $\mathbf{\Pi} \equiv \mathbf{\Gamma}^{-1}\mathbf{\Phi}$ is a $P \times Q$ matrix of **reduced form parameters** π_{pq} in its entries indexed by p (rows) and q (columns), and $\boldsymbol{\eta}_i \equiv \mathbf{\Gamma}^{-1}\boldsymbol{\varepsilon}_i$.

Example 9.5. Demand and Supply. Consider a standard microeconomic model of partial equilibrium in a single market, which an econometrician is trying to analyze by looking at a sample of N different *markets*, which contain information about prices and quantities, and that – similarly to example 9.2 – are indexed by $i = 1, \ldots, N$. We know that both demand Q_i^D and supply Q_i^S of a specific good are functions of price P_i. The econometrician assumes that, in particular, both the demand and the supply functions are linear up to some unobserved factors expressed as $\left(\upsilon_i^D, \upsilon_i^S\right)$.

$$\begin{aligned} Q_i^D &= \alpha_0 + \alpha_1 P_i + \upsilon_i^D \\ Q_i^S &= \beta_0 + \beta_1 P_i + \upsilon_i^S \end{aligned} \tag{9.9}$$

We learn from economic theory that, in a market, demand and supply meet in equilibrium, thus:

$$Q_i^D = Q_i^S = Q_i$$

and that both equilibrium prices P_i and quantities Q_i are determined simultaneously and interdependently, hence they are both *endogenous*.

The parameters $\boldsymbol{\theta} = (\alpha_0, \alpha_1, \beta_0, \beta_1)$ of model (9.9) are **not identified**. This is easily shown via the **reduced form** of the structural model:

$$\begin{aligned} Q_i &= \frac{\beta_1\alpha_0 - \alpha_1\beta_0}{\beta_1 - \alpha_1} + \frac{\beta_1 \upsilon_i^D - \alpha_1 \upsilon_i^S}{\beta_1 - \alpha_1} \\ P_i &= \frac{\alpha_0 - \beta_0}{\beta_1 - \alpha_1} + \frac{\upsilon_i^D - \upsilon_i^S}{\beta_1 - \alpha_1} \end{aligned} \tag{9.10}$$

exhibiting how, *by construction*, $\mathbb{E}\left[\upsilon_i^D, \upsilon_i^S \middle| P_i\right] \neq 0$ and $\mathbb{E}\left[\upsilon_i^D, \upsilon_i^S \middle| Q_i\right] \neq 0$. Thus, the parameters of the two bivariate regression models featured in the structural form (9.9) **cannot be identified**, neither in fully parametric nor in semi-parametric environments (Theorems 9.1 and 9.2). The best one can do is to *exploit the reduced form* to estimate the two unconditional moments $\mathbb{E}[Q_i] = (\beta_1\alpha_0 - \alpha_1\beta_0)/(\beta_1 - \alpha_1)$ and $\mathbb{E}[P_i] = (\alpha_0 - \beta_0)/(\beta_1 - \alpha_1)$ that are implied by (9.10), which obviously do not contain enough information about each element of $\boldsymbol{\theta}$: the system is not identified in the sense that there is an *infinite number* of $\boldsymbol{\theta}$ combinations that predict these two unconditional averages. This is represented in graphical form in Figure 9.1.

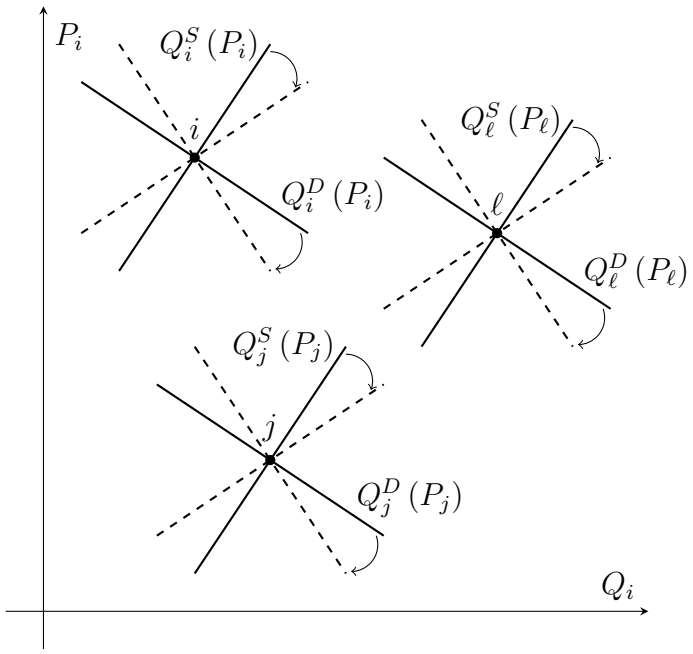


Figure 9.1: Infinite supply and demand curves given the sample $\{i, j, \ell\}$

The economic intuition behind this negative result is that changes in the equilibrium price and quantity in one market cannot be attributed to either demand or supply factors in isolation, absent any further information that is specific to either supply or demand. ■

In situation of this kind, which are ubiquitous in econometrics, the task that econometricians face is to elaborate extensions or **restrictions** of their model such that certain information contained in the data can improve on identification. Depending on the circumstances, this may as well result only on some parameters being identified, or rather on some parameters being "redundantly" identified, or both! Here, more definitions are in order.

Definition 9.11. Exact identification. An econometric model is *exactly* or *just* identified (the two expressions are interchangeable) if there exists a *unique* association from the data to the parameter set $\boldsymbol{\theta}$.

Definition 9.12. Partial identification. An econometric model is *partially* identified if there exists a *unique* association from the data to a *subset* of the parameter set $\boldsymbol{\theta}$ ($\boldsymbol{\theta}^* \subset \boldsymbol{\theta}$), but not so for the other parameters.

Definition 9.13. Overidentification. In an econometric model, a *subset* $\boldsymbol{\theta}^{**}$ of the parameter set $\boldsymbol{\theta}$ ($\boldsymbol{\theta}^{**} \subset \boldsymbol{\theta}$) is *overidentified* if there exist *multiple* associations from the data to the parameter subset in question.

In a partially identified model, there may as well be some overidentified parameters coexisting with non-identified ones. This is again best illustrated via the previous example about markets in partial equilibrium.

Example 9.6. Demand and Supply (continued). Introduce a new variable to Example 9.5: the *exogenous* income M_i of all consumers in a market. By standard microeconomic theory, consumers' demand depends positively on M_i. Furthermore, there is no theoretical reason why consumers' income should affect the production process and the supply function *directly*. Write:

$$\begin{aligned} Q_i &= \alpha_0 + \alpha_1 P_i + \alpha_2 M_i + \upsilon_i^D \\ Q_i &= \beta_0 + \beta_1 P_i + \upsilon_i^S \end{aligned} \tag{9.11}$$

with reduced form:

$$\begin{aligned} Q_i &= \frac{\beta_1\alpha_0 - \alpha_1\beta_0}{\beta_1 - \alpha_1} + \frac{\beta_1\alpha_2}{\beta_1 - \alpha_1} M_i + \frac{\beta_1\upsilon_i^D - \alpha_1\upsilon_i^S}{\beta_1 - \alpha_1} \\ P_i &= \frac{\alpha_0 - \beta_0}{\beta_1 - \alpha_1} + \frac{\alpha_2}{\beta_1 - \alpha_1} M_i + \frac{\upsilon_i^D - \upsilon_i^S}{\beta_1 - \alpha_1} \end{aligned} \tag{9.12}$$

note that if $\mathbb{E}[\upsilon_D, \upsilon_S | M_i] = 0$, the two equations in (9.12) can be treated as two separate bivariate linear regressions whose parameters are identified; in particular, through the slopes one can identify the following two parameter combinations: $\beta_1\alpha_2/(\beta_1 - \alpha_1)$ as well $\alpha_2/(\beta_1 - \alpha_1)$. This revised model is **partially identified**: β_1, the slope of the structural supply equation, is backed out as the unique ratio between those two quantities.

The intuition for this result is that now there is an "exogenous demand shifter" that allows to isolate supply effects as demand changes: this idea is illustrated graphically in Figure 9.2. Specifically, the *variation* in M_i (in the Figure, from M_i to M_j to M_ℓ), taking the supply curve as given, allows to identify different market equilibrium points, which in turn help trace the supply curve itself. It is easy to see that if M_i had been included into the supply equation, identification of β_1 would not be achieved, since it would be impossible in principle to distinguish between the influence that M_i has on demand against its effect on supply. Therefore, identification hinges on a *restriction* imposed upon the model, justified here by economic theory. Also observe that other parameters of the model, such as α_1 – the slope of the demand function – are not identified (hence "partial" identification).

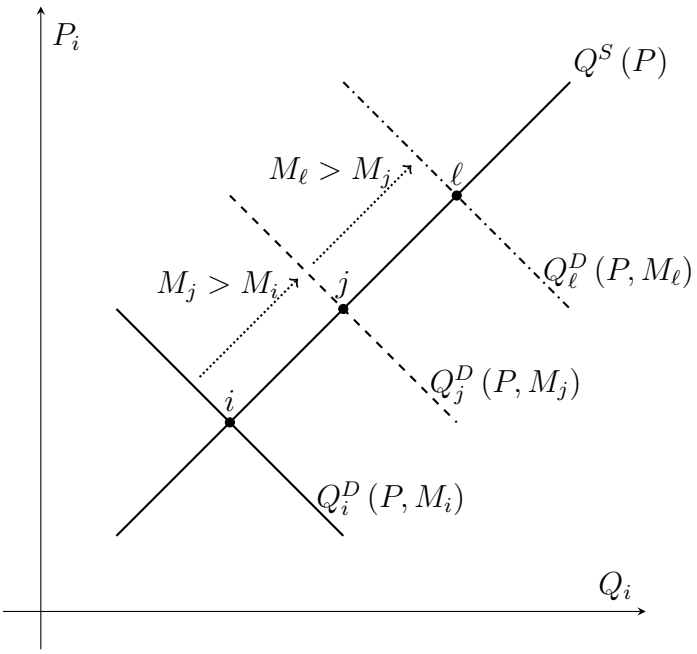


Figure 9.2: Identification of the supply curve via demand shifters

Suppose one can also observe another variable denoted by C_i that represents a synthetic index of production costs in this market. Clearly, C_i affects

supply but not demand, and therefore it can be treated as *exogenous*. The model now reads as follows.

$$\begin{aligned} Q_i &= \alpha_0 + \alpha_1 P_i + \alpha_2 M_i + \upsilon_i^D \\ Q_i &= \beta_0 + \beta_1 P_i + \beta_2 C_i + \upsilon_i^S \end{aligned} \tag{9.13}$$

The reduced form can be expressed in terms of two multivariate regressions, one for quantity Q_i and the other for price P_i, with the exogenous variables M_i and C_i showing up on the right-hand side in both cases.

$$\begin{aligned} Q_i &= \frac{\beta_1\alpha_0 - \alpha_1\beta_0}{\beta_1 - \alpha_1} + \frac{\beta_1\alpha_2}{\beta_1 - \alpha_1} M_i - \frac{\alpha_1\beta_2}{\beta_1 - \alpha_1} C_i + \frac{\beta_1\upsilon_i^D - \alpha_1\upsilon_i^S}{\beta_1 - \alpha_1} \\ P_i &= \frac{\alpha_0 - \beta_0}{\beta_1 - \alpha_1} + \frac{\upsilon_2}{\beta_1 - \alpha_1} M_i - \frac{\beta_2}{\beta_1 - \alpha_1} C_i + \frac{\upsilon_i^D - \upsilon_i^S}{\beta_1 - \alpha_1} \end{aligned} \tag{9.14}$$

If $\mathbb{E}\left[\upsilon_D, \upsilon_S | \, M_i\right] = \mathbb{E}\left[\upsilon_D, \upsilon_S | \, C_i\right] = 0$, multivariate regression techniques (as discussed later) allow to back up all the six combined parameters of (9.14). It is easy to verify that there is a unique solution that maps this set onto the set of the original "structural" parameters $\boldsymbol{\theta}_{MC} = (\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2)$. The model is thus **exactly identified**. This result fails either if C_i enters the demand equation, or if M_i does not enter the model while C_i does (in this case, however, α_1 would be identified: C_i would act as a "supply shifter" that allows to map the demand curve, symmetrically to the scenario above).

In order to appreciate an instance of **overidentification**, and how it can coexist with partial identification, let us consider one final case, that is a model without C_i – no demand shifter – but with two supply shifters: consumers' income M_i and the price of a competing product P_i^*, which is expected to affect demand positively. The model now reads as:

$$\begin{aligned} Q_i &= \alpha_0 + \alpha_1 P_i + \alpha_2 M_i + \alpha_3 P_i^* + \upsilon_i^D \\ Q_i &= \beta_0 + \beta_1 P_i + \upsilon_i^S \end{aligned} \tag{9.15}$$

and its reduced form as:

$$\begin{aligned} Q_i &= \frac{\beta_1\alpha_0 - \alpha_1\beta_0}{\beta_1 - \alpha_1} + \frac{\beta_1\alpha_2}{\beta_1 - \alpha_1} M_i + \frac{\beta_1\alpha_3}{\beta_1 - \alpha_1} P_i^* + \frac{\beta_1\upsilon_i^D - \alpha_1\upsilon_i^S}{\beta_1 - \alpha_1} \\ P_i &= \frac{\alpha_0 - \beta_0}{\beta_1 - \alpha_1} + \frac{\alpha_2}{\beta_1 - \alpha_1} M_i + \frac{\alpha_3}{\beta_1 - \alpha_1} P_i^* + \frac{\upsilon_i^D - \upsilon_i^S}{\beta_1 - \alpha_1} \end{aligned} \tag{9.16}$$

which, by reasoning analogous to those from the previous cases, has some interesting implications. First, notice that there are multiple ways to calculate β_1 (by taking the ratio of the two coefficients for M_i and P_i^*, respectively): this means that parameter β_1 is **overidentified**. Second, parameter α_1 is not identified (this is easy to verify): the intuitive reason is that there are no longer any "demand shifters" in the model. ■

This example is useful in two major respects. First, it showcases examples of both partial identification and overidentification, and as these two – as it was anticipated – can actually coexist in the same model. Some general implications of overidentification are elaborated along the analysis of the Generalized Method of Moments (Lecture 12), a framework which includes SEMs as a particular case. In particular, it is discussed how the so-called "overidentifying restrictions" – that is, the implicit constraints featured in the model that give rise to overidentification[2] – can actually be tested statistically. This may help identify problems with the structural assumptions of the model. In the last model from Example 9.6, for instance, if the two ratios that allow to identify β_1 are very different to one another, it may be a signal about "something wrong" with the setup of the model.

Second, the example outlines a strategy for identifying SEMs. The main issue is that due to the simultaneous structure of the model, the equations of the structural form do not satisfy the requirements for semi-parametric identification (a more statistical argument is developed later in Lecture 10). Conversely, the model's reduced form is identified; its P equations read as:

$$Y_{pi} = \pi_{p1} Z_{1i} + \pi_{p2} Z_{2i} + \cdots + \pi_{pQ} Z_{Qi} + \eta_{pi} \tag{9.17}$$

for $p = 1, \ldots, P$. To verify whether any SEM is identified, one must check whether some association form the parameters $\mathbf{\Pi}$ of the reduced form to the structural parameters $(\mathbf{\Gamma}, \mathbf{\Phi})$ can be established. The main problem is one of dimensionality: while $\mathbf{\Pi}$ has dimension PQ, the structural form features up to $P(P+Q)$ parameters. Unless one imposes some more "constraints" on the model, establishing such an association is mathematically impossible. Even under the typical normalization of the diagonal of $\mathbf{\Gamma}$, that is $\gamma_{pp} = 1$ for $p = 1, \ldots, P$, one needs at least $P(P-1)$ more such constraints.

The most typical kind of constraint, or **restriction**, that is placed on a SEM to obtain identification is an **exclusion restriction**, that is to force certain structural parameters to equal zero, thus ruling out specific structural relationships between variables. For example, setting some element of $\mathbf{\Phi}$ equal to zero ($\phi_{pq} = 0$) means that the q-th exogenous variable bears no effect on the p-th endogenous variable. The example on demand and supply showcases several exclusion restrictions that involve the exogenous variables M_i, C_i and P_i^*. Exclusion restrictions and the associated terminology are common in the theory and practice of Instrumental Variables (Lecture 10); the reason is that, as already hinted, the latter are grounded on SEMs.

[2] In model (9.15), overidentification of parameter β_1 actually follows from the fact that *both* supply shifters M_i and P_i^* are *restricted* to the supply function – that is, they do not show up in the structural demand function. This idea is clearly more general.

After imposing some exclusion restrictions, one can evaluate identification for each equation of a SEM through a pair of intertwined mathematical conditions. They are the so-called **order condition**, which is a necessary one, and the associated **rank condition**, which completes the other. They jointly characterize the algebraic properties that allow to solve for each row of the following $P \times (P + Q)$ matrix, which "horizontally binds" $\mathbf{\Gamma}$ and $\mathbf{\Phi}$.

$$\mathbf{F} \equiv \begin{bmatrix} \mathbf{\Gamma} & \mathbf{\Phi} \end{bmatrix}$$

In what follows, the identification conditions are formulated in terms of the number of exclusion restrictions imposed on each structural equation; they are ultimately adapted from standard (but tedious) linear algebra results.

Order condition for SEM identification. Define ϱ_p as be the number of restrictions applied to the p-th equation of the structural form.

- if $\varrho_p < P - 1$, the p-th equation is *not identified*;
- if $\varrho_p = P - 1$, the p-th equation is *exactly identified*, as long as the rank condition holds;
- if $\varrho_p > P - 1$, the p-th equation is *overidentified*, as long as the rank condition holds.

Rank condition for SEM identification. If the order condition holds, the p-th Equation is identified if at least one nonzero determinant of order $(P - 1)(P - 1)$ can be constructed out of the coefficients of the variables excluded from that equation but included in other equations in the model.

To check the rank condition for the p-th equation, one should:

1. delete from $\mathbf{F}$ the columns corresponding to the variables *included* in the p-th equation; and
2. delete row p too, which results in some submatrix $\mathbf{F}_p$.

Submatrix $\mathbf{F}_p$ should have full row rank for the rank condition to hold in the p-th equation. Note that while the two conditions are formulated with respect to the number of endogenous variables Q, but they can alternatively be expressed in terms of the number of exogenous variables Q. Together, the order and the rank condition determine a **sufficient** condition for the identification of SEMs, which is stated next. Its central implication is that the identification of SEMs rests on "equation-exclusive" exogenous variables, which are analogous to the demand and supply shifters from Example 9.6, and that are also called **instruments** for reasons that are clarified through the discussion of Instrumental Variables in Lecture 10.

Theorem 9.3. Sufficient Condition for Exact Identification. *A SEM is* at least *exactly identified if every equation of the structural form features an exogenous variable that does not show up in any other equation.*

Proof. (Exercise!) This proof is a straightforward and instructive application of the order and rank conditions, and it is best left as an exercise. □

Having learned that a SEM is identified, a researcher may want to estimate its parameters. If all the equations are exactly identified, an intuitive approach is to estimate the reduced form parameters $\mathbf{\Pi}$ via OLS and then solve for the structural parameters $\mathbf{\Gamma}$ and $\mathbf{\Phi}$; this method is called **Indirect Least Squares** (ILS) but is clearly unsuited to overidentified models. The general approach for estimating a SEM is based on the Two Stages and the Three Stages Least Squares estimators, and it is described in Lecture 10.

9.4 Causal Effects

Separable structural models that possess a reduced form representation such as (9.7) allow to characterize the concept of "causality" in econometrics. In order to define population-wide, "average" *causal effects*, it is necessary to start from the individual-level concept. Suppose then that some exogenous variable of interest Z_{qi}, $q = 1, \ldots, Q$ can be varied on its own support $\mathbb{X}_{\boldsymbol{z}q}$ independently of other exogenous variables as well as of unobservables. In this environment it is possible to define the causal effects of interest.

Definition 9.14. Individual Causal Effect. Consider the unit of observation i, whose realizations of the observable and unobservable factors are written as $(\mathbf{z}_i, \boldsymbol{\varepsilon}_i)$. Let z_{qi} be the q-th element of $\mathbf{z}_i$ and $\mathbf{z}_{-i}$ be the collection of all the other $Q-1$ elements in that vector. The *individual causal effect* of the exogenous variable Z_{qi} on the endogenous variable Y_{pi} *for unit* i is:

$$\mathcal{C}_{qpi}\left(z_{qi}, \mathbf{z}_{-qi}, \boldsymbol{\varepsilon}_i\right) = r_p\left(z'_{qi}, \mathbf{z}_{-qi}, \boldsymbol{\varepsilon}_i\right) - r_p\left(z_{qi}, \mathbf{z}_{-qi}, \boldsymbol{\varepsilon}_i\right) \tag{9.18}$$

if $\mathbb{X}_{\boldsymbol{z}q}$ is a countable discrete set with $\left(z_{qi}, z'_{qi}\right) \in \mathbb{X}^2_{\boldsymbol{z}q}$ being two consecutive values; and:

$$\mathcal{C}_{qpi}\left(z_{qi}, \mathbf{z}_{-qi}, \boldsymbol{\varepsilon}_i\right) = \frac{\partial}{\partial z_{qi}} r_p\left(z_{qi}, \mathbf{z}_{-qi}, \boldsymbol{\varepsilon}_i\right) \tag{9.19}$$

if $\mathbb{X}_{\boldsymbol{z}q}$ is a continuous set, and where $r_p(\cdot)$ is the p-th equation of the reduced form, the one that predicts Y_{pi}. Therefore, the individual causal effect can be interpreted as the "effect" of a *ceteris paribus*, marginal variation of Z_{qi} on the endogenous variable Y_{pi}, obtained by keeping all the other observable exogenous variables as well as the unobserved factors constant.

Example 9.7. Causal Effects in the Mincer Equation. Consider again equation (7.6) from the Mincer model of human capital. There, the *causal effect of education* on the log-wage of any individual i equals the parameter for education in the model: $\mathcal{C}_{SWi}(s_i, \cdot) = \beta_3$. The *causal effect of experience* is, instead:

$$\mathcal{C}_{XWi}(x_i, \cdot) = \beta_1 + 2\beta_2 x_i$$

which is a function of the current experience x_i of the i-th observation. In linear models, in general, the causal effect of variables that enter the model without higher order terms (unlike experience X_i in the Mincer equation) and without "interactions terms" with other variables (for example, dummy variables that allow for group-varying slopes, as in Example 7.9) is equal to their associated structural parameter. ■

By themselves, identification and causality are two unrelated concepts. There can be identifiable models (or, with more proper terminology, models with identified structures) in which causal effects are not defined, like non-separable models. The converse is likewise true: the model in Example (9.3) might not be identified, but the causal effect of X_i on Y_i certainly exists as a theoretical construct. The common confusion between the two terms stems from their frequent mix in the professional parlance of applied economists. In fact, in order to compute causal effects for some econometric model, it is typically necessary to estimate first some of its parameters: which is a task that requires said parameters to be identified in the first place.

When a variable of interest is a so-called binary **treatment** $S_i \in \{0, 1\}$, that takes value $S_i = 1$ if a certain condition is realized for observation i and $S_i = 0$ otherwise (in the education context, this could be, say, an indicator for the achievement of a university degree), causality is often expressed via the so-called **potential outcomes notation**, originating from the famous Rubin (1974) *causal model*, where for some endogenous variable Y_i:

$$Y_i = \begin{cases} Y_i(1) & \text{if } S_i = 1 \\ Y_i(0) & \text{if } S_i = 0 \end{cases} \tag{9.20}$$

which follows from some implicit or explicit model that makes the endogenous variable Y_i dependent on the treatment S_i. For individual i, the causal effect in question is simply given by $\mathcal{C}_{SYi} = Y_i(1) - Y_i(0)$, so long as S_i is effectively an exogenous variable and the error term is mean-independent of it. If this condition cannot be attained, the prevalent empirical practice is to look for ways to approximate it (borrowing on the statistical literature on *causal inference*, which is outside the scope of these lectures) or to search for appropriate *instruments* (see Lecture 10), implicitly treating the model as a larger structural simultaneous equations model.

The concept of individual causal effect is not very useful, since $\boldsymbol{\varepsilon}_i$ is by definition unobserved for a single individual observation. Better results are achieved through the population-wide generalization of the concept.

Definition 9.15. Average Causal Effect. In the population, the average causal effect of varying variable Z_q is the expected value of the individual causal effects *conditional* on the other exogenous variables $\boldsymbol{z}_{-q}$.

$$\begin{aligned}\mathrm{ACE}_{qp}\left(z_{qi}, \mathbf{z}_{-qi}\right) &\equiv \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\mathcal{C}_{qpi}\left(z_{qi}, \mathbf{z}_{-qi}, \boldsymbol{\varepsilon}_i\right) \middle| z_{qi}, \mathbf{z}_{-qi}\right] \\ &= \int_{\mathbb{X}_{\boldsymbol{\varepsilon}}} \mathcal{C}_{qpi}\left(z_{qi}, \mathbf{z}_{-qi}, \boldsymbol{\varepsilon}_i\right) f_{\boldsymbol{\varepsilon}|\boldsymbol{z}}\left(\boldsymbol{\varepsilon}_i \middle| z_{qi}, \mathbf{z}_{-qi}\right) d\varepsilon_{1i} \ldots d\varepsilon_{Pi}\end{aligned}$$

Notice how the expectation above is taken with respect to the unobservables $\boldsymbol{\varepsilon}_i$ (whose support is denoted here by $\mathbb{X}_{\boldsymbol{\varepsilon}}$), and conditional on the exogenous variables $\boldsymbol{z}_i$. For a binary treatment, the average causal effect is called the **Average Treatment Effect** (ATE).

$$\mathrm{ATE}_Y \equiv \mathbb{E}\left[Y_i\left(1\right) - Y_i\left(0\right) \middle| z_{1i}, \ldots, z_{(Q-1)i}\right] \tag{9.21}$$

A related quantity is the **Average Treatment on the Treated** (ATT):

$$\mathrm{ATT}_Y \equiv \mathbb{E}\left[Y_i\left(1\right) - Y_i\left(0\right) \middle| S_i = 1; z_{1i}, \ldots, z_{(Q-1)i}\right] \tag{9.22}$$

which conditions on that part of the populations that does actually receive the treatment (often a more interesting quantity for policy purposes). Note how both expectations condition on the other $Q-1$ exogenous variables.

An intermediate objective of much econometric analysis is the estimation of the Conditional Expectation Function (CEF) of some endogenous variable Y_{pi} conditional on the observable exogenous variables $\boldsymbol{z}_i$. An interesting question is whether the derivative of the CEF with respect to some q-th exogenous variable:

$$\mu^q_{Y_p|\boldsymbol{z}}\left(z_{qi}, \boldsymbol{z}_{-qi}\right) \equiv \frac{\partial}{\partial z_q}\mathbb{E}\left[Y_{pi} \middle| Z_{1i}, Z_{2i}, \ldots, Z_{Qi}\right]\Big|_{Z_{qi}=z_{qi}}$$

with $q = 1, \ldots, Q$, also equals the Average Causal Effect *in the population*. In order to answer this question, consider that in typical cases derivatives for an exogenous variable can pass through integrals taken over $\boldsymbol{\varepsilon}_i$; so:

$$\begin{aligned}\mu^q_{Y_p|\boldsymbol{z}}\left(z_{qi}, \boldsymbol{z}_{-qi}\right) &= \frac{\partial}{\partial z_{qi}}\int_{\mathbb{X}_{\boldsymbol{\varepsilon}}} r_p\left(z_{qi}, \mathbf{z}_{-qi}, \boldsymbol{\varepsilon}_i\right) f_{\boldsymbol{\varepsilon}|\boldsymbol{z}}\left(\boldsymbol{\varepsilon}_i \middle| z_{qi}, \mathbf{z}_{-qi}\right) d\varepsilon_{1i} \ldots d\varepsilon_{Pi} \\ &= \int_{\mathbb{X}_{\boldsymbol{\varepsilon}}} r_p\left(z_{qi}, \mathbf{z}_{-qi}, \boldsymbol{\varepsilon}_i\right)\left[\frac{\partial}{\partial z_{qi}} f_{\boldsymbol{\varepsilon}|\boldsymbol{z}}\left(\boldsymbol{\varepsilon}_i \middle| z_{qi}, \mathbf{z}_{-qi}\right)\right] d\varepsilon_{1i} \ldots d\varepsilon_{Pi} + \\ &\quad + \underbrace{\int_{\mathbb{X}_{\boldsymbol{\varepsilon}}}\left[\frac{\partial}{\partial z_{qi}} r_p\left(z_{qi}, \mathbf{z}_{-qi}, \boldsymbol{\varepsilon}_i\right)\right] f_{\boldsymbol{\varepsilon}|\boldsymbol{z}}\left(\boldsymbol{\varepsilon}_i \middle| z_{qi}, \mathbf{z}_{-qi}\right) d\varepsilon_{1i} \ldots d\varepsilon_{Pi}}_{=\mathrm{ACE}_{qp}(z_{qi}, \mathbf{z}_{-qi})}\end{aligned}$$

and the answer is a conditional ***no***; $\mu^{q}_{Y_p|\boldsymbol{z}}(z_{qi}, \boldsymbol{z}_{-qi}) = \text{ACE}_{qp}(z_{qi}, \mathbf{z}_{-qi})$ only if:

$$\int_{\mathbb{X}_{\boldsymbol{\varepsilon}}} r_p(z_{qi}, \mathbf{z}_{-qi}, \boldsymbol{\varepsilon}_i) \left[\frac{\partial}{\partial z_{qi}} f_{\boldsymbol{\varepsilon}|\boldsymbol{z}}(\boldsymbol{\varepsilon}_i | z_{qi}, \mathbf{z}_{-qi}) \right] d\varepsilon_{1i} \dots d\varepsilon_{Pi} = 0$$

that is, if Z_{qi} and the unobservables $\boldsymbol{\varepsilon}_i$ are independent, conditional on the other exogenous variables $\boldsymbol{z}_{-i}$. This has a proper definition in statistics.

Definition 9.16. Conditional Independence Assumption (CIA)**.** The CIA is the hypothesis that the the unobservables $\boldsymbol{\varepsilon}_i$ and a specific exogenous variable Z_{qi} are statistically independent, conditional on all the other exogenous variables $\boldsymbol{z}_{-qi}$.

$$Z_{qi} \perp \boldsymbol{\varepsilon}_i | \boldsymbol{z}_{-qi} \tag{9.23}$$

For binary treatments, the CIA is often expressed in potential outcomes notation as follows.

$$Y_i(1) - Y_i(0) \perp S_i | Z_{1i}, \dots, Z_{(Q-1)i} \tag{9.24}$$

The importance of this concept lies in the fact that it provides a clear condition to verify if the parameters of an econometric model can be interpreted causally. Suppose, for example, that the CEF of interest is linear:

$$\mathbb{E}[Y_{pi} | Z_{1i}, Z_{2i}, \dots, Z_{Qi}] = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \dots \beta_Q Z_{Qi}$$

if the CIA holds for an exogenous variable of interest Z_{qi} with $q = 1, \dots, Q$, it follows $\beta_q = \text{ACE}_{qp}(z_{qi}, \mathbf{z}_{-qi})$. A practical strategy which is instrumental to achieve such a result, is that of "enriching" a model with many exogenous variables so that conditional independence becomes a more credible hypothesis. One final observation is in order: while the CIA is weaker than full statistical independence, it strongly resembles the central condition for identification of linear models, that is mean independence of the error term (e.g. $\mathbb{E}[\varepsilon_i | X_i] = 0$). While the latter is definitely not identical to the CIA, this is definitely another cause for confusion between the two concepts of identification and causality. This mistake is to be avoided!

In a specific circumstance average causal effects are well approximated, if not exactly estimated, by a linear model. This is the case if the outcome of interest is Y_i, the exogenous variable of interest is **binary** (write it S_i), and the other exogenous variables $\boldsymbol{x}_i$ satisfy $\mathbb{E}[S_i | \boldsymbol{x}_i] = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\pi}_0$: linearity of the CEF (this holds trivially in some cases – for example if $\boldsymbol{x}_i$ represents a full dummy variable group partition, like in fully saturated regressions). Then, by estimating the following linear model:

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \delta_0 s_i + \varepsilon_i$$

the interpretation of the optimal linear predictor for δ_0 can be based on the Yitzakhi-Angrist-Krueger decomposition (7.56), where the derivative of the CEF conditional on $\boldsymbol{x}_i$ is precisely the ATE:

$$\mu'_{Y|S,\boldsymbol{x}} = \mathbb{E}\left[Y_i \middle| S_i = 1, \boldsymbol{x}_i\right] - \mathbb{E}\left[Y_i \middle| S_i = 0, \boldsymbol{x}_i\right] \equiv \Delta\left(\boldsymbol{x}_i\right) \tag{9.25}$$

while the weighting factor here is $\phi\left(\boldsymbol{x}_i\right) = \mathbb{P}\left(S_i = 1 \middle| \boldsymbol{x}_i\right)\left[1 - \mathbb{P}\left(S_i = 1 \middle| \boldsymbol{x}_i\right)\right]$. In this case, (7.56) becomes as follows.

$$\delta^* = \frac{\mathbb{E}_{\boldsymbol{x}}\left[\Delta\left(\boldsymbol{x}_i\right)\mathbb{P}\left(S_i = 1 \middle| \boldsymbol{x}_i\right)\left[1 - \mathbb{P}\left(S_i = 1 \middle| \boldsymbol{x}_i\right)\right]\right]}{\mathbb{E}_{\boldsymbol{x}}\left[\mathbb{P}\left(S_i = 1 \middle| \boldsymbol{x}_i\right)\left[1 - \mathbb{P}\left(S_i = 1 \middle| \boldsymbol{x}_i\right)\right]\right]} \tag{9.26}$$

By inspecting the expression above, it appears that the linear projection δ^* identifies the average causal effect of S_i on Y_i in one of two cases:

- the causal effect does not vary with $\boldsymbol{x}_i$ ($\Delta\left(\boldsymbol{x}_i\right)$ is a constant);
- the probability to "take up the treatment" ($S_i = 1$) is constant for $\boldsymbol{x}_i$.

While these conditions are unlikely to hold in practice, δ^* still carries an interpretation as an average causal effect that is weighted by $\phi\left(\boldsymbol{x}_i\right)$; as in the general case, in a practical setting it is important to learn about these weights in order to inform the interpretation of one's estimates.

Finally, it is important to observe that while causality is best framed in terms of effects of exogenous variables on endogenous ones in the setting of reduced form models, a selected class of structural models – the so-called **triangular models** – allows for "endogenous-to-endogenous" causal effects.

Definition 9.17. Triangular Models. A *triangular* structural model is one where its P equations and its P endogenous variables can be ordered in such a way that, for any natural number $P' < P$, the first P' endogenous variables never enter the last $P - P'$ equations, or vice versa.

Possibly, the simplest triangular model is the following trivariate model:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2 Z_i + \varepsilon_i \\ X_i &= \pi_0 + \pi_1 Z_i + \eta_i \end{aligned} \tag{9.27}$$

where (Y_i, X_i) are the endogenous variables while Z_i is the only exogenous one: notice that Y_i does not enter the second equation. In general, a SEM is triangular if matrix $\boldsymbol{\Gamma}$ is either upper- or lower-triangular (hence the name). A Mincer model enriched with a linear equation for education where wages are absent, such as (7.7), is triangular too. In the case of triangular models, it is sensible to talk about the **causal effects** of endogenous variables upon other endogenous variables (X_i on Y_i, education on wages, *et cetera*); all definitions and considerations made above apply.

Lecture 10

Instrumental Variables

This lecture is chiefly devoted to the most defining element of econometrics: the method of Instrumental Variables, which is intended for addressing the most typical problem of empirical economic studies: endogeneity. Following an introduction to the different types of endogeneity, Instrumental Variables and the related ideas are discussed through the presentation of increasingly more general estimators for linear models: IV regression, Two Stages Least Squares, and finally Three Stages Least Squares for simultaneous equations.

10.1 Endogeneity Problems

In the previous lectures, a single concept has been stated repeatedly and in different forms: that if the "exogeneity" condition $\mathbb{E}[\varepsilon_i | \boldsymbol{x}_i] = 0$ (White's Assumption 2) fails, the OLS estimator is inconsistent. As already argued, this is the main condition against which the quality of an econometric analysis is evaluated (beyond the relevance of the research question, of course) and for a very good reason: this condition is **likely to fail** in most observational data that are not generated via quasi-experiment or actual randomized experiments. The circumstance where exogeneity fails:

$$\mathbb{E}[\varepsilon_i | \boldsymbol{x}_i] \neq 0 \tag{10.1}$$

is called **endogeneity** or **failure of identification**. Both expressions are largely conventional; as hinted, the former comes from the theory of simultaneous equation models while the latter is due to the intimate relationship between the exogeneity condition and identification in linear models (see example 9.3 and the subsequent discussion), although technically (10.1) is really an issue of conditional moments. Another implication of endogeneity is that it implies a failure of the Conditional Independence Assumption too, and hence the impossibility to discuss about **causal effects** of $\boldsymbol{x}_i$ on Y_i.

It is worthwhile to provide a **taxonomy of endogeneity problems**, so to learn to easily recognize these issues in empirical studies. What follows is a separate discussion of the four scenarios that comprise a taxonomy of endogeneity: 1. the **omitted variable bias** (including a discussion of **fixed effects**); 2. **simultaneity**; 3. **measurement error** and 4. other forms of **structural endogeneity**.

Omitted Variable Bias and Fixed Effects

The omitted variable bias is described earlier in Lecture 7, and it is certainly the most common cause for concern about research papers in economics: many of these only feature one structural equation, which is often linear. It is worthwhile to summarize the problem again: suppose that the "true" CEF is $\mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i, S_i\right] = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + \delta_0 S_i$ where S_i is some relevant variable **omitted** from the empirical model; then, the probability limit of the OLS estimator is:

$$\widehat{\boldsymbol{\beta}}_{OLS} \xrightarrow{p} \boldsymbol{\beta}_0 + \delta_0 \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right]^{-1} \mathbb{E}\left[\boldsymbol{x}_i S_i\right]$$

which is affected by a **bias term** that can be decomposed as the product between: *i.* δ_0, that is the coefficient of S_i in the "true" CEF of Y_i, which represents the "effect" of the omitted variable on the dependent variable; and *ii.* the population linear projection of S_i on $\boldsymbol{x}_i$ – that is $\mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right]^{-1} \mathbb{E}\left[\boldsymbol{x}_i S_i\right]$ – which relates to the population correlation between the omitted variable and the explanatory variables.[1] Recall that if there is good reason to believe that either term is zero, omitting S_i is not a problem at all! Lecture 7 already provides a generalization of this concept to multiple omitted variables, as well as an illustrative interpretation in terms of the Mincer equation.

Fixed Effects. In applied economic research, a specific form of omitted variable bias is routinely addressed with **panel data**. Suppose that the true model is:

$$y_{it} = \alpha_i + \mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}_0 + \varepsilon_{it} \tag{10.2}$$

where $\mathbb{E}\left[\varepsilon_{it} \middle| \boldsymbol{x}_{it}\right] = 0$, while α_i is an additional unobserved error, typically called the **individual fixed effect**, which is **constant in time** and possibly "endogenous" in the following sense.

$$\mathbb{E}\left[\alpha_i \middle| \boldsymbol{x}_{it}\right] \neq 0 \tag{10.3}$$

Individual fixed effects represent the **"unobserved heterogeneity"** of the data, that is factors that are constant in time, unobserved but vary across units of observations (e.g. the ability of workers, the "know-how" of firms).

[1] If, say, $\mathbb{E}\left[S_i \middle| \boldsymbol{x}_i\right] = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma}_0$, then clearly $\mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right]^{-1} \mathbb{E}\left[\boldsymbol{x}_i S_i\right] = \boldsymbol{\gamma}_0$.

With panel data, fixed effects can be easily addressed: even if T is small and thus the individual effects α_i cannot be consistently estimated by brute force (e.g. as separate dummy variables) we know from the Frisch-Waugh-Lovell theorem that an estimate of $\boldsymbol{\beta}_0$ based on a "demeaned" model:

$$y_{it} - \bar{y}_i = \left(\mathbf{x}_{it} - \bar{\mathbf{x}}_i\right)^{\mathrm{T}} \boldsymbol{\beta}_0 + \left(\varepsilon_{it} - \bar{\varepsilon}_i\right) \tag{10.4}$$

where $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, $\bar{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{it}$, and $\bar{\varepsilon}_i = \frac{1}{T}\sum_{t=1}^{T} \varepsilon_{it}$, is numerically equivalent to the one from a brute force estimate that includes unit-specific dummies; moreover, such an estimate would be consistent since α_i is absent from (10.4). An alternative, which is generally asymptotically equivalent, is to estimate a model in "first differences:"

$$\Delta y_{it} = \Delta \mathbf{x}_{it}^{\mathrm{T}} \boldsymbol{\beta}_0 + \Delta \varepsilon_{it} \tag{10.5}$$

where Δ is the first-differences operator ($\Delta a_{it} = a_{it} - a_{i(t-1)}$, and similarly for vectors: $\Delta \mathbf{a}_{it} = \mathbf{a}_{it} - \mathbf{a}_{i(t-1)}$).

In panel data, (10.4) is called the **within transformation** of (10.2), while (10.5) is called the **between transformation**. The two approaches result in the removal of fixed effects; however, it is typically observed that the resulting models are "identified from the time-series variation in the explanatory variables." What this means is that the information on $\boldsymbol{\beta}_0$ is now inferred from how variations of $\boldsymbol{x}_{it}$ in time relate to variations of Y_{it} in time. This naturally shrinks the set of applications to the analysis of explanatory variables that show variability over time. For example, this rules out panel data as a solution to the omitted-ability bias of the Mincer equation, since the main variable of interest – education – is typically constant in panels of workers, and would thus disappear from both (10.4) and (10.5).

Simultaneity

The problem of simultaneity is perhaps the archetypical type of endogeneity problem, and it derives from the classical analysis of linear simultaneous equations in the early days of econometrics. While already elaborated in Lecture 9, this problem is revisited here from a more statistical angle. Consider a Simultaneous Equations Model (SEM) written in compact form (9.2): $\boldsymbol{\Gamma}\mathbf{y}_i = \boldsymbol{\Phi}\mathbf{z}_i + \boldsymbol{\varepsilon}_i$. Suppose that the exogenous variables $\boldsymbol{z}_i$ are defensibly exogenous, that is conditionally mean independent of the P error terms $\boldsymbol{\varepsilon}_i$ of the model. Write:

$$\mathbb{E}\left[\boldsymbol{\varepsilon}_i \middle| \boldsymbol{z}_i\right] = \mathbf{0} \tag{10.6}$$

implying $\mathbb{E}\left[\boldsymbol{z}_i \varepsilon_{pi}\right] = \mathbf{0}$ for $p = 1, \ldots, P$ by (10.6) and by the Law of Iterated Expectations; this can be written more compactly as $\mathbb{E}\left[\boldsymbol{z}_i \boldsymbol{\varepsilon}_i^{\mathrm{T}}\right] = \mathbf{0}$.

Unfortunately, **by construction** the same cannot be argued for the P endogenous variables; by exploiting the reduced form (9.8), also expressed as $\boldsymbol{y}_i = \boldsymbol{\Pi}\boldsymbol{z}_i + \boldsymbol{\Gamma}^{-1}\boldsymbol{\varepsilon}_i$, it is easy to show how identification fails.

$$\mathbb{E}\left[\boldsymbol{y}_i\boldsymbol{\varepsilon}_i^{\mathrm{T}}\right] = \boldsymbol{\Pi}\cdot\underbrace{\mathbb{E}\left[\boldsymbol{z}_i\boldsymbol{\varepsilon}_i^{\mathrm{T}}\right]}_{=\mathbf{0}} + \boldsymbol{\Gamma}^{-1}\cdot\mathbb{E}\left[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^{\mathrm{T}}\right] = \boldsymbol{\Gamma}^{-1}\cdot\mathbb{V}\mathrm{ar}\left[\boldsymbol{\varepsilon}_i\right] \neq \mathbf{0} \tag{10.7}$$

The intuition is represented in Figure 10.1 below, which displays the graph of structural relationships implied by two equations of a SEM, where any two endogenous variables Y_{1i} and Y_{2i} show up in both equations. Since both endogenous variables are also affected by the respective error terms ε_{1i} and ε_{2i}, the latter are by construction correlated with *both* Y_{1i} and Y_{2i}, directly or indirectly; this is an abstract representation of the identification problem illustrated in Examples 9.5 and 9.6. Here the denomination **simultaneity** is predicated on the concomitance of all these structural relationships.

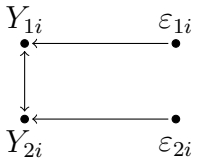


Figure 10.1: Graphical representation of Simultaneity

The analysis conducted in the previous Lecture elaborates upon conditions for the identification of a SEM. The final part of this Lecture completes the discussion of SEMs by discussing methods for their estimation.

Measurement Error

It is quite common that some variables contained in a dataset are, to some degree, **measured with error**. This problem typically leads to inconsistent estimates whenever it affects the model's explanatory (exogenous) variables. To illustrate, consider a bivariate regression model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ with "exogenous" X_i ($\mathbb{E}\left[\varepsilon_i \middle| X_i\right] = 0$). However, *the researcher cannot observe the true variable* X_i, but only *its error-ridden version*:

$$X_i^* = X_i + \upsilon_i \tag{10.8}$$

where υ_i denotes the *error in the measurement* of X_i (with $\mathbb{E}\left[\upsilon_i\right] = 0$). In addition, suppose that the error υ_i is also **completely random**, that it is independent from both the "true" X_i as well as of the original error term ε_i.

$$\mathbb{E}\left[X_i\upsilon_i\right] = \mathbb{E}\left[\varepsilon_i\upsilon_i\right] = 0$$

This circumstance is known as **classical measurement error**. For a more agile notation, write $\sigma_x^2 \equiv \mathbb{V}\mathrm{ar}[X_i]$ and $\sigma_\upsilon^2 \equiv \mathbb{V}\mathrm{ar}[\upsilon_i]$.

Given actual data $\{(y_i, x_i^*)\}_{i=1}^N$ (with $\bar{x}^* = \frac{1}{N}\sum_{i=1}^N x_i^*$), the OLS estimator of the regression slope is inconsistent because:

$$\widehat{\beta}_{1,OLS} = \frac{\sum_{i=1}^N (x_i^* - \bar{x}^*) y_i}{\sum_{i=1}^N (x_i^* - \bar{x}^*)^2} \xrightarrow{p} \frac{\mathbb{C}\mathrm{ov}[X_i + \upsilon_i, \beta_0 + \beta_1 X_i + \varepsilon]}{\mathbb{V}\mathrm{ar}[X_i + \upsilon_i]}$$
$$= \frac{\beta_1 \mathbb{V}\mathrm{ar}[X_i]}{\mathbb{V}\mathrm{ar}[X_i] + \mathbb{V}\mathrm{ar}[\upsilon_i]} = \beta_1 \left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_\upsilon^2} \right)$$

its probability limit is actually *smaller* than β_1, to an extent that depends on the following multiplicative constant.

$$\frac{\sigma_x^2}{\sigma_x^2 + \sigma_\upsilon^2} = \frac{1}{1 + \sigma_\upsilon^2 \sigma_x^{-2}} = 1 - \frac{\sigma_\upsilon^2 \sigma_x^{-2}}{1 + \sigma_\upsilon^2 \sigma_x^{-2}} \in (0, 1)$$

This is the infamous **attenuation bias** of classical measurement error; the intuition for this result is that, even if υ_i is completely random, the relative importance of the covariance between Y_i and X_i^* is obfuscated by the error-inflated variance of X_i^*. The magnitude of the problem depends on the one of the term $\sigma_\upsilon^2 \sigma_x^{-2}$, which is called **noise-to-signal ratio**.

Classical measurement error is an instance of failure of the exogeneity assumption as per (10.1). To see this, consider the model:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \varepsilon_i \\ &= \beta_0 + \beta_1 X_i^* + \varepsilon_i - \beta_1 \upsilon_i \end{aligned}$$

where the *actual regressor being used in practice* by the analyst is X_i^* and the actual error term is therefore $\varepsilon_i - \beta_1 \upsilon_i$. Clearly, by construction it is:

$$\mathbb{E}[\varepsilon_i - \beta_1 \upsilon_i | X_i^*] \neq 0$$

because X_i^* incorporates υ_i. No similar problem affects the dependent variable Y_i: if the researcher really observes the latter with error: $Y_i^* = Y_i + \upsilon_i$, then the model:

$$Y_i^* = \beta_0 + \beta_1 X_i + \varepsilon_i + \upsilon_i$$

is still estimated consistently under the previous assumptions about υ_i being "completely random" (although certainly the additional source of noise would make the estimates overall less precise).

These facts are easily generalized to the multivariate model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$. Each explanatory variable X_{ki} may or may not be affected by measurement error, in the sense that the actually observed variables are:

$$X_{ki}^* = X_{ki} + \upsilon_{ki} \tag{10.9}$$

for $k = 1, \ldots, K$ and $i = 1, \ldots, N$. In compact matrix notation, the actual realizations of the explanatory variables as well as the measurement errors are collected by two $N \times K$ matrices:

$$\mathbf{X}^* = \begin{bmatrix} x_{11}^* & x_{21}^* & \ldots & x_{K1}^* \\ x_{12}^* & x_{22}^* & \ldots & x_{K2}^* \\ \vdots & \vdots & \ddots & \vdots \\ x_{1N}^* & x_{2N}^* & \ldots & x_{KN}^* \end{bmatrix}; \; \mathbf{U}^* = \begin{bmatrix} \upsilon_{11}^* & \upsilon_{21}^* & \ldots & \upsilon_{K1}^* \\ \upsilon_{12}^* & \upsilon_{22}^* & \ldots & \upsilon_{K2}^* \\ \vdots & \vdots & \ddots & \vdots \\ \upsilon_{1N}^* & \upsilon_{2N}^* & \ldots & \upsilon_{KN}^* \end{bmatrix}$$

implying the following "true estimated model."

$$\mathbf{y} = \mathbf{X}^* \boldsymbol{\beta}_0 + \boldsymbol{\varepsilon} - \mathbf{U}\boldsymbol{\beta}_0 \tag{10.10}$$

The "actual" error term is now $\varepsilon_i - \sum_{k=1}^{K} \beta_{k,0} \upsilon_{ki}$; again by construction, it is:

$$\mathbb{E}\left[\varepsilon_i - \sum_{k=1}^{K} \beta_{k,0} \upsilon_{ki} \middle| \boldsymbol{x}_i^* = \mathbf{x}_i^* \right] \neq 0 \tag{10.11}$$

where $\mathbf{x}_i^*$ is the i-th row of $\mathbf{X}^*$, even under the maintained hypothesis that the errors υ_{ki} are completely independent of both the "true" explanatory variables and the primitive error term:

$$\mathbb{E}\left[X_{k'i} \upsilon_{ki} \right] = \mathbb{E}\left[\varepsilon_{ki} \upsilon_i \right] = 0 \tag{10.12}$$

for $k, k' = 1, \ldots, K$. For simplicity, define the following probability limits:

$$\begin{aligned} \frac{1}{N} \mathbf{X}^{\mathrm{T}} \mathbf{X} &\xrightarrow{p} \boldsymbol{\Sigma}_{\boldsymbol{x}} \\ \frac{1}{N} \mathbf{U}^{\mathrm{T}} \mathbf{U} &\xrightarrow{p} \boldsymbol{\Sigma}_{\boldsymbol{\upsilon}} \end{aligned} \tag{10.13}$$

and note that the "actual" OLS estimator can be written as follows.

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{OLS} &= \left(\mathbf{X}^{*\mathrm{T}} \mathbf{X}^* \right)^{-1} \mathbf{X}^{*\mathrm{T}} \mathbf{y} \\ &= \boldsymbol{\beta}_0 + \left[\frac{1}{N} \left(\mathbf{X} + \mathbf{U} \right)^{\mathrm{T}} \left(\mathbf{X} + \mathbf{U} \right) \right]^{-1} \frac{1}{N} \left(\mathbf{X} + \mathbf{U} \right)^{\mathrm{T}} \left(\boldsymbol{\varepsilon} - \mathbf{U}\boldsymbol{\beta}_0 \right) \end{aligned} \tag{10.14}$$

Since, under the maintained assumptions, it is $\frac{1}{N} \mathbf{U}^{\mathrm{T}} \boldsymbol{\varepsilon} \xrightarrow{p} \mathbf{0}$; and similarly the probability limit of $\frac{1}{N} \mathbf{X}^{\mathrm{T}} \mathbf{U}$ is a matrix full of zeros, the probability limit of the OLS estimator can be written as:

$$\widehat{\boldsymbol{\beta}}_{OLS} \xrightarrow{p} \left[\mathbf{I} - \left(\boldsymbol{\Sigma}_{\boldsymbol{x}} + \boldsymbol{\Sigma}_{\boldsymbol{\upsilon}} \right)^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\upsilon}} \right] \boldsymbol{\beta}_0 \leq \boldsymbol{\beta}_0 \tag{10.15}$$

which is a generalization of the earlier formula about the attenuation bias to the multivariate case.

Structural Endogeneity

The final subcategory of endogeneity is somewhat residual, and it collects those kinds of endogeneity problems that are somewhat "built-in" the very structural model. Consider, for example, the following time-series model:

$$\begin{aligned} y_t &= \beta_0 + \beta_1 y_{t-1} + \mathbf{x}_t^{\mathrm{T}}\boldsymbol{\gamma}_0 + \mathbf{x}_{t-1}^{\mathrm{T}}\boldsymbol{\gamma}_1 + \varepsilon_t \\ \varepsilon_t &= \rho\varepsilon_{t-1} + \xi_t \end{aligned} \tag{10.16}$$

where the current realizations of the dependent variable y_t depend on its past, as well as on current and past realizations of some explanatory variables $\mathbf{x}_t$; in addition, the error term presents an AR(1) structure with autoregressive parameter $\rho \in (0, 1)$ and "innovation" shock ξ_t – not autocorrelated; this is a more specialized version of model (8.44). It is quite obvious that:

$$\mathbb{E}\left[\varepsilon_t \middle| Y_{t-1}\right] = \mathbb{E}\left[\rho\varepsilon_{t-1} + \xi_t \middle| Y_{t-1}\right] = \rho\,\mathbb{E}\left[\varepsilon_{t-1} \middle| Y_{t-1}\right] + \mathbb{E}\left[\xi_t \middle| Y_{t-1}\right] \neq 0$$

even if ξ_t is conditionally mean-independent, the grand error term ε_t is not, as its lag ε_{t-1} affects the lag of the dependent variable Y_t *by construction.*

Another not too dissimilar case is a **spatial model** written in compact matrix notation as:

$$\mathbf{y} = \beta_0\boldsymbol{\iota} + \beta_1\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\gamma}_0 + \mathbf{W}\mathbf{X}\boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon} \tag{10.17}$$

where $\mathbf{W}$ is a $N \times N$ non-stochastic **spatial weighting matrix** with zero diagonal:

$$\mathbf{W} = \begin{bmatrix} 0 & w_{12} & \dots & w_{1N} \\ w_{21} & 0 & \dots & w_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & \dots & 0 \end{bmatrix}$$

which collects the w_{ij} **distances** between two distinct units i and j in the sample. A model of this sort is common in urban and regional econometrics, as well as in the econometrics of networks and social interactions. In general, it can be rewritten in terms of its solution for $\mathbf{y}$ as:

$$\mathbf{y} = \left(\mathbf{I} - \beta_1\mathbf{W}\right)^{-1}\left(\beta_0\boldsymbol{\iota} + \mathbf{X}\boldsymbol{\gamma}_0 + \mathbf{W}\mathbf{X}\boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon}\right)$$

which suggests the existence of an endogeneity problem due to the feedback mechanisms, that are built in the model, between the dependent variables (and the error terms) of different economic units.

$$\mathbb{E}\left[\varepsilon_i \middle| Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_N\right] \neq 0$$

A careful reader will easily note an analogy between the endogeneity problem of spatial models and the issue of simultaneity in SEMs!

10.2 Instrumental Variables in Theory

Instrumental variables are the chief solution that econometric studies adopt to address endogeneity problems. Some methods employed in econometrics and causal statistics (like the different "discontinuity" designs) can be seen as specific applications of instrumental variables. As it is detailed in later lectures, Instrumental Variables (IVs) or more simply **instruments** can be generalized to multi-equation non-linear models; however, it is more useful to start characterizing them in the setting of single-equation linear regression models such as (7.1). In that context, IVs are **exogenous variables** Z_i such that the regression error is conditionally mean-independent of them:

$$\mathbb{E}\left[\varepsilon_i \middle| Z_i\right] = 0 \tag{10.18}$$

while at the same time, IVs Z_i do not show up in the assumed model that "explains" the endogenous variable of interest Y_i. This last statement needs some more qualification. It is useful for illustrative purposes to think of IVs as part of an augmented structural model which features relationships like the ones exemplified by the following graph:

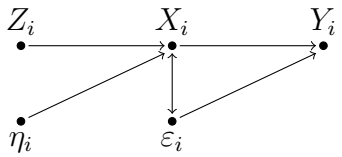


Figure 10.2: Graphical representation of IVs

where X_i is the main explanatory variable that a researcher is interested in, and that is possibly endogenous (as indicated by the bidirectional arrow that connects it with the error term ε_i), and which is itself "explained" by both the IV Z_i as well as some other unobserved factor η_i. Observe that Z_i is itself unrelated to both error terms (ε_i, η_i) and – importantly – it does not itself "explain" Y_i, at least not directly: any "effect" that Z_i might have occurs through the X_i channel, which is indirectly represented through the lack of a direct arrow starting from Z_i and terminating in Y_i. In addition, even the error term for X_i, that is η_i, is conditionally mean independent of the instrument Z_i, in the following sense.

$$\mathbb{E}\left[\eta_i \middle| Z_i\right] = 0 \tag{10.19}$$

A model with such features would clearly be a **triangular** structural model as defined at the end of Lecture 9, and the plausibility of such a scenario would depend on the socio-economic context of interest.

Trivariate triangular models

In order to explain how IVs can help address endogeneity problems, it is useful to start from the simplest triangular model, to move gradually towards the analysis of the more general **Two-stages Least Squares estimator**. Consider the very simple "trivariate" triangular model (9.27), by imposing the additional **restriction** that the exogenous variable (here, the IV) does not "explain" Y_i in the structural model, as per Figure 10.2.

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \varepsilon_i \\ X_i &= \pi_0 + \pi_1 Z_i + \eta_i \end{aligned} \tag{10.20}$$

In econometric terminology, the first of these two equations is named the **structural form** (implicitly, the structural form for Y_i – that is the main relationship of interest) while the second equation is called the **first stage** for X_i. Let us examine the properties of this model under the assumption that X_i is endogenous, that is $\mathbb{E}[\varepsilon_i | X_i] \neq 0$.

Since the IV is exogenous, it is $\mathbb{E}[\varepsilon_i, \eta_i | Z_i] = 0$, hence (π_0, π_1) are consistently estimated via OLS performed on the first stage equation. It turns out that, in this model, even parameters (β_0, β_1) are identified! To see this, consider the reduced form of the model:

$$\begin{aligned} Y_i &= \beta_0 + \pi_0\beta_1 + \beta_1\pi_1 Z_i + \varepsilon_i + \beta_1\eta_i \\ X_i &= \pi_0 + \pi_1 Z_i + \eta_i \end{aligned} \tag{10.21}$$

the first equation too can be consistently estimated via OLS! Moreover, there clearly is a unique mapping from the reduced form to the structural parameters, making the model exactly identified. In particular, a consistent estimate of β_1 is obtained – for a given sample $\{(y_i, x_i, z_i)\}_{i=1}^N$ – as the ratio between the OLS estimates of the two coefficients for Z_i from the reduced form:

$$\widehat{\beta}_{1,IV} = \frac{\widehat{\beta_1\pi_1}_{OLS}}{\widehat{\pi}_{1OLS}} = \frac{\frac{\sum_{i=1}(z_i-\bar{z})(y_i-\bar{y})}{\sum_{i=1}(z_i-\bar{z})^2}}{\frac{\sum_{i=1}(z_i-\bar{z})(x_i-\bar{x})}{\sum_{i=1}(z_i-\bar{z})^2}} = \frac{\sum_{i=1}(z_i-\bar{z})(y_i-\bar{y})}{\sum_{i=1}(z_i-\bar{z})(x_i-\bar{x})} \tag{10.22}$$

where $\bar{z} = \frac{1}{N}\sum_{i=1}^N z_i$ while $\bar{y}$ and $\bar{x}$ are analogous, as usual. The expression above implicitly defines the **IV estimator** for β_1 in this simple trivariate model. Instead, the intercept β_0 of the structural form is estimated by IV as:

$$\widehat{\beta}_{0,IV} = \bar{y} - \widehat{\pi}_{0,OLS}\widehat{\beta}_{1,IV} - \widehat{\pi}_{1,OLS}\widehat{\beta}_{1,IV} \cdot \bar{z} \tag{10.23}$$

that is, its consistent IV estimator is obtained by plugging the appropriate estimates for (π_0, π_1) and β_1 in the "average" reduced form equation for Y_i.

That the IV estimators (β_0, β_1) are consistent should appear self-evident by the properties of probability limits upon acknowledging that the reduced form estimators are consistent too; it is useful to also show it as follows:

$$\begin{aligned}\widehat{\beta}_{1,IV} = \frac{\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(y_i - \bar{y})}{\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(x_i - \bar{x})} &\xrightarrow{p} \frac{\mathbb{C}\text{ov}[Z_i, Y_i]}{\mathbb{C}\text{ov}[Z_i, X_i]} \\ &= \beta_1 + \frac{\mathbb{C}\text{ov}[Z_i, \varepsilon_i]}{\mathbb{C}\text{ov}[Z_i, X_i]} = \beta_1\end{aligned} \tag{10.24}$$

where $\mathbb{C}\text{ov}[Z_i, \varepsilon_i] = \mathbb{E}[Z_i\varepsilon_i] = 0$ follows by (10.18) and the Law of Iterated Expectations;[2] the consistency of $\widehat{\beta}_{0,IV}$ follows easily. The decomposition in the second line of (10.24) and an analysis akin to the one from Example 6.7 imply that, if the *observations are independently, not identically distributed* (i.n.i.d.), it follows that:

$$\sqrt{N}\left(\widehat{\beta}_{1,IV} - \beta_1\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left[\varepsilon_i^2 (Z_i - \mathbb{E}[Z_i])^2\right] \cdot \mathbb{C}\text{ov}[Z_i, X_i]^{-2}\right) \tag{10.25}$$

while, if the conditional variance of the error term ε_i – given the IV Z_i – is independent of the latter (*homoscedasticity*), the above simplifies as:

$$\sqrt{N}\left(\widehat{\beta}_{1,IV} - \beta_1\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_0^2 \cdot \mathbb{V}\text{ar}[Z_i]^2 \cdot \mathbb{C}\text{ov}[Z_i, X_i]^{-2}\right) \tag{10.26}$$

where $\sigma_0^2 = \mathbb{E}[\varepsilon_i^2]$. The asymptotic distributions as well as the estimators of their variances are obtained accordingly. Expressions (10.25) and (10.26), however, both show an important point: the asymptotic variance of the IV estimator of the structural form's slope is **inversely proportional to the covariance between the instrument Z_i and the endogenous variable X_i**. This is a crucial aspect that is discussed later at length.

Note that an analogous results would not obtain if the IV Z_i showed up as an explanatory variable for Y_i in the structural form. This is shown via the reduced form of the *unrestricted* triangular model (9.27), that is:

$$\begin{aligned}Y_i &= \beta_0 + \pi_0\beta_1 + (\beta_1\pi_1 + \beta_2) Z_i + \varepsilon_i + \beta_1\eta_i \\ X_i &= \pi_0 + \pi_1 Z_i + \eta_i\end{aligned} \tag{10.27}$$

which is easily shown not to be identified, due to the additional parameter β_2 (the one associated with Z_i in the unrestricted structural equation for Y_i). An analogous failure of identification would follow if the model were not triangular, and Y_i showed up on the right-hand side of the second structural equation. These observations help illustrate the intuition about the

[2]The second-to-last equality follows from $\mathbb{C}\text{ov}[Z_i, Y_i] = \beta_1\mathbb{C}\text{ov}[Z_i, X_i] + \mathbb{C}\text{ov}[Z_i, \varepsilon_i]$.

identification and the estimation of β_1: the "effect" of X_i on Y_i is backed up by the component of the variation in X_i that is predicted by the exogenous instrument Z_i. Obviously, this would be impossible to disentangle from any "direct" structural effect of Z_i on Y_i, if this were not assumed to be zero; or if an additional "effect" of Y_i on X_i were to be accounted for.

It is worth to summarize the four conditions that Instrumental Variables must conform to in order to be adequate and effective in practical contexts. These are expressed as follows in the context of the simple trivariate triangular model, but they extend easily to higher-dimensional environments.

1. **Exogeneity**: conditions (10.18)-(10.19) hold, i.e. $\mathbb{E}[\varepsilon_i, \eta_i | Z_i] = 0$.
2. **Exclusion Restriction**: the instrument Z_i does not affect the main endogenous variable Y_i of the structural form, that is $\beta_2 = 0$ in (9.27).
3. **No Reverse Causality**: the structural relationship between the two endogenous variables is unidirectional: Y_i does not affect X_i directly, that is, the system is indeed triangular.
4. **Relevance**: the covariance $\mathbb{C}\text{ov}[Z_i, X_i]$ between the endogenous explanatory variable X_i and the IV Z_i must be "sufficiently strong," or else the IV estimates would be so imprecise to be useless (a problem commonly referred to as the one of **weak instruments**).

This terminology is typically adopted in the applied practice of econometrics; it commonly appears whenever the adequacy of a specific instrumental variable for estimation is under scrutiny.

The Multivariate IV Estimator

The IV estimator and the related ideas are easily generalized to multivariate regression models, with the usual aid of vector- and matrix-based notation. Our departure point is the usual linear model $y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \varepsilon_i$ with $K \geq 2$, however, now the set of explanatory variables is **partitioned** as:

$$\mathbf{x}_i = \begin{bmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \end{bmatrix}$$

where $\mathbf{x}_{i1}$ is a subset of K_1 **exogenous** regressors with:

$$\mathbb{E}[\varepsilon_i | \boldsymbol{x}_{i1}] = 0$$

while $\mathbf{x}_{i2}$ is a subset of K_2 possibly **endogenous** regressors such that:

$$\mathbb{E}[\varepsilon_i | \boldsymbol{x}_{i2}] \neq 0$$

is allowed by the researcher, and $K_1 + K_2 = K$.

Suppose that a vector of K_2 **instrumental variables**, which is written as $\mathbf{z}_{i2}$, is available, and these IVs are exogenous in the sense that:

$$\mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{z}_{i2}\right] = 0$$

if one groups these along the original K_1 exogenous regressors, writing their realizations as:

$$\mathbf{z}_i = \begin{bmatrix} \mathbf{x}_{i1} \\ \mathbf{z}_{i2} \end{bmatrix}$$

one obtains that upon conditioning on the resulting vector of dimension K, the expectation of the error term is zero as well:

$$\mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{z}_i\right] = 0 \tag{10.28}$$

with the usual covariance implication by the Law of Iterated Expectations.

$$\mathbb{C}\mathrm{ov}\left[\boldsymbol{z}_i, \varepsilon_i\right] = \mathbb{E}\left[\boldsymbol{z}_i \varepsilon_i\right] = \mathbb{E}_{\boldsymbol{z}}\left[\boldsymbol{z}_i \cdot \mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{z}_i\right]\right] = \mathbf{0}$$

In this context, the (**just-identified**) **IV estimator** is defined as:

$$\widehat{\boldsymbol{\beta}}_{IV} = \left(\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i^{\mathrm{T}}\right)^{-1} \sum_{i=1}^{N} \mathbf{z}_i y_i \tag{10.29}$$

which generalizes (10.22), and which is itself a special case of the more general "overidentified" Two-Stages Least Squares estimator, as discussed below. By writing the $N \times K$ matrix that collects the $\mathbf{z}_i$ vectors of exogenous variables as:

$$\mathbf{Z} \equiv \begin{bmatrix} \mathbf{z}_1^{\mathrm{T}} \\ \mathbf{z}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{z}_N^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1K_1} & z_{11} & \dots & z_{1K_2} \\ x_{21} & \dots & x_{2K_1} & z_{21} & \dots & z_{2K_2} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{N1} & \dots & x_{NK_1} & z_{N1} & \dots & z_{NK_2} \end{bmatrix} \tag{10.30}$$

the IV estimator can be elegantly written in compact matrix notation.

$$\widehat{\boldsymbol{\beta}}_{IV} = \left(\mathbf{Z}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{Z}^{\mathrm{T}}\mathbf{y} \tag{10.31}$$

Both representations show straightforwardly that the IV estimator is well-defined so long as matrix $\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i^{\mathrm{T}} = \mathbf{Z}^{\mathrm{T}}\mathbf{X}$ is invertible, and that it admits a decomposition in terms of the "true" parameters $\boldsymbol{\beta}_0$ and of the error terms which is analogous to that of the OLS estimator, (8.2)-(8.3).

$$\widehat{\boldsymbol{\beta}}_{IV} = \boldsymbol{\beta}_0 + \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i^{\mathrm{T}}\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i \varepsilon_i \tag{10.32}$$

$$= \boldsymbol{\beta}_0 + \left(\frac{1}{N}\mathbf{Z}^{\mathrm{T}}\mathbf{X}\right)^{-1} \frac{1}{N}\mathbf{Z}^{\mathrm{T}}\boldsymbol{\varepsilon} \tag{10.33}$$

Clearly, under the exogeneity assumption (10.28) the "remainder" terms on the right-hand sides of both (10.32) and (10.33) converge in probability to $\mathbb{E}[\boldsymbol{z}_i\varepsilon_i] = \mathbf{0}$: **the IV estimator is consistent**.

$$\widehat{\boldsymbol{\beta}}_{IV} \xrightarrow{p} \boldsymbol{\beta}_0 \tag{10.34}$$

Example 10.1. Mincer, Revisited. Let us return to the Mincer Equation from Example 7.2, which is rewritten as (7.41) in Lecture 7:

$$\log W_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 S_i + \varepsilon_i$$

where the error term $\varepsilon_i = \alpha_i + \epsilon_i$ is the sum of unobserved "ability" α_i and some additional residual error ϵ_i. It is, however, very difficult to justify any hypothesis of the form $\mathbb{E}[\alpha_i | S_i] = 0$ (which would imply $\mathbb{E}[\varepsilon_i | S_i] = 0$ if ϵ_i is conditionally mean independent as well). It is clear that education and individual ability are correlated: more skilled individuals tend to get more education, and vice versa – this is essentially a case of omitted variable bias. Hence, one cannot consistently estimate the Mincer equation by OLS.

Suppose that, however, researchers have some *exogenous instrument* Z_i at their disposal, one that:

1. is *exogenous*, in the sense that it does not correlate with unobserved ability and (10.18) holds;
2. satisfies the *exclusion restriction*: it does not affect wages directly;
3. fits a setting that rules out *reverse causality*: education is itself hardly affected by future individual wages;
4. is *relevant*, that is, it correlates with education.

In the literature, there is an innumerable amount of instruments proposed for the education variable S_i; one famous example for Z_i is the "distance of one's home from a college" from a celebrated study by Card (1995). A good exercise is to ask oneself if the four conditions above hold in this example. Regardless of the specific instrument Z_i being chosen, a consistent estimate of the Mincer equation's parameters $\boldsymbol{\beta}_0 = (\beta_0, \beta_1, \beta_2, \beta_3)$ can be obtained by setting:

$$\mathbf{x}_i = \begin{bmatrix} 1 \\ x_i \\ x_i^2 \\ s_i \end{bmatrix} \quad \text{and} \quad \mathbf{z}_i = \begin{bmatrix} 1 \\ x_i \\ x_i^2 \\ z_i \end{bmatrix}$$

where x_i is the experience X_i specific to observation i; s_i is instead his or her education S_i, while z_i is her or his specific value of the instrument Z_i. IV estimation would then proceed as per (10.29) or (10.31). ■

Until now, the formula for the IV estimator – and its asymptotic properties – have been presented without much of a motivation. To gain intuition, it is useful to once again represent the structural relationships between the endogenous and the exogenous variables in the form of a triangular model of simultaneous equations – in a simple case. Suppose, in particular, that $K_1 = K-1$ and $K_2 = 1$, with $\mathbf{x}_{i2} = x_{Ki} = s_i$ (this is similar to an analogous partition from Lecture 7) and $\mathbf{z}_{i0} = z_i$. Thus, only one variable of the main linear model is suspected to be endogenous, with only one instrument z_i to compensate for it. The triangular model representation of this setup is:

$$\begin{aligned} y_i &= \mathbf{x}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_{0\backslash K} + \delta_0 s_i + \varepsilon_i \\ s_i &= \mathbf{x}_{i1}^{\mathrm{T}}\boldsymbol{\pi}_{0\backslash K} + \tau_0 z_i + \eta_i \end{aligned} \tag{10.35}$$

where $\boldsymbol{\beta}_0 = \begin{bmatrix}\boldsymbol{\beta}_{0\backslash K}^{\mathrm{T}} & \delta_0\end{bmatrix}^{\mathrm{T}}$; the reduced form of this model is as follows.

$$\begin{aligned} y_i &= \mathbf{x}_{i1}^{\mathrm{T}}\left(\boldsymbol{\beta}_{0\backslash K} + \delta_0\boldsymbol{\pi}_{0\backslash K}\right) + \delta_0\tau_0 z_i + \varepsilon_i + \delta_0\eta_i \\ s_i &= \mathbf{x}_{i1}^{\mathrm{T}}\boldsymbol{\pi}_{0\backslash K} + \tau_0 z_i + \eta_i \end{aligned} \tag{10.36}$$

This model is, like (10.20), identified, thanks to the **exclusion restriction** whereby the instrument z_i does not enter the structural equation for y_i in (10.35); note that a consistent estimator for δ_0 can be obtained, in analogy with the trivariate case above (10.22), as:

$$\widehat{\delta}_{IV} = \frac{\widehat{\delta\tau}_{OLS}}{\widehat{\tau}_{OLS}} = \frac{\mathbf{z}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{y}}{\mathbf{z}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{z}}\left(\frac{\mathbf{z}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{s}}{\mathbf{z}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{z}}\right)^{-1} = \frac{\mathbf{z}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{y}}{\mathbf{z}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{s}} \tag{10.37}$$

where $\mathbf{y}$, $\mathbf{s}$ and $\mathbf{z}$ are the vectors of length N that collect, respectively, y_i, s_i and z_i; while $\mathbf{M}_{\mathbf{X}1} = \mathbf{I} - \mathbf{X}_1\left(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathrm{T}}$ is the **residual-maker matrix** obtained from the first $K-1$ (exogenous) explanatory regressors $\mathbf{x}_{i1}$. Note that expression (10.37) is best understood as an application of (7.30).

It turns out that the estimator for δ_0 given in (10.37) is **numerically equivalent** to the one be obtained via the IV estimator; it is a good exercise to develop the proof of this result, which is in its essence a variation of the Frisch-Waugh-Lovell Theorem, but applied to the *partitioned* IV estimator. The intuition is likewise analogous: the IV estimator for multivariate linear models extends the simple IV estimator of the slope (10.22) by partialing out the empirical correlations of the instruments with the endogenous variables, as well as the empirical correlation of the instrument with the dependent variable, from the empirical correlations of the other explanatory variables included in the structural form. In this respect, the IV estimator inherits the desirable properties of the least squares solution and the OLS estimator that have been discussed in the previous lectures.

How to perform statistical hypothesis tests based on IV-based estimates of a linear model? In fact, the **asymptotic properties** of the IV estimator are obtained under very general conditions that can be deduced as special cases of the Two-Stages Least Squares estimator and the even more general GMM estimators to be discussed later, and that resemble the six White's assumptions for the large sample properties of OLS. Semi-formally, if matrix $\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i^{\mathrm{T}} = \mathbf{Z}^{\mathrm{T}}\mathbf{X}$ is invertible, if the following probability limits:

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i^{\mathrm{T}} \xrightarrow{p} \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\boldsymbol{z}_i \boldsymbol{x}_i^{\mathrm{T}}\right] \equiv \widetilde{\mathbf{P}}_0 \tag{10.38}$$

$$\frac{1}{N}\sum_{i=1}^{N} \varepsilon_i^2 \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}} \xrightarrow{p} \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\varepsilon_i^2 \boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}}\right] \equiv \widetilde{\mathbf{\Psi}}_0 \tag{10.39}$$

are finite and of full rank, and if *the observations are independent* so that a suitable Central Limit Theorem can be extended to the random sequence $\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \mathbf{z}_i \varepsilon_i$, then the limiting distribution of the IV estimator is:

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \widetilde{\mathbf{P}}_0^{-1}\widetilde{\mathbf{\Psi}}_0\widetilde{\mathbf{P}}_0^{\mathrm{T}-1}\right) \tag{10.40}$$

from which the asymptotic distribution follows accordingly. The asymptotic variance is estimated, for inference purposes, through the following analogue of the heteroscedasticity-consistent (HC) formula of OLS.

$$\widehat{\mathbb{A}\mathrm{var}}_{HC}\left(\widehat{\boldsymbol{\beta}}_{IV}\right) = \left[\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i^{\mathrm{T}}\right]^{-1} \left[\sum_{i=1}^{N} \left(y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{IV}\right)^2 \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}}\right] \left[\sum_{i=1}^{N} \mathbf{x}_i \mathbf{z}_i^{\mathrm{T}}\right]^{-1} \tag{10.41}$$

It is a good exercise to work out the expression of the limiting variance under homoscedasticity as well. Under instances of *dependent observations*, (10.41) would obviously no longer work; in analogy with OLS, the clustering case (CCE) allows the following estimator of the asymptotic variance:

$$\widehat{\mathbb{A}\mathrm{var}}_{CCE}\left(\widehat{\boldsymbol{\beta}}_{IV}\right) = \left[\sum_{c=1}^{C} \mathbf{Z}_c^{\mathrm{T}}\mathbf{X}_c\right]^{-1} \left[\sum_{c=1}^{C} \mathbf{Z}_c^{\mathrm{T}}\mathbf{e}_c\mathbf{e}_c^{\mathrm{T}}\mathbf{Z}_c\right] \left[\sum_{c=1}^{C} \mathbf{X}_c^{\mathrm{T}}\mathbf{Z}_c\right]^{-1} \tag{10.42}$$

where $\mathbf{e}_c \equiv \mathbf{y}_c - \mathbf{X}_c\widehat{\boldsymbol{\beta}}_{IV}$ whereas $\mathbf{Z}_c$ is the cluster-specific sub-matrix of $\mathbf{Z}$, in a similar way as $\mathbf{y}_c$ and $\mathbf{X}_c$ are the cluster-specific collections of the dependent and explanatory variables, respectively. Similarly, HAC estimators for time-series dependence, spatial correlation, or a combination of both, are easily extended to the just-identified IV estimator as well. Importantly, observe how in all these cases the estimated standard errors are inversely proportional to the empirical covariance between the endogenous variables and the instruments, which is subsumed in the matrix $\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i^{\mathrm{T}}$.

Two-Stages Least Squares

It turns out, as already hinted, that the IV estimator is a particular case of the more general (**overidentified**) **Two-Stages Least Squares** (2SLS) **estimator**. This estimator typically applies to cases where the researcher has available **more instruments than endogenous variables**. In other terms, the vector of exogenous instruments has redundant dimensions: this could be written as $|\mathbf{z}_{i2}| = J_2 > K_2$ or equivalently as $|\mathbf{z}_i| = J > K$. While one could in principle obtain different IV estimators for each appropriate subset of $\mathbf{z}_i$ – in this respect the model is **overidentified** – intuitively it appears useful to exploit the additional information that is contained in the "redundant" instruments, simultaneously. The 2SLS estimator was designed to perform precisely this task; it is traditionally attributed to Theil (1953).

The estimator is best illustrated in terms of the two abstract steps, or **stages**, by which it is constructed, and which give it its name.

1. In the **first stage**, one shall perform a set of linear projections, one for each (endogenous) regressors $\mathbf{x}_i$ of the main model of interest, onto the set of exogenous variables (including the instrumental variables) $\mathbf{z}_i$. This results in a set of "projection" vectors $\widehat{\mathbf{x}}_i$ equal to:

$$\widehat{\mathbf{x}}_i^{\mathrm{T}} = \mathbf{z}_i^{\mathrm{T}} \left(\sum_{i=1}^{N} \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}} \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i^{\mathrm{T}} \right) \tag{10.43}$$

   to be collected, in compact matrix notation, as the $N \times K$ matrix:

$$\widehat{\mathbf{X}} = \mathbf{Z} \left(\mathbf{Z}^{\mathrm{T}} \mathbf{Z} \right)^{-1} \mathbf{Z}^{\mathrm{T}} \mathbf{X} = \mathbf{P}_{\mathbf{Z}} \mathbf{X} \tag{10.44}$$

   where $\mathbf{P}_{\mathbf{Z}} \equiv \mathbf{Z} \left(\mathbf{Z}^{\mathrm{T}} \mathbf{Z} \right)^{-1} \mathbf{Z}^{\mathrm{T}}$ is the **projection matrix** based on the exogenous variables collected by the matrix $\mathbf{Z}$ of dimension $N \times J$. This is equivalent to running some **first stage regressions**:

$$x_{ki} = \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\pi}_{k0} + \eta_{ki} \tag{10.45}$$

   and calculating the fitted values $\widehat{x}_{ki} = \mathbf{z}_i^{\mathrm{T}} \widehat{\boldsymbol{\pi}}_{kOLS}$. Observe that if x_{ki} is contained in $\mathbf{z}_i$, it must be $\widehat{x}_{ki} = x_{ki}$, since a vector projected onto itself returns the input vector. Also note the similarity between (10.45) and the second equations of (10.20) and (10.35). This explains why, in the terminology of triangular SEMs, the relationships that "explain" endogenous regressors x_{ki} are called **First Stage equations**.

2. In the **second stage**, the 2SLS estimator of $\boldsymbol{\beta}_0$ is obtained by running an OLS regression of y_i onto the "projected regressors" $\widehat{\mathbf{x}}_i$.

$$y_i = \widehat{\mathbf{x}}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \varepsilon_i \tag{10.46}$$

A 2SLS estimator can be written as simply the OLS estimator of (10.46):

$$\widehat{\boldsymbol{\beta}}_{2SLS} = \left(\sum_{i=1}^{N} \widehat{\mathbf{x}}_i \widehat{\mathbf{x}}_i^{\mathrm{T}} \right)^{-1} \sum_{i=1}^{N} \widehat{\mathbf{x}}_i y_i \tag{10.47}$$

however, it is way more convenient to write it in compact matrix notation:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{2SLS} &= \left(\widehat{\mathbf{X}}^{\mathrm{T}} \widehat{\mathbf{X}} \right)^{-1} \widehat{\mathbf{X}}^{\mathrm{T}} \mathbf{y} \\ &= \left(\mathbf{X}^{\mathrm{T}} \mathbf{P}_{\mathbf{Z}} \mathbf{X} \right)^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{P}_{\mathbf{Z}} \mathbf{y} \end{aligned} \tag{10.48}$$

to understand why the two expressions above are equivalent, recall that a projection matrix is symmetric and idempotent. As it has been mentioned, in the just-identified case ($J = K$), the 2SLS and the IV estimators coincide:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{2SLS} &= \left(\mathbf{X}^{\mathrm{T}} \mathbf{P}_{\mathbf{Z}} \mathbf{X} \right)^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{P}_{\mathbf{Z}} \mathbf{y} \\ &= \left[\mathbf{X}^{\mathrm{T}} \mathbf{Z} \left(\mathbf{Z}^{\mathrm{T}} \mathbf{Z} \right)^{-1} \mathbf{Z}^{\mathrm{T}} \mathbf{X} \right]^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{Z} \left(\mathbf{Z}^{\mathrm{T}} \mathbf{Z} \right)^{-1} \mathbf{Z}^{\mathrm{T}} \mathbf{y} \\ &= \left(\mathbf{Z}^{\mathrm{T}} \mathbf{X} \right)^{-1} \mathbf{Z}^{\mathrm{T}} \mathbf{Z} \left(\mathbf{X}^{\mathrm{T}} \mathbf{Z} \right)^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{Z} \left(\mathbf{Z}^{\mathrm{T}} \mathbf{Z} \right)^{-1} \mathbf{Z}^{\mathrm{T}} \mathbf{y} \\ &= \left(\mathbf{Z}^{\mathrm{T}} \mathbf{X} \right)^{-1} \mathbf{Z}^{\mathrm{T}} \mathbf{y} \\ &= \widehat{\boldsymbol{\beta}}_{IV} \end{aligned}$$

where the third line is only possible if $\mathbf{X}$ and $\mathbf{Z}$ have the same (column) dimensions. The usual decomposition gives:

$$\widehat{\boldsymbol{\beta}}_{2SLS} = \boldsymbol{\beta}_0 + \left(\frac{1}{N} \sum_{i=1}^{N} \widehat{\mathbf{x}}_i \widehat{\mathbf{x}}_i^{\mathrm{T}} \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} \widehat{\mathbf{x}}_i \varepsilon_i \tag{10.49}$$

$$= \boldsymbol{\beta}_0 + \left(\frac{1}{N} \mathbf{X}^{\mathrm{T}} \mathbf{P}_{\mathbf{Z}} \mathbf{X} \right)^{-1} \frac{1}{N} \mathbf{X}^{\mathrm{T}} \mathbf{P}_{\mathbf{Z}} \boldsymbol{\varepsilon} \tag{10.50}$$

which shows, in conjuction with the geometric interpretation of the 2SLS estimator, why the latter is consistent. In fact, the operation of projecting the possibly endogenous regressors $\mathbf{x}_i$ onto the space which is spanned by the exogenous variables $\mathbf{z}_i$ – call it $\mathcal{S}(\mathbf{Z})$ – generates a set of "fitted" regressors $\widehat{\mathbf{x}}_{ki}$ that, **by construction**, lie on $\mathcal{S}(\mathbf{Z})$, and are consequently orthogonal to the disturbance vector $\boldsymbol{\varepsilon}$. Again, this is by construction because of (10.28); see figure 10.3 for the geometric intuition. Since $\mathbb{E}[\widehat{\boldsymbol{x}}_i \varepsilon_i] = \mathbf{0}$, it is:

$$\frac{1}{N} \mathbf{X}^{\mathrm{T}} \mathbf{P}_{\mathbf{Z}} \boldsymbol{\varepsilon} = \frac{1}{N} \sum_{i=1}^{N} \widehat{\mathbf{x}}_i \varepsilon_i \xrightarrow{p} \mathbf{0} \tag{10.51}$$

implying consistency of 2SLS estimator, i.e. $\widehat{\boldsymbol{\beta}}_{2SLS} \xrightarrow{p} \boldsymbol{\beta}_0$.

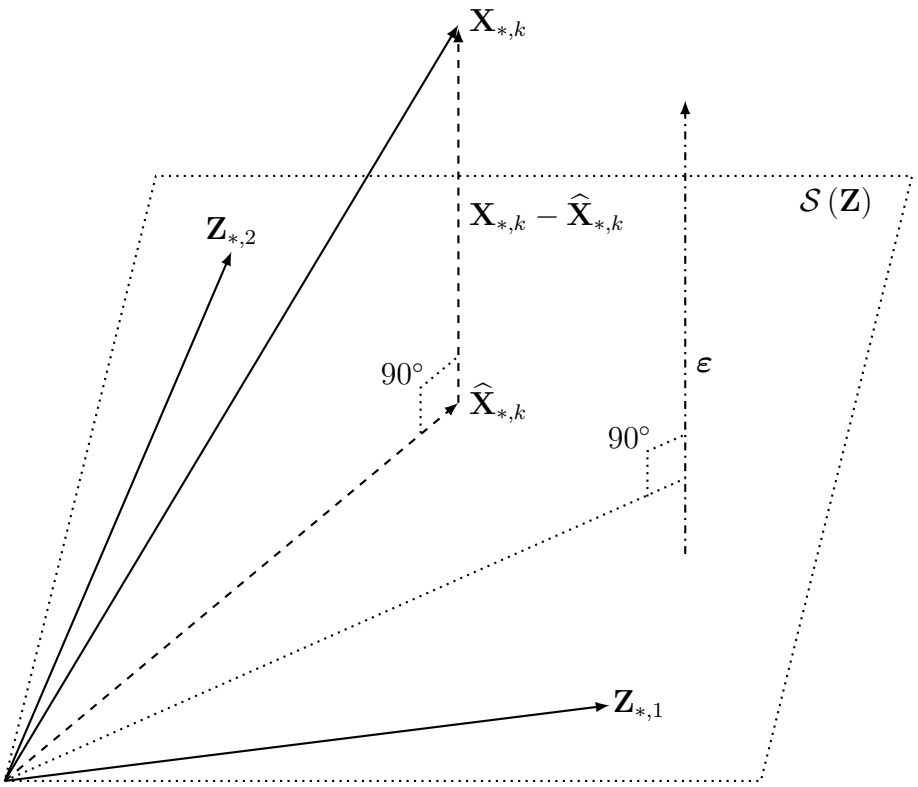


Figure 10.3: The geometric interpretation of the 2SLS estimator

Example 10.2. Mincer, Revisited (again). Let us return yet once more to the Mincer equation, and the attempt to address the endogeneity bias due to the omission of ability α_i which is explored in example 10.1. One can obtain the IV estimator in two alternative ways; both require specifying a **first stage equation** for education, like:

$$S_i = \pi_0 + \pi_1 X_i + \pi_2 X_i^2 + \pi_3 Z_i + \eta_i \tag{10.52}$$

note that this is a linear version of the structural equation for ability (7.7) from example 7.2, with the inclusion of a squared term for experience (as in the "structural" Mincer equation) but the exclusion of ability itself. The resulting reduced form of the model is:

$$\begin{aligned} \log W_i &= \beta_0 + \beta_3\pi_0 + (\beta_1 + \beta_3\pi_1) X_i + (\beta_2 + \beta_3\pi_2) X_i^2 \\ &\quad + \beta_3\pi_3 Z_i + \alpha_i + \epsilon_i + \beta_3\eta_i \\ S_i &= \pi_0 + \pi_1 X_i + \pi_2 X_i^2 + \pi_3 Z_i + \eta_i \end{aligned} \tag{10.53}$$

which, once again, is just identified as long as $\mathbb{E}\left[\alpha_i, \epsilon_i, \eta_i \middle| X_i, Z_i\right] = 0$.

By the Frisch-Waugh-Lovell Theorem, a consistent estimate of β_3 which is numerically identical to the IV estimator can be obtained – see (10.37) – as:

$$\widehat{\beta}_3 = \frac{\widehat{\beta_3 \pi_3}}{\widehat{\pi}_3} = \frac{\mathbf{z}^{\mathrm{T}} \mathbf{M}_{\mathbf{X}} \mathbf{y}}{\mathbf{z}^{\mathrm{T}} \mathbf{M}_{\mathbf{X}} \mathbf{s}}$$

and the other coefficients from the Mincer equation can similarly be backed up via through the reduced form estimates. Alternatively, one could obtain exactly the same numerical estimate through the linear projection of S_i that results from the First Stage estimation of (10.52):

$$\widehat{S}_i = \widehat{\pi}_0 + \widehat{\pi}_1 X_i + \widehat{\pi}_2 X_i^2 + \widehat{\pi}_3 Z_i$$

and the just-identified IV-2SLS estimator results from running OLS on the following model.

$$\log W_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 \widehat{S}_i + \varepsilon_i$$

However, the 2SLS estimator would work just as well in an overidentified setting where the researcher has access to redundant instruments. Suppose that there are two additional valid instruments G_i and F_i; the First Stage equation would then read as:

$$S_i = \pi_0 + \pi_1 X_i + \pi_2 X_i^2 + \pi_3 Z_i + \pi_4 G_i + \pi_5 F_i + \eta_i \tag{10.54}$$

and 2SLS estimation would proceed as described. In Lecture 12, which is about the more general GMM framework, this case is revisited in order to develop an example of an *overidentification test*, while making at the same time some substantive examples of "additional" instruments G_i and F_i. ■

It remains to show the remaining asymptotic properties of the 2SLS estimator, especially with regard to the variance. These are now presented more formally than how it was done in the context of the just-identified IV estimator; it is a good exercise to derive the asymptotic properties of the latter as a particular case. In what follows, some additional assumptions of the **generalized linear model** – one that allows for instrumental variables and possibly overidentification – are stated more rigorously.

Assumption 9. Independently but not identically distributed IVs. The observations in the sample $\{(y_i, \mathbf{x}_i, \mathbf{z}_i)\}_{i=1}^N$ are *independently*, but *not* necessarily *identically*, distributed (i.n.i.d.).

Assumption 10. Exogeneity of the Instruments. Conditional on the $J \geq K$ regressors $\boldsymbol{z}_i$, the error term ε_i has mean zero.

$$\mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{z}_i\right] = 0 \tag{10.55}$$

Assumption 11. Asymptotics of the Projected Regressors. The following probability limit exists, is finite, and has full rank.

$$\frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{X} \xrightarrow{p} \mathbf{P}_0 \tag{10.56}$$

Assumption 12. Heteroscedastic, Independent Errors. The variance of the error term ε_i, conditional on the instruments $\boldsymbol{z}_i$, is unrestricted (*heteroscedasticity*), while the conditional covariance between two error terms from two different observations $i, j = 1, \ldots, N$ is zero.

$$\mathbb{E}\left[\varepsilon_i^2 \middle| \boldsymbol{z}_i\right] = \sigma^2(\boldsymbol{z}_i) \equiv \sigma_i^2 \tag{10.57}$$

$$\mathbb{E}\left[\varepsilon_i \varepsilon_j \middle| \boldsymbol{z}_i, \boldsymbol{z}_j\right] = 0 \tag{10.58}$$

Assumption 13. Asymptotics of Projected Regressors interacted with the Errors. Given the following diagonal matrix of squared errors:

$$\mathbf{E} \equiv \begin{bmatrix} \varepsilon_1^2 & 0 & \cdots & 0 \\ 0 & \varepsilon_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \varepsilon_N^2 \end{bmatrix} \tag{10.59}$$

the following probability limit exists, is finite, and has full rank.

$$\frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{E}\mathbf{P}_{\mathbf{Z}}\mathbf{X} \xrightarrow{p} \mathbf{\Psi}_0 \tag{10.60}$$

In addition, conditions hold so that the following Central Limit Theorem result applies.

$$\frac{1}{\sqrt{N}}\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\boldsymbol{\varepsilon} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{\Psi}_0) \tag{10.61}$$

These assumption are worth to be discussed briefly in relationship with the analogous (White's) hypotheses of OLS. Assumption 9 simply extends Assumption 2 to the instruments listed in $\boldsymbol{z}_i$. Assumption 10 was discussed multiple times; it is necessary to obtain (10.51) and hence consistency of the 2SLS estimator. Assumption 11 allows to establish a condition analogous to (8.6), for the sake of establishing and estimating the asymptotic variance. Assumption 12 characterizes the concept of heteroscedasticity in the context of instrumental variables. Finally, Assumption 13 ensures that some Central Limit Theorem can be appropriately extended to the 2SLS estimator as well. Observe that some of these assumptions might be founded – more rigorously – onto more primitive hypotheses (like conditions on specific moments, e.g. Ljapunov's); this is not pursued here for the sake of conciseness.

In light of these assumptions, one can finally establish the asymptotic properties of the 2SLS estimator.

Theorem 10.1. Large Sample properties of the 2SLS Estimator. *Under Assumptions 1, 3 and 9-13, the 2SLS estimator is consistent:*

$$\widehat{\boldsymbol{\beta}}_{2SLS} \xrightarrow{p} \boldsymbol{\beta}_0 \tag{10.62}$$

and it is asymptotically normal, that is:

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{2SLS} - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{P}_0^{-1}\boldsymbol{\Psi}_0\mathbf{P}_0^{-1}\right) \tag{10.63}$$

hence its asymptotic distribution, for a given N, is as follows.

$$\widehat{\boldsymbol{\beta}}_{2SLS} \overset{A}{\sim} \mathcal{N}\left(\boldsymbol{\beta}_0, \frac{1}{N}\mathbf{P}_0^{-1}\boldsymbol{\Psi}_0\mathbf{P}_0^{-1}\right) \tag{10.64}$$

Proof. The proof is analogous to the one for the OLS case; it exploits the decomposition (10.50), the asymptotic results from Assumptions 11 and 13, as well as Slutskij's Theorem and the Cramér-Wold device. □

At this point, asymptotic properties such as these should appear familiar, at least by comparison with OLS. In this case, though, the expressions for the *estimators* of the asymptotic variance are less straightforward. The heteroscedasticity-robust estimator of the asymptotic variance follows from (10.56) and (10.60), and it reads:

$$\widehat{\mathbb{A}\text{var}}\left(\widehat{\boldsymbol{\beta}}_{2SLS}\right) = \frac{1}{N}\widehat{\mathbf{P}}_N^{-1}\widehat{\boldsymbol{\Psi}}_N\widehat{\mathbf{P}}_N^{-1} \tag{10.65}$$

where:

$$\widehat{\mathbf{P}}_N \equiv \frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{X} \tag{10.66}$$

$$\widehat{\boldsymbol{\Psi}}_N \equiv \frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\widehat{\mathbf{E}}_N\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{X} \tag{10.67}$$

with $\widehat{\mathbf{P}}_N \xrightarrow{p} \mathbf{P}_0$, $\widehat{\boldsymbol{\Psi}}_N \xrightarrow{p} \boldsymbol{\Psi}_0$; while $\widehat{\mathbf{E}}_N$ is the following diagonal matrix:

$$\widehat{\mathbf{E}}_N \equiv \begin{bmatrix} e_1^2 & 0 & \cdots & 0 \\ 0 & e_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e_N^2 \end{bmatrix} \tag{10.68}$$

where $e_i \equiv y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{2SLS}$ for $i = 1, \ldots, N$. Naturally, this would not work with dependent observations as Assumption 13 would fail; estimating the appropriate "meat" matrix $\boldsymbol{\Psi}_0$ would require a CCE approach, as follows:

$$\widehat{\boldsymbol{\Psi}}_{CCE} \equiv \frac{1}{N}\left[\sum_{c=1}^{C}\mathbf{X}_c^{\mathrm{T}}\mathbf{Z}_c\left(\mathbf{Z}_c^{\mathrm{T}}\mathbf{Z}_c\right)^{-1}\mathbf{Z}_c^{\mathrm{T}}\mathbf{e}_c\mathbf{e}_c^{\mathrm{T}}\mathbf{Z}_c\left(\mathbf{Z}_c^{\mathrm{T}}\mathbf{Z}_c\right)^{-1}\mathbf{Z}_c^{\mathrm{T}}\mathbf{X}_c\right] \xrightarrow{p} \boldsymbol{\Psi}_0 \tag{10.69}$$

where $\mathbf{e}_c \equiv \mathbf{y}_c - \mathbf{X}_c\widehat{\boldsymbol{\beta}}_{2SLS}$. Clearly, analogous HAC estimators also exist.

The case of homoscedasticity is of particular interest in the 2SLS setting. Here, homoscedasticity implies the variance-independence of the error term with respect to the instruments $\boldsymbol{z}_i$ in the following sense:

$$\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\, \boldsymbol{Z}\right] = \sigma_0^2\mathbf{I}$$

and one can show that:

$$\frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{E}\mathbf{P}_{\mathbf{Z}}\mathbf{X} \xrightarrow{p} \sigma_0^2 \cdot \lim_{N\to\infty}\frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{X} = \sigma_0^2\mathbf{P}_0 \tag{10.70}$$

hence, $\boldsymbol{\Psi}_0 = \sigma_0^2\mathbf{P}_0$ and:

$$\widehat{\boldsymbol{\beta}}_{2SLS} \overset{A}{\sim} \mathcal{N}\left(\boldsymbol{\beta}_0, \frac{\sigma_0^2}{N}\mathbf{P}_0^{-1}\right) \tag{10.71}$$

which, for $\mathbf{e} \equiv \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{2SLS}$, is estimated as follows.

$$\widehat{\mathbb{A}\mathrm{var}}\left(\widehat{\boldsymbol{\beta}}_{2SLS}\right) = \frac{\mathbf{e}^{\mathrm{T}}\mathbf{e}}{N}\left[\mathbf{X}^{\mathrm{T}}\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{X}\right]^{-1} \tag{10.72}$$

A result analogous to the Gauss-Markov Theorem for OLS shows that the under homoscedasticity, the 2SLS estimator is the **most efficient** linear estimator that can be constructed under the "exogeneity" hypotheses (10.55). The *traditional* practice of econometrics emphasizes the importance of gathering as many instrumental variables as possible in order to approximate as close as possible the efficiency bound represented by (10.71).

Example 10.3. 2SLS and structural endogeneity. This efficiency result is especially useful to address issues of "structural endogeneity." Consider the time series model (10.16); through iterated substitution, it can be restated as:

$$y_t = \beta_0 + \sum_{s=0}^{t-1}\beta_1^s\left(\mathbf{x}_{t-s}^{\mathrm{T}}\boldsymbol{\gamma}_0 + \mathbf{x}_{t-1-s}^{\mathrm{T}}\boldsymbol{\gamma}_1\right) + \sum_{s=0}^{t-1}\sum_{z=0}^{t-s}\beta_1^s\rho^z\xi_{t-s-z}$$

(this assumes the existence of some $(\mathbf{x}_0, \xi_0)$ at $t = 0$). The above relationship, lagged by one further period, applies to the endogenous y_{t-1} variable as well. Hence, **all valid lags** $\mathbf{x}_{t-s}$ for $s \geq 2$ can be combined into the instruments vector $\mathbf{z}_t$ in a 2SLS framework. If, in addition, ξ_t is homoscedastic, this leads to efficient estimates. In a similar vein, the solution of the spatial model (10.17) can be rephrased, by standard results of linear algebra, as:

$$\mathbf{y} = \sum_{s=0}^{\infty}\beta_1^s\mathbf{W}^s\left(\beta_0\boldsymbol{\iota} + \mathbf{X}\boldsymbol{\gamma}_0 + \mathbf{W}\mathbf{X}\boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon}\right)$$

thus, **all the linearly independent** columns of the matrices in $\{\mathbf{W}^s\mathbf{X}\}_{s=2}^{\infty}$ can enter the instruments matrix $\mathbf{Z}$ (Kelejian and Prucha, 1998; Bramoullé et al., 2009); with homoscedastic errors, this leads to efficient estimates. ■

10.3 Instrumental Variables in Practice

The previous section develops the basic theory of IV-2SLS estimation. Due to the importance of these estimators in the applied econometric practice, it is worth to separately discuss a series of technical aspects that are often useful in actual econometric research. In this section, the following topics are discussed: **control function approaches** and **tests for endogeneity** (these are *options* available to researchers), as well as the **the small sample bias of IV-2SLS** and the issue of **weak instruments** (these, instead, are *cautionary warnings* about the possible misuse of instrumental variables).

Control function approaches

A **control function approach** is a method to address endogeneity in an econometric model which is predicated on a specification of the endogeneity problem that is "built-in" the error term ε_i. This definition is quite general, but in the case of linear models, the control function approach is actually complementary to IV-2SLS, and it is possible so long as there are at least as many exogenous instruments as there are endogenous explanatory variables. With $J_2 \geq K_2$ instruments, it works as follows:

1. one shall first run some K_2 first stage regressions like (10.45) (one for each of the endogenous variables) and **calculates the residuals** from these equations: $\widehat{\eta}_{ki} \equiv x_{ki} - \mathbf{z}_i^{\mathrm{T}}\widehat{\boldsymbol{\pi}}_{kOLS} = x_{ki} - \widehat{x}_{ki}$ for $k = 1, \ldots, K_2$;

2. in the alternative second stage one would run OLS regressions on the structural form **augmented with the estimated residuals**:

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + \widehat{\boldsymbol{\eta}}_i^{\mathrm{T}}\boldsymbol{\rho}_0 + \varsigma_i \tag{10.73}$$

   where $\widehat{\boldsymbol{\eta}}_i = \begin{bmatrix}\widehat{\eta}_{1i} & \widehat{\eta}_{2i} & \ldots & \widehat{\eta}_{K_2 i}\end{bmatrix}^{\mathrm{T}}$ while $\boldsymbol{\rho}_0$ collects the K_2 parameters associated with each set of residuals and ς_i is some new error term.

The OLS estimates of (10.73) are actually consistent for both $\boldsymbol{\beta}_0$ and $\boldsymbol{\rho}_0$. A semi-formal argument, intuitive if convoluted, is provided next.

Consider, for $k = 1, \ldots, K_2$, the first stage model for the k-th endogenous variable (10.45); and note that, by construction:

$$\mathbb{E}\left[\eta_{ki}\varepsilon_i\right] = \mathbb{E}\left[\left(X_{ki} - \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\pi}_0\right)\varepsilon_i\right] = \underbrace{\mathbb{E}\left[X_{ki}\varepsilon_i\right]}_{\neq 0} - \underbrace{\mathbb{E}\left[\boldsymbol{z}_i\varepsilon_i\right]^{\mathrm{T}}}_{=0}\boldsymbol{\pi}_0 \neq 0$$

therefore, the error term of the k-th equation η_{ki} appears to contain some statistical information about what makes variable X_{ki} endogenous in the

first place (in fact this is so by construction, since η_{ki} is the residual from the projection of X_{ki} onto the exogenous instruments). One might want to extend this intuition to all the other K_2 first stage residuals η_{ki} by specifying a **statistical model for the error term**, also called a **control function**. If one supposes that such a model is linear:

$$\varepsilon_i = \boldsymbol{\eta}_i^{\mathrm{T}} \boldsymbol{\rho}_0 + \xi_i \tag{10.74}$$

where, *in the population*:

$$\boldsymbol{\rho}_0 \equiv \mathbb{E}\left[\boldsymbol{\eta}_i \boldsymbol{\eta}_i^{\mathrm{T}}\right]^{-1} \mathbb{E}\left[\boldsymbol{\eta}_i \varepsilon_i\right]$$

is the **linear projection** of ε_i onto $\boldsymbol{\eta}_i = \begin{bmatrix} \eta_{1i} & \eta_{2i} & \dots & \eta_{K_2 i} \end{bmatrix}^{\mathrm{T}}$, while ξ_i is defined residually. Expression $\boldsymbol{\eta}_i^{\mathrm{T}} \boldsymbol{\rho}_0$ represents the **extent of endogeneity** contained in the model. Note that, by assumption:

$$\mathbb{E}\left[\boldsymbol{z}_i \xi_i\right] = \underbrace{\mathbb{E}\left[\boldsymbol{z}_i \varepsilon_i\right]}_{=\mathbf{0}} - \underbrace{\mathbb{E}\left[\boldsymbol{z}_i \boldsymbol{\eta}_i^{\mathrm{T}}\right]}_{=\mathbf{0}} \boldsymbol{\rho}_0 = \mathbf{0}$$

because the instruments are exogenous; moreover, for $k = 1, \dots, K_2$:

$$\mathbb{E}\left[\eta_{ki} \xi_i\right] = 0$$

by definition of linear projection, that is by construction, and:

$$\mathbb{E}\left[X_{ki} \xi_i\right] = \underbrace{\mathbb{E}\left[\boldsymbol{z}_i \xi_i\right]^{\mathrm{T}}}_{=\mathbf{0}} \boldsymbol{\pi}_0 + \underbrace{\mathbb{E}\left[\eta_{ki} \xi_i\right]}_{=0} = 0$$

as implied by the observations above. Substituting (10.74) into the original structural equation $y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \varepsilon_i$ gives:

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \boldsymbol{\eta}_i^{\mathrm{T}} \boldsymbol{\rho}_0 + \xi_i \tag{10.75}$$

which could be consistently estimated by OLS if only $\boldsymbol{\eta}_i$ could be observed. While $\boldsymbol{\eta}_i$ cannot be observed by definition, it can be definitely estimated in the first stage, hence (10.73) matches (10.75) for

$$\varsigma_i \equiv \left(\boldsymbol{\eta}_i - \widehat{\boldsymbol{\eta}}_i\right)^{\mathrm{T}} \boldsymbol{\rho}_0 + \xi_i = \mathbf{z}_i^{\mathrm{T}} \cdot \sum_{k=0}^{K_0} \left(\boldsymbol{\pi}_{k0} - \widehat{\boldsymbol{\pi}}_{k,OLS}\right) \rho_{k0} + \xi_i$$

and it can be shown that the first component of this expression is conditionally mean independent of $(\boldsymbol{x}_i, \widehat{\boldsymbol{\eta}}_i)$, because it only depends on the statistical noise in the estimation of the first stage models. Consequently:

$$\mathbb{E}\left[\boldsymbol{x}_i^{\mathrm{T}} \varsigma_i\right] = \mathbf{0} \quad \text{and} \quad \mathbb{E}\left[\widehat{\boldsymbol{\eta}}_i^{\mathrm{T}} \varsigma_i\right] = \mathbf{0}$$

hence, OLS estimation of (10.73) is both consistent and, for $\boldsymbol{\beta}_0$, equivalent to IV-2SLS. The classical, full-fledged proof of this result is based on a variation of the Frisch-Waugh-Lovell Theorem and the algebra of projections.

In practice, control function approaches are seldom used for linear models. With respect to IV-2SLS, in fact, they entail a few shortcomings: first, they might not work too well if the endogenous variables entered the structural form non-linearly (with higher-order terms, interactions etc.); second, they can be shown to be less efficient and to produce larger standard errors. One might wonder, then, what are control function approaches useful for. Not only they play a role in the **tests for endogeneity**, as it is mentioned below; but they can be actually convenient for extending instrumental variables to **non-linear models**. In fact, IVs can be combined with non-linear models in a variety of ways (see e.g. the discussion in Lecture 12 about the GMM approach) but in practical terms, these often entail complications of computational or statistical kind. Conversely, control function approaches are very flexible; they typically entail augmenting a non-linear model with the inclusion of some function of the residuals obtained from the first-stage regressions of the endogenous variables.

Tests for endogeneity

After performing IV-2SLS estimation, researchers might ask themselves if anything was gained in terms of the overall quality of their estimates, that is, if they differ from the OLS results in a way that reveals the presence of substantial endogeneity in the original model. This can be formulated as a **test** where the null hypothesis is as follows.

$$\mathrm{H}_0 : \mathbb{E}\left[\boldsymbol{x}_i^{\mathrm{T}} \varepsilon_i\right] = \mathbf{0}$$

The above implies that OLS and IV-2SLS have the same probability limit; thus, an operationally more convenient null hypothesis can be:

$$\mathrm{H}_0 : \operatorname{plim} \widehat{\boldsymbol{\beta}}_{OLS} - \operatorname{plim} \widehat{\boldsymbol{\beta}}_{2SLS} = \mathbf{0}$$

which suggests a natural test statistic.

$$\widetilde{\mathcal{H}}_{\mathrm{H}_0} = \left[\widehat{\boldsymbol{\beta}}_{OLS} - \widehat{\boldsymbol{\beta}}_{2SLS}\right]^{\mathrm{T}} \widehat{\mathbb{A}\mathrm{var}}\left[\widehat{\boldsymbol{\beta}}_{OLS} - \widehat{\boldsymbol{\beta}}_{2SLS}\right] \left[\widehat{\boldsymbol{\beta}}_{OLS} - \widehat{\boldsymbol{\beta}}_{2SLS}\right] \xrightarrow{p} \chi^2_K$$

Under the null hypothesis, this quadratic form should be close to zero. The problem is that, in general, it is hard to derive an exact expression for the variance of the difference between two estimators, like:

$$\mathbb{V}\mathrm{ar}\left[\widehat{\boldsymbol{\beta}}_{OLS} - \widehat{\boldsymbol{\beta}}_{2SLS}\right] = \mathbb{V}\mathrm{ar}\left[\widehat{\boldsymbol{\beta}}_{OLS}\right] + \mathbb{V}\mathrm{ar}\left[\widehat{\boldsymbol{\beta}}_{2SLS}\right] - 2\,\mathbb{C}\mathrm{ov}\left[\widehat{\boldsymbol{\beta}}_{OLS}, \widehat{\boldsymbol{\beta}}_{2SLS}\right]$$

since the covariance component is generally unknown. However, Hausman (1978) showed that if one of the two estimator is **efficient**, as in the case of

the OLS estimator under i.i.d. homoscedastic errors (by the Gauss-Markov Theorem), the unknown covariance is actually equal to the variance of the less efficient estimator, therefore:

$$\mathbb{C}\text{ov}\left[\widehat{\boldsymbol{\beta}}_{OLS}, \widehat{\boldsymbol{\beta}}_{2SLS}\right] = \mathbb{V}\text{ar}\left[\widehat{\boldsymbol{\beta}}_{2SLS}\right]$$

which allows to formulate the **Hausman test statistic**:[3]

$$\mathcal{H}_{\text{H}_0} = \left[\widehat{\boldsymbol{\beta}}_{OLS} - \widehat{\boldsymbol{\beta}}_{2SLS}\right]^{\text{T}} \left\{\widehat{\mathbb{A}\text{var}}\left[\widehat{\boldsymbol{\beta}}_{OLS}\right] - \widehat{\mathbb{A}\text{var}}\left[\widehat{\boldsymbol{\beta}}_{2SLS}\right]\right\} \left[\widehat{\boldsymbol{\beta}}_{OLS} - \widehat{\boldsymbol{\beta}}_{2SLS}\right]$$

with $\mathcal{H} \xrightarrow{p} \chi^2_K$ asymptotically. This statistic is easily calculated in the data.

Unfortunately, in regression analysis the Hausman test is limited to the case of i.i.d. homoscedastic errors, or to other scenarios where an alternative estimator is evaluated against an efficient benchmark (such as fixed effects vs. random effects in panel data). Even with i.i.d. errors, however, it can be shown that the Hausman test converges in probability to a Wald statistic formed out of the $\widehat{\boldsymbol{\rho}}_{OLS}$ from the control function estimator of (10.73)! The intuition is simple in light of the earlier discussion about control function: the null hypothesis of "no endogeneity" is equivalent to:

$$\text{H}_0 : \boldsymbol{\rho}_0 = \mathbf{0}$$

since it implies $\mathbb{E}\left[\boldsymbol{\eta}_i^{\text{T}} \varepsilon_i\right] = \mathbb{E}\left[\boldsymbol{x}_i^{\text{T}} \varepsilon_i\right] = \mathbf{0}$. This observation is not only useful as an alternative route for calculating the Hausman test. In fact, it suggest methods for performing **"regression-based" tests for endogeneity** with heteroscedastic or dependent errors, which are allowed by control function approaches. Various tests of this sort exist; see e.g. Wooldridge (1995).

Small sample bias of IV-2SLS

One might think that, since the IV-2SLS estimator is consistent, in analogy with OLS it is also unbiased in small samples. This is false. Observe that:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\boldsymbol{\beta}}_{2SLS}\right] &= \boldsymbol{\beta}_0 + \mathbb{E}\left[\left(\boldsymbol{X}^{\text{T}}\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\text{T}}\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{\varepsilon}\right] \\
&= \boldsymbol{\beta}_0 + \mathbb{E}_{\boldsymbol{X},\boldsymbol{Z}}\left[\mathbb{E}\left[\left(\boldsymbol{X}^{\text{T}}\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\text{T}}\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{\varepsilon}\,\middle|\,\boldsymbol{X},\boldsymbol{Z}\right]\right] \\
&= \boldsymbol{\beta}_0 + \mathbb{E}_{\boldsymbol{X},\boldsymbol{Z}}\left[\left(\boldsymbol{X}^{\text{T}}\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\text{T}}\boldsymbol{P}_{\boldsymbol{Z}} \cdot \mathbb{E}\left[\boldsymbol{\varepsilon}|\,\boldsymbol{X},\boldsymbol{Z}\right]\right]
\end{aligned}$$

and it is impossible to simplify the expression further, since $\mathbb{E}\left[\boldsymbol{\varepsilon}|\,\boldsymbol{X},\boldsymbol{Z}\right] \neq \mathbf{0}$ because of endogeneity: the IV-2SLS estimator is **biased in small samples** and the bias goes towards the direction of the inconsistent OLS estimator.

[3]The Hausman test is sometimes also called Durbin-Wu-Hausman test, in light of two earlier contributions (Durbin, 1954; Wu, 1973) that proposed tests similar to Hausman's.

Some theoretical research has attempted to quantify this bias and to develop procedures for testing it. The practice of empirical research which is based on IV-2SLS estimation emphasizes the use of large datasets in order to rely on the better asymptotic properties of the estimator.

Weak instruments

It was observed in several instances that the limiting variance of IV-2SLS estimators is inversely proportional to the correlation between the endogenous regressors and the exogenous instruments. This relates to the intuition for identification in IV estimation: the "effect" of some endogenous variable X_i on Y_i is obtained through the indirect effect that the instrument Z_i has on Y_i, since Z_i affects X_i directly but it affects Y_i only through X_i. It is then self-evident that if the direct relationship between X_i and Z_i is statistically weak, the main effect of interest is hard to "capture" and it will be at best imprecisely estimated. This is a problem of **weak instruments**.

Weak instruments have two main implications. First, IV-2SLS estimates obtained with weak instruments might make predictions worse than the ones obtained with inconsistent OLS estimates, in a Mean Squared Error sense. Asymptotically, the latter reads (particularly for IV-2SLS) as:

$$\operatorname{plim} \mathrm{MSE}_{IV-2SLS} = \underbrace{\left(\operatorname{plim} \widehat{\boldsymbol{\beta}}_{IV-2SLS} - \boldsymbol{\beta}_0\right)^2}_{= \textit{squared asymptotic bias}} + \underbrace{\mathbb{A}\mathrm{var}\left(\widehat{\boldsymbol{\beta}}_{IV-2SLS}\right)}_{= \textit{asymptotic variance}}$$

hence, the gains obtained in terms of lower bias might be more than offset by the losses due to a higher variance. Second, if the instruments are weak and *only slightly endogenous*, the "cure" to endogeneity achieved by IV-2SLS estimation might be worse then the disease. To appreciate this, consider the ratio between the asymptotic bias of the IV estimator (10.22) of the trivariate triangular model, and the OLS estimator (5.15) from bivariate regression:

$$\begin{aligned} \frac{\operatorname{plim} \beta_{1,IV} - \beta_1}{\operatorname{plim} \beta_{1,OLS} - \beta_1} &= \frac{\mathbb{C}\mathrm{ov}\left(Z_i, \varepsilon_i\right)}{\mathbb{C}\mathrm{ov}\left(Z_i, X_i\right)} \cdot \frac{\mathbb{V}\mathrm{ar}\left(X_i\right)}{\mathbb{C}\mathrm{ov}\left(X_i, \varepsilon_i\right)} \\ &= \frac{\mathbb{C}\mathrm{orr}\left(Z_i, \varepsilon_i\right)}{\mathbb{C}\mathrm{orr}\left(X_i, \varepsilon_i\right)} \cdot \frac{1}{\mathbb{C}\mathrm{orr}\left(Z_i, X_i\right)} \end{aligned}$$

which is obtained by elaborating the probability limit of the two estimators under the assumptions that $\mathbb{C}\mathrm{ov}\left(Z_i, \varepsilon_i\right) \neq 0$ and $\mathbb{C}\mathrm{ov}\left(X_i, \varepsilon_i\right) \neq 0$. Observe that even if the instrument is, while not completely exogenous, somewhat "less endogenous" – in the sense that $\mathbb{C}\mathrm{ov}\left(Z_i, \varepsilon_i\right) < \mathbb{C}\mathrm{ov}\left(X_i, \varepsilon_i\right)$ – a weak

instrument might actually amplify the endogeneity problem. This intuition is easily generalized to higher dimensional problems. Consider for example a triangular model with one endogenous regressor s_i, similar to (10.35) but possibly overidentified:

$$\begin{aligned} y_i &= \mathbf{x}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_{0\setminus K} + \delta_0 s_i + \varepsilon_i \\ s_i &= \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\pi}_0 + \eta_i \end{aligned}$$

and assume further that it has i.i.d. errors. It is possible to show that:

$$\frac{\operatorname{plim}\delta_{IV} - \delta_0}{\operatorname{plim}\delta_{OLS} - \delta_0} = \frac{\mathbb{C}\mathrm{orr}\left(\widehat{S}_i, \varepsilon_i\right)}{\mathbb{C}\mathrm{orr}\left(S_i, \varepsilon_i\right)} \cdot \frac{1}{\operatorname{plim}\mathcal{R}^2_{s,\mathbf{z}|\mathbf{x}}}$$

where $\mathcal{R}^2_{s,\mathbf{z}|\mathbf{x}}$ is the following **partialed out R-squared coefficient**.

$$\mathcal{R}^2_{s,\mathbf{z}|\mathbf{x}} = \frac{\mathbf{s}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{M}_{\mathbf{X}1}\mathbf{P}_{\mathbf{Z}}\mathbf{s}}{\mathbf{s}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{s}}$$

Notice that, by the Frisch-Waugh-Lovell Theorem, this is the R^2 coefficient that would be obtained from a regression of s_i on $\mathbf{z}_i$, after **partialing out** the exogenous regressors $\mathbf{x}_{i1}$, as follows (see Lecture 7).

$$\mathbf{M}_{\mathbf{X}1}\mathbf{s} = \mathbf{M}_{\mathbf{X}1}\mathbf{Z}\boldsymbol{\pi}_0 + \mathbf{M}_{\mathbf{X}1}\boldsymbol{\eta}$$

In light of this analysis, it may appear that embarking into an empirical study based on Instrumental Variables is very risky, due to the high chance of ending up with mildly endogenous and fairly weak instruments. For the sake of mitigating this risk, it is best to follow some general guidelines.

1. It is always useful to **test** the **statistical power** of the instruments via estimates of the First Stage models (10.45). In the applied econometric practice, some rules of thumb apply: t-statistics for the exogenous instruments higher than 3, or model-wide F-statistics higher than 10, are considered signs that the instruments are "satisfactorily strong." These numbers appear to be based on simulation studies and surveys, see e.g. Stock et al. (2002); however, they must be taken with a grain of salt, since the conditions that make an instrument "strong enough" are really context- and data-dependent.

2. The earlier observation that 2SLS is more likely to hit the efficiency bound the more instruments are used must be revisited. While this is true in theory, in practice chances are that the more instruments one is employing, the higher the probability to include mildly endogenous, weak instruments – it is advisable to **drop instruments** from overidentified 2SLS estimators whenever they are suspected to be weak.

To summarize, IV-2SLS are indeed powerful "instruments," but they must be used with care. Any researcher employing Instrumental Variables should first make sure that the exogeneity assumptions can be credibly defended in the context at hand, and then show that the instruments are strong enough.

10.4 Estimation of Simultaneous Equations

The 2SLS estimator is not only a solution to the endogeneity problem in single-equation models, but is also the traditional tool for the estimation of linear simultaneous equations models. As it has been argued, the very usage of the word "endogeneity" originates in econometrics from the terminology of linear SEMs. This section describes how SEMs are estimated via 2SLS or, possibly better, via its extension which is specifically tailored to multiple equations: the so-called **Three-Stages Least Squares** (3SLS) estimator. The ensuing discussion builds on the analysis of the identification of SEMs which is developed in Lecture 9.

A **single equation** of a SEM that is identified can be easily estimated via 2SLS. To see this, rewrite the p-th equation of interest as:

$$y_{pi} = \mathbf{x}_{pi}^{\mathrm{T}} \boldsymbol{\beta}_{p0} + \varepsilon_{pi} \tag{10.76}$$

where the normalized endogenous variable y_{pi} is isolated on the left-hand side, while on the right-hand side, $\mathbf{x}_{pi}$ collects the realizations of all variables – **both exogenous and endogenous** – that are **not excluded** from the structural form; the appropriately **restricted** parameter set $\boldsymbol{\beta}_{p0}$ is defined accordingly. If (10.76) has $K_2 \leq P-1$ endogenous variables, it is associated to as many **first stage** models from the **reduced form** of the SEM:

$$x_{ki} = \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\pi}_{k0} + \eta_{ki} \tag{10.77}$$

for $k = 1, \ldots, K_2$. Whether the p-th equation (10.76) is exactly identified or overidentified, it can always be estimated by IV-2SLS (while the Indirect Least Squares approach cannot work under overidentification).

A problem with the 2SLS estimation of SEMs, if separately performed equation-by-equation, has to do to the estimation of the variance of $\boldsymbol{\beta}_{p0}$. In fact, separate estimation disregards any potential statistical variation that is common across equations. To put it more concretely, in the likely circumstance where the errors $(\varepsilon_{1i}, \ldots, \varepsilon_{Pi})$ are correlated across the P equations, 2SLS estimation is *inefficient*. An analogous problem arises when estimating a set of P linear regressions with no endogenous variables on the right hand side – basically, SEMs with $\boldsymbol{\Gamma} = \mathbf{I}$ – but with correlated error terms, models which are known as SURs (Seemingly Unrelated Regressions).

Example 10.4. Household labor supply. Consider the following model:

$$H_{hi} = \alpha_0 + \alpha_1 H_{wi} + \alpha_2 S_{hi} + \alpha_3 S_{wi} + \alpha_4 \log W_{hi} + \alpha_5 \log W_{wi} + \varepsilon_{hi}$$
$$H_{wi} = \beta_0 + \beta_1 H_{hi} + \beta_2 S_{hi} + \beta_3 S_{wi} + \beta_4 \log W_{hi} + \beta_5 \log W_{wi} + \varepsilon_{wi}$$

where subscript i denotes a household (the unit of observation), w denotes variables relative to the wife, h to the husband, and for $s \in \{h, w\}$, H_{si}, S_{si} and $\log W_{hi}$ denote "hours worked," education and the logarithm of the wage respectively. This model, which characterizes the interdependence of labor supply choices between the two members of the household, is identified with at least one restriction per equation (in particular, if $\alpha_1 = \beta_1 = 0$ this SEM becomes a SUR). If, for example, both the husband's and the wife's labor supply choices are not influenced by their partner's level of education S_{si}, one can impose the restrictions $\alpha_3 = \beta_2 = 0$ and separate 2SLS estimation of both $\boldsymbol{\alpha} = (\alpha_0, \alpha_2, \alpha_4, \alpha_5)^{\mathrm{T}}$ and $\boldsymbol{\beta} = (\beta_0, \beta_2, \beta_3, \beta_5)^{\mathrm{T}}$ is possible, by using the wife's education as an instrument of her labor supply – and vice versa. It is quite likely, however, that the members of a family experience shared external circumstances, such as for example the quality of social connections or individual and cultural attitudes about work. In this case:

$$\mathbb{C}\mathrm{ov}\left(\varepsilon_{hi}, \varepsilon_{wi}\right) \neq 0$$

therefore, separate 2SLS estimation of $\widehat{\boldsymbol{\alpha}}_{2SLS}$ and $\widehat{\boldsymbol{\beta}}_{2SLS}$ would be inefficient: intuitively, statistical inferences would not take into account that both sets of point estimates are generated by some common statistical variation. ■

Thus, methods for the joint estimation of SEMs (or SURs) have been developed. These methods are known as **full information** approaches, which contrast with equation-by-equation **limited information** approaches (such as the equation-by-equation 2SLS). The most straightforward full information method is the so-called "Three Stages Least Squares" (3SLS) estimator, which extends 2SLS by adding a further "third stage" meant to obtain more efficient estimates. This extended procedure aims to address cross-equation error correlation, and it is analogous to a GLS approach to correct for heteroscedasticity or error dependence in single equation models.

The 3SLS estimator is best described using compact matrix notation. Write any SEM whose equations are all at least exactly identified as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon} \tag{10.78}$$

or:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_P \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}_P \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{10} \\ \boldsymbol{\beta}_{20} \\ \vdots \\ \boldsymbol{\beta}_{P0} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_P \end{bmatrix} \tag{10.79}$$

where $\mathbf{y}_p$ is the N-dimensional vector obtained from stacking all the observations y_{pi} for $i = 1, \ldots, N$, while $\boldsymbol{\varepsilon}_p$ is constructed analogously. Similarly, matrix $\mathbf{X}_p$ results from vertically stacking vectors $\mathbf{x}_{pi}^{\mathrm{T}}$ for $i = 1, \ldots, N$; thus (10.80) can be rephrased as follows.

$$\mathbf{y}_p = \mathbf{X}\boldsymbol{\beta}_{p0} + \boldsymbol{\varepsilon}_p \tag{10.80}$$

Furthermore, consider the stacked instruments matrix $\mathbf{Z}$ as in (10.30); the associated projection matrix $\mathbf{P}_{\mathbf{Z}}$, and construct the P equation-specific matrices of projected regressors as:

$$\widehat{\mathbf{X}}_p = \mathbf{P}_{\mathbf{Z}}\mathbf{X}_p$$

while

$$\widehat{\mathbf{X}} \equiv \begin{bmatrix} \widehat{\mathbf{X}}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{X}}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \widehat{\mathbf{X}}_P \end{bmatrix}$$

results from diagonally stacking, block-by-block, all the $\widehat{\mathbf{X}}_p$ matrices. Given this notation, the 2SLS estimator for **all** P equations can be written more compactly as follows.

$$\widehat{\boldsymbol{\beta}}_{2SLS} = \left(\widehat{\mathbf{X}}^{\mathrm{T}}\widehat{\mathbf{X}}\right)^{-1}\widehat{\mathbf{X}}^{\mathrm{T}}\mathbf{y} \tag{10.81}$$

Assume for simplicity that the error terms $\boldsymbol{\varepsilon}$ are i.i.d. *within equations*, but are correlated *across equations*:

$$\mathbb{E}\left[\boldsymbol{\varepsilon} \middle| \boldsymbol{Z}\right] = 0$$
$$\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}} \middle| \boldsymbol{Z}\right] = \boldsymbol{\Sigma} \otimes \mathbf{I}$$

where $\boldsymbol{\Sigma}$ is the symmetric $P \times P$ matrix containing the equation-specific variance of each equation along the diagonal, and the cross-equation covariance terms outside the diagonal.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1P} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{P1} & \sigma_{P2} & \cdots & \sigma_{PP} \end{bmatrix} \tag{10.82}$$

Therefore, by the definition of Kronecker product:

$$\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}} \middle| \boldsymbol{Z}\right] = \boldsymbol{\Sigma} \otimes \mathbf{I} = \begin{bmatrix} \sigma_{11}\mathbf{I} & \sigma_{12}\mathbf{I} & \cdots & \sigma_{1P}\mathbf{I} \\ \sigma_{21}\mathbf{I} & \sigma_{22}\mathbf{I} & \cdots & \sigma_{2P}\mathbf{I} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{P1}\mathbf{I} & \sigma_{P2}\mathbf{I} & \cdots & \sigma_{PP}\mathbf{I} \end{bmatrix}$$

meaning that, in addition to the equation-specific variance terms σ_{pp} along the diagonal, for any observation i ($i = 1, \ldots, N$) the conditional covariance of the two shocks ε_{pi} and ε_{qi} for the p-th and the q-th equations respectively is equal to σ_{pq}. All the other cross-equation covariance terms, that is those for any two different observations, are assumed to be equal to zero.

Thus, the 3SLS estimator can be easily defined as just the GLS generalization of the 2SLS estimator for this model. It is computed by iterating the following "three stages" (the first two of which corresponding to the two stages of the 2SLS estimator):

1. project $\mathbf{x}_{pi}$ ($\mathbf{X}_p$) onto $\mathbf{z}_i$ ($\mathbf{Z}$) equation-by-equation for $p = 1, \ldots, P$;

2. compute the 2SLS estimator $\widehat{\boldsymbol{\beta}}_{2SLS}$ from (10.81), as well as estimates for the $P(P-1)$ parameters contained in matrix $\boldsymbol{\Sigma}$:

$$\widehat{\sigma}_{pq} = \frac{\left(\mathbf{y}_p - \mathbf{X}_p \widehat{\boldsymbol{\beta}}_{p2SLS}\right)^{\mathrm{T}} \left(\mathbf{y}_q - \mathbf{X}_q \widehat{\boldsymbol{\beta}}_{q2SLS}\right)}{N} \tag{10.83}$$

   for $p, q = 1, \ldots, P$, resulting in an estimate $\widehat{\boldsymbol{\Sigma}}_N$ of matrix $\boldsymbol{\Sigma}$;

3. finally, compute the 3SLS estimator as:

$$\widehat{\boldsymbol{\beta}}_{3SLS} = \left[\widehat{\mathbf{X}}^{\mathrm{T}} \left(\widehat{\boldsymbol{\Sigma}}_N^{-1} \otimes \mathbf{I}\right) \widehat{\mathbf{X}}\right]^{-1} \widehat{\mathbf{X}}^{\mathrm{T}} \left(\widehat{\boldsymbol{\Sigma}}_N^{-1} \otimes \mathbf{I}\right) \mathbf{y} \tag{10.84}$$

   and its asymptotic variance as follows – compare it with (8.40).

$$\widehat{\mathbb{A}\mathrm{var}}\left(\widehat{\boldsymbol{\beta}}_{3SLS}\right) = \left[\widehat{\mathbf{X}}^{\mathrm{T}} \left(\widehat{\boldsymbol{\Sigma}}_N^{-1} \otimes \mathbf{I}\right) \widehat{\mathbf{X}}\right]^{-1} \tag{10.85}$$

As hinted later in Lecture 12, there exist versions of the 3SLS estimator that are robust to heteroscedasticity and to wider forms of error dependence.

The 3SLS estimator is the most efficient among all the semi-parametric estimators of SEMs. Just like the 2SLS estimator, as it is discussed later, it corresponds to the solution of a Generalized Method of Moments (GMM) problem. Nonetheless, in the fully parametric case other methods are available for the estimation of SEMs (the so-called **LIML**, Limited Information Maximum Likelihood, and **FIML**, Full Information Maximum Likelihood methods). These maximum likelihood methods however, are *not* more efficient than GMM-based or otherwise semi-parametric methods, and in addition they are liable to violations of the parametric assumptions. For this reason, the current practice favors the use of semi-parametric methods for the estimation of linear simultaneous equations.

Lecture 11

Maximum Estimation

This lecture illustrates the general estimation framework that encompasses the three most common estimation methods in econometrics: Least Squares methods (which include OLS, its generalizations, as well as their non-linear versions), Maximum Likelihood Estimation, and the Generalized Method of Moments. This framework is usually referred to as "Maximum Estimation" (in short, "M-Estimation") or as "Extremum Estimation." All M-Estimators are based on the optimization of some objective function, hence their name. This lecture develops the theoretical and statistical framework that is common to all M-Estimators; whole doing so, it introduces the more specialized Non-Linear Least Squares (NLLS) and especially the Maximum Likelihood Estimation (MLE) framework, along with illustrative applied examples.

11.1 Criterion Functions

Consider a structural model such as (9.1). Suppose that given some specific assumptions about the joint probability distribution of $(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{\varepsilon}_i)$ (either fully parametric or semi-parametric), the model's **true** parametric structure $\boldsymbol{\theta}_0$ *in the population* is the solution to a maximization problem of the kind:

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max}\, \mathcal{Q}_0(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max} \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[q(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta})\right] \tag{11.1}$$

where $q(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta})$ is an observation-specific **criterion** for $i = 1, \ldots, N$. For the sake of a simpler notation, this function is heretofore written as $q(\boldsymbol{x}_i; \boldsymbol{\theta})$, where $\boldsymbol{x}_i = (\boldsymbol{y}_i, \boldsymbol{z}_i)$. Observe that with *identically distributed observations*, it is $\mathcal{Q}_0(\boldsymbol{\theta}) = \mathbb{E}\left[q(\boldsymbol{x}_i; \boldsymbol{\theta})\right]$. To anticipate the discussion about identification, notice how the definition in (11.1) implies that *at the limit*, function $\mathcal{Q}_0(\boldsymbol{\theta})$ should have a unique maximum, or else $\boldsymbol{\theta}_0$ would be ill-defined.

An appropriate estimator of $\boldsymbol{\theta}_0$ is the maximizer of the sample version of (11.1), which is known as the **sample criterion function** $\widehat{\mathcal{Q}}_N(\boldsymbol{\theta})$:

$$\widehat{\boldsymbol{\theta}}_M = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max}\, \widehat{\mathcal{Q}}_N(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max}\, \frac{1}{N}\sum_{i=1}^{N} q(\mathbf{x}_i;\boldsymbol{\theta}) \tag{11.2}$$

such an object is called an **M-Estimator**. The analysis of **M-Estimation** that follows concerns the identification conditions of M-Estimators and their asymptotic properties. The full-fledged analysis is provided by Newey and McFadden (1994) in a chapter of the *Handbook of Econometrics*.

Example 11.1. Ordinary Least Squares (OLS). Let the CEF of some endogenous variable Y_i, given some K exogenous variables $\boldsymbol{x}_i$, be linear.

$$\mathbb{E}[Y_i|\,\boldsymbol{x}_i] = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 \tag{11.3}$$

In such a case, the "true" parameter vector $\boldsymbol{\beta}_0$ is shown by an extension of Theorem 7.1, to minimize the "limiting" mean squared error (MSE):

$$\boldsymbol{\beta}_0 = \underset{\boldsymbol{\beta}\in\mathbb{R}^K}{\arg\max} \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} -\mathbb{E}\left[\left(Y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2\right] \tag{11.4}$$

thus the OLS estimator is just the minimizer of the sample analog:

$$\widehat{\boldsymbol{\beta}}_{OLS} = \underset{\boldsymbol{\beta}\in\mathbb{R}^K}{\arg\max} -\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 \tag{11.5}$$

corresponding to an M-Estimator for $q(Y_i,\boldsymbol{x}_i;\boldsymbol{\beta}) = -\left(Y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2$. ■

Example 11.2. Non-Linear Least Squares (NLLS). If, instead, the CEF is non-linear, governed by some function denoted as $h(\boldsymbol{x}_i;\boldsymbol{\theta})$:

$$\mathbb{E}[Y_i|\,\boldsymbol{x}_i] = h(\boldsymbol{x}_i;\boldsymbol{\theta}_0) \tag{11.6}$$

one can show again by extending Theorem 7.1 that:

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max} \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} -\mathbb{E}\left\{\left[Y_i - h(\boldsymbol{x}_i;\boldsymbol{\theta})\right]^2\right\} \tag{11.7}$$

and the **Non-Linear Least Squares** (NLLS) estimator is defined as the sample analog of the above problem, so long as $h(\boldsymbol{x}_i;\boldsymbol{\theta})$ is invertible in $\boldsymbol{\theta}$:

$$\widehat{\boldsymbol{\theta}}_{NLLS} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max} -\frac{1}{N}\sum_{i=1}^{N}\left[y_i - h(\mathbf{x}_i;\boldsymbol{\theta})\right]^2 \tag{11.8}$$

where here $q(Y_i,\boldsymbol{x}_i;\boldsymbol{\theta}) = -\left[Y_i - h(\boldsymbol{x}_i;\boldsymbol{\theta})\right]^2$. In typical applications, (11.8) lacks an explicit solution; estimates must be thus obtained numerically. ■

For the present discussion, it is useful to define the following two objects. Borrowing from Maximum Likelihood terminology, the observation-specific **score** $\mathbf{s}_i(\mathbf{x}_i;\boldsymbol{\theta})$ for $i=1,\ldots,N$ of an M-estimator is defined as the vector of first derivatives of $q(\mathbf{x}_i;\boldsymbol{\theta})$, with respect to the parameter set $\boldsymbol{\theta}$:

$$\mathbf{s}_i(\mathbf{x}_i;\boldsymbol{\theta}) = \frac{\partial q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial \theta_1} \\ \frac{\partial q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial \theta_2} \\ \vdots \\ \frac{\partial q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial \theta_K} \end{bmatrix}$$

while $\mathbf{H}_i(\mathbf{x}_i;\boldsymbol{\theta})$ is the $K \times K$ **Hessian matrix** of the second derivatives of $q(\mathbf{x}_i;\boldsymbol{\theta})$, or – equivalently – of the score's first derivatives (the Jacobian matrix), with respect to the parameter set $\boldsymbol{\theta}$:

$$\mathbf{H}_i(\mathbf{x}_i;\boldsymbol{\theta}) = \frac{\partial \mathbf{s}_i(\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}} = \frac{\partial^2 q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^{\mathrm{T}}} = \begin{bmatrix} \frac{\partial^2 q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_1} & \frac{\partial^2 q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_K} \\ \frac{\partial^2 q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial^2 \theta_2 \partial \theta_1} & \frac{\partial^2 q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_2} & \cdots & \frac{\partial^2 q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial \theta_K \partial \theta_1} & \frac{\partial^2 q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial \theta_K \partial \theta_2} & \cdots & \frac{\partial^2 q(\mathbf{x}_i;\boldsymbol{\theta})}{\partial \theta_K \partial \theta_K} \end{bmatrix}$$

again for $i=1,\ldots,N$. Notice that both the score vector and the Hessian matrix only exist under certain conditions, specifically if the M-Estimation objective function is, respectively, at least once or twice continuously differentiable. These conditions might not be respected for the objective function of some important econometric estimators, such as the quantile regression. The score and the Hessian matrix are instrumental for the characterization of the **identification** conditions for M-Estimators.

Theorem 11.1. Identification of M-Estimators. *In any M-Estimation environment, the "true" parameter set $\boldsymbol{\theta}_0$ is locally point identified if the following limiting average Hessian matrix evaluated at $\boldsymbol{\theta}_0$ has full K rank.*

$$\mathbf{Q}_0 \equiv \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\mathbf{H}_i(\boldsymbol{x}_i;\boldsymbol{\theta}_0)\right]$$

Proof. This is indeed a simple application of the Implicit Function Theorem. In a well-defined M-Estimator, the true parameter vector $\boldsymbol{\theta}_0$ sets the K First Order Conditions of the empirical criterion function $\widehat{\mathcal{Q}}_N(\boldsymbol{\theta}_0)$ equal to zero *at the probability limit*, that is, at some *limiting average score*.

$$\frac{\partial}{\partial \boldsymbol{\theta}} \widehat{\mathcal{Q}}_N(\boldsymbol{\theta}_0) = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{\partial}{\partial \boldsymbol{\theta}} q(\mathbf{x}_i;\boldsymbol{\theta}_0)\right] \xrightarrow{p} \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\mathbf{s}_i(\boldsymbol{x}_i;\boldsymbol{\theta}_0)\right] = \mathbf{0}$$

If the Jacobian $\mathbf{Q}_0$ has full rank, there is a unique local solution $\boldsymbol{\theta}_0$. □

Example 11.3. Identification of OLS. In OLS, the observation-specific score and the Hessian matrix are as follows.

$$\begin{aligned}\mathbf{s}_i\left(y_i,\mathbf{x}_i;\boldsymbol{\beta}\right)&=2\mathbf{x}_i\varepsilon_i\\ \mathbf{H}_i\left(y_i,\mathbf{x}_i;\boldsymbol{\beta}\right)&=-2\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\end{aligned}$$

Consequently, the limiting average Hessian of OLS is just:

$$\mathbf{Q}_0=-2\mathbf{K}_0=\lim_{N\to\infty}-\frac{2}{N}\sum_{i=1}^{N}\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right]$$

and the requirement that the above matrix must have full rank in order for the OLS estimator to be identified is quite a familiar condition. ■

Example 11.4. Identification of NLLS. In the NLLS case, by denoting the *error term* by $\varepsilon_i\equiv y_i-h\left(\mathbf{x}_i;\boldsymbol{\theta}\right)$, the score and the Hessian are:

$$\begin{aligned}\mathbf{s}_i\left(y_i,\mathbf{x}_i;\boldsymbol{\theta}\right)&=2\frac{\partial}{\partial\boldsymbol{\theta}}h\left(\mathbf{x}_i;\boldsymbol{\theta}\right)\cdot\varepsilon_i\\ \mathbf{H}_i\left(y_i,\mathbf{x}_i;\boldsymbol{\theta}\right)&=-2\frac{\partial}{\partial\boldsymbol{\theta}}h\left(\mathbf{x}_i;\boldsymbol{\theta}\right)\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}h\left(\mathbf{x}_i;\boldsymbol{\theta}\right)+2\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}h\left(\mathbf{x}_i;\boldsymbol{\theta}\right)\cdot\varepsilon_i\end{aligned}$$

and note that, by the Law of Iterated Expectations:

$$\mathbb{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}h\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\cdot\varepsilon_i\right]=\mathbb{E}_{\boldsymbol{x}}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}h\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\cdot\mathbb{E}\left[\varepsilon_i|\,\boldsymbol{x}_i\right]\right]=\mathbf{0}$$

as $\mathbb{E}\left[\varepsilon_i|\,\boldsymbol{x}_i\right]=\mathbb{E}\left[Y_i|\,\boldsymbol{x}_i\right]-h\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)=0$. Thus, any expected Hessian matrix of NLLS is just:

$$\mathbb{E}\left[\mathbf{H}_i\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right]=-2\,\mathbb{E}\left[\boldsymbol{h}_{0i}\boldsymbol{h}_{0i}^{\mathrm{T}}\right]$$

where:

$$\boldsymbol{h}_{0i}\equiv\frac{\partial}{\partial\boldsymbol{\theta}}h\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)$$

is the derivative of the CEF evaluated at $\mathbf{x}_i$ and at the true parameters $\boldsymbol{\theta}_0$. The identification of NLLS is generally evaluated in terms of the matrices $\mathbb{E}\left[\boldsymbol{h}_{0i}\boldsymbol{h}_{0i}^{\mathrm{T}}\right]$, and the probability limits of their sample averages. ■

In practical applications, it is customary to verify that the *sample mean* of the Hessian (like $N^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}$ in the OLS case) has full rank, as an indication that the model is identified. In addition, it is useful to check that the rows or columns of the Hessian's sample mean are not *too correlated*; otherwise, identification is said to be *weak*, and the estimates are usually very imprecise with large standard errors. This problem is called **quasi-multicollinearity**

and is intuitively due to the statistical difficulty of distinguishing between two "factors" (like different explanatory variables, columns in $\mathbf{X}$) if they are very similar. In the IV/2SLS case, this corresponds to the problem of *weak instruments* which appears if the inverse of $\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{X}$ is too large.

A relevant subclass of M-Estimation is constituted by the **Maximum Likelihood Estimation** (MLE) framework. The general approach to MLE in Statistics is introduced in Lecture 5; in econometrics, this framework is utilized to construct estimators for **fully parametric structural models**. To substantiate, suppose that a structural model reads, for every unit of observation $i = 1, \ldots, N$, as in (9.1): $\boldsymbol{y}_i = \boldsymbol{s}(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{\varepsilon}_i; \boldsymbol{\theta})$. If an analyst either knows *a priori* or can confidently assume the joint probability distribution $F_{\boldsymbol{z},\boldsymbol{\varepsilon}}(\boldsymbol{z}_i, \boldsymbol{\varepsilon}_i)$ of the exogenous variables (observable and unobservable), it is possible to characterize the distribution of $\boldsymbol{x}_i = (\boldsymbol{y}_i, \boldsymbol{z}_i)$ as:

$$f_{\boldsymbol{z},\boldsymbol{\varepsilon}}(\boldsymbol{z}_i, \boldsymbol{\varepsilon}_i; \boldsymbol{\theta}) = f_{\boldsymbol{z},\boldsymbol{\varepsilon}}\left(\boldsymbol{z}_i; \boldsymbol{s}_{\boldsymbol{\varepsilon}}^{-1}(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta}); \boldsymbol{\theta}\right)$$

where here $f_{\boldsymbol{z},\boldsymbol{\varepsilon}}(\cdot)$ is the probability mass or density function associated with $F_{\boldsymbol{z},\boldsymbol{\varepsilon}}(\cdot)$, whereas $\boldsymbol{s}_{\boldsymbol{\varepsilon}}^{-1}(\cdot)$ is the solution of the structural relationship with respect to the unobservable factors $\boldsymbol{\varepsilon}_i$ (assuming that such a unique inverse exists). In this environment, MLE corresponds with the M-Estimator that is defined for a criterion function equaling the logarithm of $f_{\boldsymbol{z},\boldsymbol{\varepsilon}}(\cdot)$; write this function succintly as $\ell(\boldsymbol{x}_i; \boldsymbol{\theta})$ where again $\boldsymbol{x}_i = (\boldsymbol{y}_i, \boldsymbol{z}_i)$.

$$\begin{aligned} q(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta}) &= \log f_{\boldsymbol{z},\boldsymbol{\varepsilon}}\left(\boldsymbol{z}_i; \boldsymbol{s}_{\boldsymbol{\varepsilon}}^{-1}(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta}); \boldsymbol{\theta}\right) \\ &\equiv \ell(\boldsymbol{x}_i; \boldsymbol{\theta}) \end{aligned}$$

This characterization perfectly complies with the definition of M-Estimators since the true value of the parameters $\boldsymbol{\theta}_0$, by definition, satisfies:

$$\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\ell(\boldsymbol{x}_i; \boldsymbol{\theta})\right] \tag{11.9}$$

which is a natural consequence of choosing this particular criterion function. In fact, for $i = 1, \ldots, N$ one can maintain the following relationships, thanks to an application of Jensen's inequality (second line below):

$$\begin{aligned} \mathbb{E}\left[\ell(\boldsymbol{x}_i; \boldsymbol{\theta})\right] - \mathbb{E}\left[\ell(\boldsymbol{x}_i; \boldsymbol{\theta}_0)\right] &= \mathbb{E}\left[\log \frac{f_{\boldsymbol{x}}(\boldsymbol{x}_i; \boldsymbol{\theta})}{f_{\boldsymbol{x}}(\boldsymbol{x}_i; \boldsymbol{\theta}_0)}\right] \\ &\leq \log \mathbb{E}\left[\frac{f_{\boldsymbol{x}}(\boldsymbol{x}_i; \boldsymbol{\theta})}{f_{\boldsymbol{x}}(\boldsymbol{x}_i; \boldsymbol{\theta}_0)}\right] \\ &= \log \int_{\mathbb{X}} \frac{f_{\boldsymbol{x}}(\mathbf{x}_i; \boldsymbol{\theta})}{f_{\boldsymbol{x}}(\mathbf{x}_i; \boldsymbol{\theta}_0)} f_{\boldsymbol{x}}(\mathbf{x}_i; \boldsymbol{\theta}_0)\, \mathrm{d}\mathbf{x}_i \\ &= 0 \end{aligned}$$

where $\mathbb{X}$ is the joint support of $\boldsymbol{x}_i = (\boldsymbol{y}_i, \boldsymbol{z}_i)$, and $\mathrm{d}\mathbf{x}_i$ is the joint differential. Note that the expectation must be evaluated by integrating over $f_{\boldsymbol{x}}(\boldsymbol{x}_i; \boldsymbol{\theta}_0)$, as this is the function which is assumed to generate the data. These relationships hold for any for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$; hence over the entire parameter space it is $\mathbb{E}[\ell(\boldsymbol{x}_i; \boldsymbol{\theta}_0)] \geq \mathbb{E}[\ell(\boldsymbol{x}_i; \boldsymbol{\theta})]$ for $i = 1, \ldots, N$ and (11.9) must hold.

A variant of this approach which is equally valid, and at the same time typically more practical, is that where a researcher only specifies the *conditional* distribution of the unobserved factors $\boldsymbol{\varepsilon}_i$, *given* the realizations $\mathbf{y}_i$ of the exogenous variables: thus, the criterion function is specified as follows.

$$q(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta}) = \log f_{\boldsymbol{z},\boldsymbol{\varepsilon}}\left(\boldsymbol{s}_{\boldsymbol{\varepsilon}}^{-1}(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta}) \middle| \boldsymbol{z}_i\right)$$

This method is called **Conditional Maximum Likelihood Estimation** (CMLE) and in econometrics it prevails over its unconditional version; it is best illustrated via an example.

Example 11.5. Maximum Likelihood and Linear Regression. Consider a linear regression model like $y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i$. Suppose there is reason to believe that the error term of this model is homoscedastic, independent across observations and **normally distributed**:

$$\boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$$

where σ^2 is an unknown parameter of the model. In addition, assume that the right-hand side variables $\boldsymbol{x}_i$ are "fixed" (or *non-stochastic*): that is, one realization $\mathbf{x}_i$ appears with probability one in every sample. Together, these assumptions specify the entire joint probability distribution of $(\boldsymbol{x}_i, \varepsilon_i)$, characterizing the so-called "classical" regression model with spherical/normal disturbances. Notice that one implication is that the covariance between the error term and the regressors is zero, that is $\mathbb{E}[\boldsymbol{x}_i\varepsilon_i] = \mathbf{0}$. The probability density function for the error term can be then written as:

$$f_\varepsilon\left(\varepsilon_i \middle| \boldsymbol{\beta}, \sigma^2\right) = f_\varepsilon\left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} \middle| \boldsymbol{\beta}, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2}{2\sigma^2}\right)$$

which is just the expression of a univariate normal distribution after having applied a change of variable. Write the collection of the unknown parameters as $\boldsymbol{\theta} = (\boldsymbol{\beta}; \sigma^2)$. In this model, the likelihood function is:

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{\theta} \middle| \{y_i, \mathbf{x}_i\}_{i=1}^N\right) &= \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} \exp\left(-\frac{\sum_{i=1}^N \left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2}{2\sigma^2}\right)\end{aligned}$$

and correspondingly, the log-likelihood function is:

$$\log \mathcal{L}\left(\boldsymbol{\theta}|\left\{y_i, \mathbf{x}_i\right\}_{i=1}^N\right) = -\frac{N}{2}\left(\log 2\pi + \log \sigma^2\right) - \frac{\sum_{i=1}^N \left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2}{2\sigma^2}$$

which has clearly a unique maximum (the Maximum Likelihood Estimator of $\boldsymbol{\theta}$ is well defined). In fact, the First Order Conditions are:

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log \mathcal{L}\left(\boldsymbol{\theta}|\left\{y_i, \mathbf{x}_i\right\}_{i=1}^N\right) = \begin{bmatrix} \sum_{i=1}^N \mathbf{x}_i \left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right) \sigma^{-2} \\ -\frac{N}{2}\sigma^{-2} + \frac{1}{2}\sum_{i=1}^N \left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 \sigma^{-4} \end{bmatrix} = \mathbf{0}$$

whose solution is the ML estimator $\widehat{\boldsymbol{\theta}}_{MLE} = \left(\widehat{\boldsymbol{\beta}}_{MLE}; \widehat{\sigma}^2_{MLE}\right)$ given by:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{MLE} &= \left(\sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\right)^{-1} \sum_{i=1}^N \mathbf{x}_i y_i \\ \widehat{\sigma}^2_{MLE} &= \frac{\sum_{i=1}^N \left(y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{MLE}\right)^2}{N} \end{aligned}$$

and the Second Order Conditions are satisfied for a maximum. Therefore, this estimator exists and is unique as long as matrix $\sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}$ has full rank. Note that the Maximum Likelihood estimator of $\boldsymbol{\beta}$ is identical to the corresponding OLS estimator of the linear regression model. The estimator for the error variance parameter $\widehat{\sigma}^2_{MLE}$ differs, however, from the unbiased estimator $\widehat{\sigma}^2$ from the small sample analysis of OLS, as the latter is larger by the factor $\frac{N}{N-K}$. This is just one particular example of a general feature of MLE: while this approach might produced biased estimators, these are in general consistent and at least as efficient as their unbiased counterparts.

This Maximum Likelihood estimator can alternatively be obtained under more general assumptions. Suppose that $\boldsymbol{x}_i$ is not fixed; without specifying its full data generation process, assume that *conditional on any realization* $\mathbf{x}_i$, the error term is normal with constant variance: $\varepsilon_i|\,\mathbf{x}_i \sim \mathcal{N}\left(0, \sigma^2\right)$. Since $\varepsilon_i = y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$, this implies the following conditional density function:

$$f_{Y|\boldsymbol{x}}\left(y_i|\,\mathbf{x}_i; \boldsymbol{\beta}, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2}{2\sigma^2}\right)$$

with the same associated likelihood function as above. This shows that the CMLE approach here delivers the same result as the simple (but unrealistic) assumption that $\boldsymbol{x}_i$ is fixed. In fact, $\boldsymbol{x}_i$ is allowed to follow any distribution, so long as ε_i is normal when conditioning on it. ■

11.2 Asymptotics of Maximum Estimators

In their leading article, Newey and McFadden (1994) establish conditions for consistency and asymptotic normality of all M-Estimators; their results are summarized here. In order to prove consistency, it is necessary to show that the maximizer of the sample average criterion function converges in probability to the *unique* maximizer of the *population expected* criterion.

$$\widehat{\boldsymbol{\theta}}_M = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \widehat{\mathcal{Q}}_N(\boldsymbol{\theta}) \xrightarrow{p} \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{Q}_0(\boldsymbol{\theta}) = \boldsymbol{\theta}_0 \tag{11.10}$$

Intuitively, **pointwise convergence** of $\widehat{\mathcal{Q}}_N(\boldsymbol{\theta})$ to $\mathcal{Q}_0(\boldsymbol{\theta})$ for all the possible values $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is necessary:

$$\left| \frac{1}{N} \sum_{i=1}^{N} \{q(\mathbf{x}_i; \boldsymbol{\theta}) - \mathbb{E}[q(\boldsymbol{x}_i; \boldsymbol{\theta})]\} \right| \xrightarrow{p} 0 \tag{11.11}$$

still, it is not sufficient. A sufficient condition is **uniform convergence**:

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| \frac{1}{N} \sum_{i=1}^{N} \{q(\mathbf{x}_i; \boldsymbol{\theta}) - \mathbb{E}[q(\boldsymbol{x}_i; \boldsymbol{\theta})]\} \right| \xrightarrow{p} 0 \tag{11.12}$$

this condition is stronger than (11.11) as it requires that $\widehat{\mathcal{Q}}_N(\boldsymbol{\theta})$ converges in probability towards $\mathcal{Q}_0(\boldsymbol{\theta})$ "at the same speed" over the entire parameter space $\boldsymbol{\Theta}$. The intuition is graphically represented in Figure 11.1.

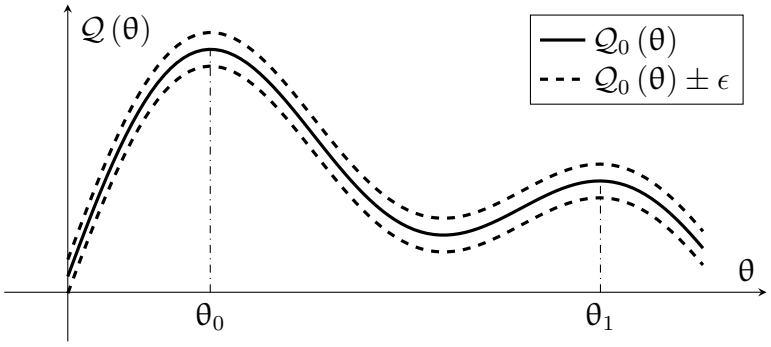


Figure 11.1: Uniform convergence: intuition (for $\boldsymbol{\Theta} = \mathbb{R}_+$)

For (11.12) to hold it is necessary that $\widehat{\mathcal{Q}}_N(\boldsymbol{\theta}) \in [\mathcal{Q}_0(\boldsymbol{\theta}) - \epsilon, \mathcal{Q}_0(\boldsymbol{\theta}) + \epsilon]$ for any small $\epsilon > 0$ and *for all parameter values* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$: the sample criterion must remain confined within a "sleeve" of the population expected criterion over the entire parameter space $\boldsymbol{\Theta}$, like the area represented in Figure 11.1.

If, in this example, the sampling error increases at higher values of θ, the local maximum θ_1 could be mistaken for the actual global maximum θ_0.

Uniform convergence is ensured if these four conditions hold: *i.* $q(\mathbf{x}_i; \boldsymbol{\theta})$ is continuous; *ii.* $\boldsymbol{\Theta}$ is a compact set; *iii.* $\mathbb{E}[|q(\boldsymbol{x}_i; \boldsymbol{\theta})|] < \infty$: that is, $q(\mathbf{x}_i; \boldsymbol{\theta})$ has a bounded first absolute moment; *iv.* $q(\mathbf{x}_i; \boldsymbol{\theta})$ is Borel-measureable on its support. These conditions, together, allow to invoke a result known as **Uniform Weak Law of Large Numbers**, implying uniform convergence. These conditions are technical; notice, however, that *i.* and *ii.* relate to the fact that M-Estimators are, in fact, maximum points; while *iii.* and *iv.* are analogous to similar conditions from other Laws of Large Numbers. Armed with the notion of uniform convergence, one can replicate the original proof of M-Estimators' consistency as it was given by Newey and McFadden (their Theorem 2.1, which is reported here with minor variations).

Theorem 11.2. Consistency of M-Estimators. *If i.* $\mathcal{Q}_0(\boldsymbol{\theta})$ *is uniquely maximized at* $\boldsymbol{\theta}_0$*, ii.* $\boldsymbol{\Theta}$ *is a compact set, iii.* $\mathcal{Q}_0(\boldsymbol{\theta})$ *is a continuous function, and iv.* $\widehat{\mathcal{Q}}_N(\boldsymbol{\theta})$ *uniformly converges in probability to* $\mathcal{Q}_0(\boldsymbol{\theta})$*, then it follows that M-Estimators are consistent as per (11.10).*

Proof. For any $\epsilon > 0$, with probability approaching 1 (w.p.a. 1);

$$\text{by } i.: \quad \widehat{\mathcal{Q}}_N\left(\widehat{\boldsymbol{\theta}}_M\right) > \widehat{\mathcal{Q}}_N(\boldsymbol{\theta}_0) - \frac{\epsilon}{3} \tag{a}$$

$$\text{by } iv.: \quad \mathcal{Q}_0\left(\widehat{\boldsymbol{\theta}}_M\right) > \widehat{\mathcal{Q}}_N\left(\widehat{\boldsymbol{\theta}}_M\right) - \frac{\epsilon}{3} \tag{b}$$

$$\text{by } iv.: \quad \widehat{\mathcal{Q}}_N(\boldsymbol{\theta}_0) > \mathcal{Q}_0(\boldsymbol{\theta}_0) - \frac{\epsilon}{3} \tag{c}$$

therefore, w.p.a. 1:

$$\mathcal{Q}_0\left(\widehat{\boldsymbol{\theta}}_M\right) \overset{(b)}{>} \widehat{\mathcal{Q}}_N\left(\widehat{\boldsymbol{\theta}}_M\right) - \frac{\epsilon}{3} \overset{(a)}{>} \widehat{\mathcal{Q}}_N(\boldsymbol{\theta}_0) - \frac{2\epsilon}{3} \overset{(c)}{>} \mathcal{Q}_0(\boldsymbol{\theta}_0) - \epsilon$$

hence, $\mathcal{Q}_0\left(\widehat{\boldsymbol{\theta}}_M\right) > \mathcal{Q}_0(\boldsymbol{\theta}_0) - \epsilon$ w.p.a. 1. Now, denote by $\mathbb{U}$ any given open neighborhood of $\boldsymbol{\theta}_0$ and by $\mathbb{U}^c$ its complement in $\boldsymbol{\Theta}$. Also define,

$$\mathcal{Q}_0(\boldsymbol{\theta}^*) = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta} \cap \mathbb{U}^c} \mathcal{Q}_0(\boldsymbol{\theta})$$

for some $\boldsymbol{\theta}^*$, and notice that $\mathcal{Q}_0(\boldsymbol{\theta}^*) < \mathcal{Q}_0(\boldsymbol{\theta}_0)$ by *i.*-*ii.*-*iii.*: thus, by setting:

$$\epsilon = \mathcal{Q}_0(\boldsymbol{\theta}_0) - \mathcal{Q}_0(\boldsymbol{\theta}^*)$$

it follows that:

$$\mathcal{Q}_0\left(\widehat{\boldsymbol{\theta}}_M\right) > \mathcal{Q}_0(\boldsymbol{\theta}_0) - \epsilon \Rightarrow \mathcal{Q}_0\left(\widehat{\boldsymbol{\theta}}_M\right) > \mathcal{Q}_0(\boldsymbol{\theta}^*)$$

implying that $\widehat{\boldsymbol{\theta}}_M \in \mathbb{U}$ for any open neighborhood $\mathbb{U}$. Thus, $\widehat{\boldsymbol{\theta}}_M \xrightarrow{p} \boldsymbol{\theta}_0$. □

The result about asymptotic normality, which is instrumental for statistical inference, is derived in a perhaps more familiar way.

Theorem 11.3. Asymptotic Normality of M-Estimators. *A generic M-Estimator $\widehat{\boldsymbol{\theta}}_M$ follows an asymptotically normal distribution if the following five conditions hold simultaneously:*

i. $\widehat{\boldsymbol{\theta}}_M$ *is a consistent estimator of* $\boldsymbol{\theta}_0$*;*

ii. $q(\mathbf{x}_i; \boldsymbol{\theta})$ *is a concave and twice continuously differentiable function in an open neighborhood of* $\boldsymbol{\theta}_0$*;*

iii. $\frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E}[q(\boldsymbol{x}_i; \boldsymbol{\theta})] = \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} q(\boldsymbol{x}_i; \boldsymbol{\theta})\right]$*: the derivative can pass through the expectation integral;*

iv. the data meet the requirements for the application of a Central Limit Theorem (the data are "well behaved");

v. the Hessian matrix is nonsingular, it is continuous in $\boldsymbol{\theta}$ *and it has a bounded absolute first moment.*

The limiting distribution is:

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_M - \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{Q}_0^{-1} \boldsymbol{\Upsilon}_0 \mathbf{Q}_0^{-1}\right) \tag{11.13}$$

where $\mathbf{Q}_0$ *and* $\boldsymbol{\Upsilon}_0$ *are defined as the following probability limits:*

$$\lim_{N\to\infty} \mathbb{V}\text{ar}\left[\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{s}_i(\boldsymbol{x}_i; \boldsymbol{\theta}_0)\right] \xrightarrow{p} \boldsymbol{\Upsilon}_0$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[\mathbf{H}_i(\boldsymbol{x}_i; \boldsymbol{\theta}_0)] \xrightarrow{p} \mathbf{Q}_0$$

which implies the following asymptotic distribution, for a fixed N*.*

$$\widehat{\boldsymbol{\theta}}_M \overset{A}{\sim} \mathcal{N}\left(\boldsymbol{\theta}_0, \frac{1}{N} \mathbf{Q}_0^{-1} \boldsymbol{\Upsilon}_0 \mathbf{Q}_0^{-1}\right) \tag{11.14}$$

Proof. This derivation is reminiscent of the proofs for Theorems 6.17 and 6.18, respectively for MM and MLE estimators, in Lecture 6; this one is in a way more general as it allows for possibly non i.i.d. data. Since, by condition *ii.* the score function is assumed to be continuous and differentiable, then by the Mean Value Theorem one can write:

$$\mathbf{s}_i\left(\mathbf{x}_i; \widehat{\boldsymbol{\theta}}_M\right) = \mathbf{s}_i(\mathbf{x}_i; \boldsymbol{\theta}_0) + \mathbf{H}_i\left(\mathbf{x}_i; \widetilde{\boldsymbol{\theta}}_N\right)\left(\widehat{\boldsymbol{\theta}}_M - \boldsymbol{\theta}_0\right)$$

where $\widetilde{\boldsymbol{\theta}}_N$ is some convex combination of $\widehat{\boldsymbol{\theta}}_M$ and $\boldsymbol{\theta}_0$. By summing over the N observations and dividing by $\sqrt{N}$, one gets:

$$\begin{aligned}\mathbf{0} &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{s}_i \left(\mathbf{x}_i; \widehat{\boldsymbol{\theta}}_M \right) \\ &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{s}_i \left(\mathbf{x}_i; \boldsymbol{\theta}_0 \right) + \left[\frac{1}{N} \sum_{i=1}^{N} \mathbf{H}_i \left(\mathbf{x}_i; \widetilde{\boldsymbol{\theta}}_N \right) \right] \sqrt{N} \left(\widehat{\boldsymbol{\theta}}_M - \boldsymbol{\theta}_0 \right)\end{aligned}$$

by recalling that the sample score evaluated at the solution is equal to zero by definition of M-Estimators. The expression above can be rewritten as:

$$\sqrt{N} \left(\widehat{\boldsymbol{\theta}}_M - \boldsymbol{\theta}_0 \right) = - \left[\frac{1}{N} \sum_{i=1}^{N} \mathbf{H}_i \left(\mathbf{x}_i; \widetilde{\boldsymbol{\theta}}_N \right) \right]^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{s}_i \left(\mathbf{x}_i; \boldsymbol{\theta}_0 \right) \quad (11.15)$$

as *v.* lets invert the average Hessian matrix. Next, consider the following.

1. By *i.* and *v.* one can apply some suitable Law of Large Numbers to the "sample-averaged" Hessian matrix, showing that:

$$\frac{1}{N} \sum_{i=1}^{N} \mathbf{H}_i \left(\mathbf{x}_i; \widetilde{\boldsymbol{\theta}}_N \right) \xrightarrow{p} \mathbf{Q}_0 \quad (11.16)$$

   which follows from the Continuous Mapping Theorem since $\widetilde{\boldsymbol{\theta}}_N \xrightarrow{p} \boldsymbol{\theta}_0$.

2. Condition *iii.* implies $\frac{\partial \mathcal{Q}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left[\mathbf{s}_i \left(\mathbf{x}_i; \boldsymbol{\theta}_0 \right) \right] = \mathbf{0}$ and:

$$\frac{1}{N} \sum_{i=1}^{N} \mathbf{s}_i \left(\mathbf{x}_i; \boldsymbol{\theta}_0 \right) \xrightarrow{p} \mathbf{0}$$

   hence, by condition *iv.* and the Continuous Mapping Theorem, it is as follows.

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{s}_i \left(\mathbf{x}_i; \boldsymbol{\theta}_0 \right) \xrightarrow{d} \mathcal{N} \left(\mathbf{0}, \boldsymbol{\Upsilon}_0 \right)$$

Again, Slutskij's Theorem and the Cramér-Wold Device allow to recombine these intermediate results so to show (11.13). □

As usual, the asymptotic sandwiched variance-covariance is not immediately workable for statistical inference since matrices $\mathbf{Q}_0$ and $\boldsymbol{\Upsilon}_0$ are unknown and must be estimated. The "bread" $\mathbf{Q}_0$ is asymptotically evaluated as:

$$\widehat{\mathbf{Q}}_N \equiv \frac{1}{N} \sum_{i=1}^{N} \mathbf{H}_i \left(\mathbf{x}_i; \widehat{\boldsymbol{\theta}}_M \right) \xrightarrow{p} \mathbf{Q}_0 \quad (11.17)$$

which is straightforward; like in linear models, it is the estimation of the "meat" matrix $\boldsymbol{\Upsilon}_0$ that requires more care. Note that condition *iv.* from the Theorem guarantees that *some* Central Limit Theorem can be applied, but it is silent as to *which* version of it is being invoked. This, in turn, depends on the assumptions regarding the data that the researcher feels confident about making. If the observations are assumed *independent* (but possibly not identically distributed) it is:

$$\boldsymbol{\Upsilon}_0 = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\mathbf{s}_i\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right) \mathbf{s}_i^{\mathrm{T}}\left(\boldsymbol{x}_i, \boldsymbol{\theta}_0\right)\right] \tag{11.18}$$

which specializes further to $\boldsymbol{\Upsilon}_0 = \mathbb{E}\left[\mathbf{s}_i\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right) \mathbf{s}_i^{\mathrm{T}}\left(\boldsymbol{x}_i, \boldsymbol{\theta}_0\right)\right]$ if, in addition, the data are *identically distributed* (i.i.d.). In such cases, $\boldsymbol{\Upsilon}_0$ is consistently estimated as follows.

$$\widehat{\boldsymbol{\Upsilon}}_N = \frac{1}{N} \sum_{i=1}^{N} \mathbf{s}_i\left(\mathbf{x}_i; \widehat{\boldsymbol{\theta}}_M\right) \mathbf{s}_i^{\mathrm{T}}\left(\mathbf{x}_i, \widehat{\boldsymbol{\theta}}_M\right) \xrightarrow{p} \boldsymbol{\Upsilon}_0 \tag{11.19}$$

In analogy with the discussion from Lecture 8 about the consequences of dependent observations, the above might not be a consistent estimator of $\boldsymbol{\Upsilon}_0$ if the observations are dependent. Yet the CCE estimator can be easily adapted to the case of within-group dependence even for M-Estimators:

$$\widehat{\boldsymbol{\Upsilon}}_{CCE} = \frac{1}{N} \sum_{c=1}^{C} \sum_{i=1}^{N_c} \sum_{j=1}^{N_c} \mathbf{s}_{ic}\left(\mathbf{x}_{ic}; \widehat{\boldsymbol{\theta}}_M\right) \mathbf{s}_{jc}^{\mathrm{T}}\left(\mathbf{x}_{jc}; \widehat{\boldsymbol{\theta}}_M\right) \xrightarrow{p} \boldsymbol{\Upsilon}_0 \tag{11.20}$$

where both observations and scores are also indexed by the group or cluster $c = 1, \ldots, C$ which they belong to. Similar extensions to HAC estimation do exist, although in order to account for dependent observations in practical applications of M-Estimators, CCE is overwhelmingly preferred because it is much easier to implement. For any appropriate estimator $\widehat{\boldsymbol{\Upsilon}}_N \xrightarrow{p} \boldsymbol{\Upsilon}_0$, the variance-covariance matrix of M-Estimators is estimated as follows.

$$\widehat{\mathbb{A}\mathrm{var}}\left(\widehat{\boldsymbol{\theta}}_M\right) = \frac{1}{N} \widehat{\mathbf{Q}}_N^{-1} \widehat{\boldsymbol{\Upsilon}}_N \widehat{\mathbf{Q}}_N^{-1} \tag{11.21}$$

The above expression can be used to perform inference about $\widehat{\boldsymbol{\theta}}_M$.

Example 11.6. Asymptotics of OLS. These results are best understood by making appropriate comparisons with OLS. Consistency is easily established under $\mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{x}_i\right] = 0$. The results on asymptotic normality are clearly related to the discussion above by observing that the score and the Hessian matrix of OLS are as in Example 11.3; therefore $\boldsymbol{\Upsilon}_0$ corresponds with $4\boldsymbol{\Xi}_0$; while the "bread" matrices $\mathbf{Q}_0$ are equal to $-2\mathbf{K}_0$ in the OLS case. ■

Example 11.7. Asymptotics of NLLS. For the discussion of the asymptotic properties of NNLS, it is useful to define the following $K \times 1$ vector:

$$\widehat{\mathbf{h}}_i \equiv \frac{\partial}{\partial \boldsymbol{\theta}} h\left(\mathbf{x}_i; \widehat{\boldsymbol{\theta}}_{NLLS}\right)$$

that is, the derivative of the CEF evaluated at $\mathbf{x}_i$ and at the estimate $\widehat{\boldsymbol{\theta}}_{NLLS}$. To evaluate consistency of the NLLS estimator, recall that by construction the latter sets the average score at zero for every value of N.

$$\mathbf{0} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{s}_i\left(y_i, \mathbf{x}_i; \widehat{\boldsymbol{\theta}}_{NLLS}\right) = \frac{1}{N}\sum_{i=1}^{N} 2\widehat{\mathbf{h}}_i \varepsilon_i \xrightarrow{p} \mathbf{0} \tag{11.22}$$

The NLLS is consistent under its motivating assumption about the CEF of Y_i given $\boldsymbol{x}_i$, which for clarity's sake is reported again below.

$$\mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{x}_i\right] = \mathbb{E}\left[y_i \middle| \boldsymbol{x}_i\right] - h\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right) = 0$$

From this condition, along with its direct implication (11.7), it follows that:

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} h\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right)\right)\varepsilon_i\right] = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\boldsymbol{h}_{0i}\varepsilon_i\right] = \mathbf{0}$$

which can be reconciled with (11.22) so long as $\widehat{\mathbf{h}}_i \xrightarrow{p} \boldsymbol{h}_{0i}$ for $i = 1, \ldots, N$, as per some applicable Law of Large Numbers. Thus, the Continuous Mapping Theorem also implies that:

$$\widehat{\boldsymbol{\theta}}_{NLLS} \xrightarrow{p} \boldsymbol{\theta}_0$$

that is, the NLLS estimator is indeed consistent. Regarding the asymptotic distribution, note that *with independent observations*:

$$\mathbf{Q}_0 = \lim_{N\to\infty} -\frac{1}{N}\sum_{i=1}^{N} 2 \cdot \mathbb{E}\left[\boldsymbol{h}_{0i}\boldsymbol{h}_{0i}^{\mathrm{T}}\right]$$

$$\boldsymbol{\Upsilon}_0 = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} 4 \cdot \mathbb{E}\left[\varepsilon_i^2 \boldsymbol{h}_{0i}\boldsymbol{h}_{0i}^{\mathrm{T}}\right]$$

and therefore, by defining the residual $e_i \equiv y_i - h\left(\mathbf{x}_i; \widehat{\boldsymbol{\theta}}_{NLLS}\right)$:

$$\widehat{\mathbb{A}\mathrm{var}}\left(\widehat{\boldsymbol{\theta}}_{NLLS}\right) = \left[\sum_{i=1}^{N} \widehat{\mathbf{h}}_i\widehat{\mathbf{h}}_i^{\mathrm{T}}\right]^{-1}\left[\sum_{i=1}^{N} e_i^2\widehat{\mathbf{h}}_i\widehat{\mathbf{h}}_i^{\mathrm{T}}\right]\left[\sum_{i=1}^{N} \widehat{\mathbf{h}}_i\widehat{\mathbf{h}}_i^{\mathrm{T}}\right]^{-1} \tag{11.23}$$

is the heteroscedasticity-consistent estimator of the variance-covariance of NLLS. Note how it parallels estimator (8.22) for OLS. The homoscedastic, CCE and HAC versions are analogous to their OLS counterparts too. ■

11.3 The Trinity of Asymptotic Tests

After estimating an econometric model, a researcher is often interested into performing some tests of hypothesis, which are possibly non-linear:

$$\mathrm{H}_0 : \ \boldsymbol{v}\left(\boldsymbol{\theta}_0\right) = \mathbf{0} \qquad\qquad \mathrm{H}_1 : \ \boldsymbol{v}\left(\boldsymbol{\theta}_0\right) \neq \mathbf{0}$$

where $\boldsymbol{v}\left(\cdot\right)$, a vector-valued function, has length L (for multiple hypotheses). There are three alternative methods to perform such tests; these are known together as the "Trinity." Under the unifying framework of M-Estimation, these methods have definitions that are uniform across Least Squares, MLE, GMM and other estimators. Here, these methods are briefly reviewed.

Asymptotic Test 1. The Generalized Wald Statistics. Consider the following scalar, called the *generalized* Wald Statistics:

$$\widetilde{\mathrm{W}}_{\mathrm{H}_0} = \boldsymbol{v}^{\mathrm{T}}\left(\widehat{\boldsymbol{\theta}}_M\right) \cdot \left[\widehat{\mathbf{V}} \cdot \widehat{\mathbb{A}\mathrm{var}}\left(\widehat{\boldsymbol{\theta}}_M\right) \cdot \widehat{\mathbf{V}}^{\mathrm{T}}\right]^{-1} \cdot \boldsymbol{v}\left(\widehat{\boldsymbol{\theta}}_M\right) \tag{11.24}$$

where $\widehat{\mathbf{V}} \equiv \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \boldsymbol{v}\left(\widehat{\boldsymbol{\theta}}_M\right)$. The limiting distribution of this statistic is:

$$\widetilde{\mathrm{W}}_{\mathrm{H}_0} \xrightarrow{d} \chi^2_L$$

that is, **under the null hypothesis** H_0 the test statistic has a limiting χ^2_L distribution with L degrees of freedom. This result appears more intuitive upon comparing the generalized Wald Statistic to its original version from the linear model, where $\boldsymbol{v}\left(\boldsymbol{\beta}\right) = \mathbf{R}\boldsymbol{\beta} - \mathbf{c} = \mathbf{0}$ is some linear function. There, the original Wald Statistic is just a particular case of Hotelling's t-squared statistic, which asymptotically follows the chi-squared distribution as per Observation 6.2. This non-linear case is analogously derived after applying the Delta Method to the central matrix of the quadratic form.

Example 11.8. Generalized Wald statistic and a linear constraint. Suppose that interest lies in a specific hypothesis about the linear model:

$$\mathrm{H}_0 : \ \sum_{k=1}^{K} \beta_k = 1 \qquad\qquad \mathrm{H}_1 : \ \sum_{k=1}^{K} \beta_k \neq 1$$

corresponding, for example, to the hypothesis of constant return to scale in production functions (with the constant parameter being β_0). In this case:

$$\widetilde{\mathrm{W}}_{\mathrm{H}_0} = \frac{\left(\sum_{k=1}^{K} \widehat{\beta}_{k,OLS} - 1\right)^2}{\sum_{k=1}^{K}\sum_{q=1}^{K} \widehat{\sigma}_{\beta kq}} \xrightarrow{d} \chi^2_1$$

where $\widehat{\sigma}_{\beta kq}$ is the kq-th element of the estimated variance-covariance of the OLS estimates. ■

The Wald Test is particularly easy to implement, as it only requires to recombine some already calculated estimates. It has, however, two relevant drawbacks in the case of non-linear hypotheses tests. First, it performs quite poorly in small samples: this is a consequence of applying the Delta Method (which is an asymptotic result). The second problem is that the Wald test is not transformation-invariant: in fact, it computes different values of the Wald Statistic for two equivalent hypotheses such as, say, $\mathrm{H}_0 : \beta_k = 0$ and $\mathrm{H}_0 : \exp(\beta_k) = 1$. For all these reasons, the Wald Test should be preferably used only when performing simple tests about linear hypotheses.

Asymptotic Test 2. The Distance, or Likelihood Ratio test. The "Distance Test" was and still is also called "Likelihood Ratio Test," as it was originally conceived in the context of MLE. With respect to the Generalized Wald Test, it has two major advantages: first, it is transformation-invariant; second, it deals non-linear hypotheses quite well. This comes at a cost: this is the most computationally demanding of all tests: in addition to the main estimate of $\boldsymbol{\theta}$ it requires to compute an additional "restricted" estimate

$$\widehat{\boldsymbol{\theta}}_V = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_V}{\arg\max}\, \widehat{\mathcal{Q}}_N(\boldsymbol{\theta})$$

where $\boldsymbol{\Theta}_V = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \boldsymbol{v}(\boldsymbol{\theta}) = \mathbf{0}\}$ is the "restricted parameter space." Then

$$\mathrm{D}_{\mathrm{H}_0} = N\left[\widehat{\mathcal{Q}}_N\left(\widehat{\boldsymbol{\theta}}_M\right) - \widehat{\mathcal{Q}}_N\left(\widehat{\boldsymbol{\theta}}_V\right)\right] \xrightarrow{d} \chi^2_L \tag{11.25}$$

is the expression of the "Distance" Statistic in all cases but MLE, while

$$\mathrm{LR}_{\mathrm{H}_0} = 2\left[\log \widehat{\mathcal{Q}}_N\left(\widehat{\boldsymbol{\theta}}_M\right) - \log \widehat{\mathcal{Q}}_N\left(\widehat{\boldsymbol{\theta}}_V\right)\right] \xrightarrow{d} \chi^2_L \tag{11.26}$$

is the expression of the "Likelihood Ratio" for MLE, where $\widehat{\mathcal{Q}}_N(\boldsymbol{\theta}) = \widehat{\mathcal{L}}_N(\boldsymbol{\theta})$ is the empirical likelihood function (notice a difference in the scaling factor). Intuitively, the test is comparing how much gain is there to make, in terms of explaining the data, by letting the model to be estimated "freely" without the restriction. Clearly, the unrestricted model will always perform statistically better at fitting the data; the question is "how much better" with respect to the researcher's *a priori* hypotheses.

Example 11.9. The distance test and a linear constraint. One can test the same hypothesis as in Example 11.8 through the estimation of a "restricted" model, such as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_{K-1} X_{(K-1)i} + \left(1 - \sum_{k=1}^{K-1} \beta_k\right) X_{Ki} + \epsilon_i$$

which can also be written as follows, with $\ddot{X}_{ki} \equiv X_{ki} - X_{Ki}$ for $k = 1, \ldots, K$.

$$Y_i - X_{Ki} = \beta_0 + \beta_1 \ddot{X}_{1i} + \beta_2 \ddot{X}_{2i} + \cdots + \beta_{K-1} \ddot{X}_{(K-1)i} + \varepsilon_i$$

In this example, the last coefficient of the original model is forced to conform to the restriction that is implied by the null hypothesis; yet imposing the restriction on any other coefficient (except β_0) is equivalent. The Distance Test is computed in this case as:

$$\mathrm{D}_{\mathrm{H}_0} = \left[\sum_{i=1}^{N} \left(y_i - x_{Ki} - \ddot{\mathbf{x}}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_V \right)^2 - \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{OLS} \right)^2 \right] \xrightarrow{d} \chi^2$$

where $\widehat{\boldsymbol{\beta}}_V$ are the parameter estimates from the restricted model above. ■

Asymptotic Test 3. The score – or Lagrange multiplier – test. The last type of test in the Trinity, which also features different names, presents the same advantages as the Distance/LR test, with the extra benefit that it does not require the "unrestricted" model to be estimated at all. Thus, it is computationally more parsimonious. The test is based on the properties of the sample average score function evaluated at **one specific parameter value $\boldsymbol{\theta}_{\mathbf{v}}$** implied by the null hypothesis. Recall that the average score:

$$\frac{1}{N} \sum_{i=1}^{N} \mathbf{s}_i \left(\mathbf{x}_i; \widehat{\boldsymbol{\theta}}_M \right) = \mathbf{0}$$

always equals zero when evaluated at the unrestricted estimate value $\widehat{\boldsymbol{\theta}}_M$, by the definition of M-Estimators. Consider, however, a *restricted* parameter value $\boldsymbol{\theta}_{\mathbf{v}}$ such that $\boldsymbol{v}(\boldsymbol{\theta}_{\mathbf{v}}) = \mathbf{0}$. It follows that:

$$\frac{1}{N} \sum_{i=1}^{N} \left| \mathbf{s}_i \left(\mathbf{x}_i; \boldsymbol{\theta}_{\mathbf{v}} \right) \right| > \mathbf{0} \tag{11.27}$$

the K-dimensional sample score vector deviates from zero when evaluated at any "suboptimal" parameter choice. The Lagrange Multiplier statistic is:

$$\mathrm{LM}_{\mathrm{H}_0} = N \left[\sum_{i=1}^{N} \mathbf{s}_i \left(\mathbf{x}_i; \boldsymbol{\theta}_{\mathbf{v}} \right) \right]^{\mathrm{T}} \widehat{\boldsymbol{\Upsilon}}_{\mathbf{v}}^{-1} \left[\sum_{i=1}^{N} \mathbf{s}_i \left(\mathbf{x}_i; \boldsymbol{\theta}_{\mathbf{v}} \right) \right] \xrightarrow{d} \chi^2_K \tag{11.28}$$

where, as appropriate:

$$\widehat{\boldsymbol{\Upsilon}}_{\mathbf{v}} = \widehat{\mathbb{A}\mathrm{var}} \left[\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{s}_i \left(\mathbf{x}_i; \boldsymbol{\theta}_{\mathbf{v}} \right) \right]$$

and it is indicative of how statistically relevant are the deviations in (11.27), because it computes a quadratic form of their standardized values.

Example 11.10. The score test and a linear constraint. Continuing the running Example 11.8-11.9, a set of parameters $\boldsymbol{\beta}_{\mathbf{v}}$ of dimension K that satisfies the hypothesized restriction could be computed from the estimates $\widehat{\boldsymbol{\beta}}_V$ of the restricted model as follows.

$$\boldsymbol{\beta}_{\mathbf{v}} = \left(\widehat{\beta}_{0V}, \widehat{\beta}_{1V}, \widehat{\beta}_{2V}, \ldots, \widehat{\beta}_{(K-1)V}, 1 - \sum_{k=1}^{K-1} \widehat{\beta}_{kV} \right)^{\mathrm{T}}$$

Therefore, hence, the Lagrange Multiplier Statistic would read as:

$$\mathrm{LM}_{\mathrm{H}_0} = \left[\sum_{i=1}^{N} \mathbf{x}_i e_i \left(\boldsymbol{\beta}_{\mathbf{v}} \right) \right]^{\mathrm{T}} \left[\sum_{i=1}^{N} e_i^2 \left(\boldsymbol{\beta}_{\mathbf{v}} \right) \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \right]^{-1} \left[\sum_{i=1}^{N} \mathbf{x}_i e_i \left(\boldsymbol{\beta}_{\mathbf{v}} \right) \right] \xrightarrow{d} \chi^2_K$$

given the "restricted" residuals $e_i \left(\boldsymbol{\beta}_{\mathbf{v}} \right) = y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_{\mathbf{v}}$. ■

11.4 Quasi-Maximum Likelihood

The asymptotic properties of Maximum Likelihood estimators are especially desirable in light of the considerations that are made in Lecture 6 following the analysis of Theorem 6.18. The main implications can be rephrased using the notation developed in this Lecture; *with i.i.d. data* the MLE framework has the property that:

$$\boldsymbol{\Upsilon}_0 = -\mathbf{Q}_0$$

which follows from the Information Matrix Equality. Therefore, for a fixed N the asymptotic distribution of any Maximum Likelihood estimator is:

$$\widehat{\boldsymbol{\theta}}_{MLE} \overset{A}{\sim} \mathcal{N} \left(\boldsymbol{\theta}_0, \left[\boldsymbol{I}_N \left(\boldsymbol{\theta}_0 \right) \right]^{-1} \right) \tag{11.29}$$

where $\boldsymbol{I}_N \left(\boldsymbol{\theta}_0 \right)$ is the information matrix; thus the variance-covariance hits the Cramér-Rao bound making MLE the most suitable choice for estimation *so long as the distributional assumptions can be believed*. This also delivers the convenient implication that the information matrix can be consistently estimated in two alternative ways, by either $\widehat{\boldsymbol{\Upsilon}}_N$ or $-\widehat{\mathbf{Q}}_N$, where matrix $\widehat{\boldsymbol{\Upsilon}}_N$ is evaluated according to expression (11.19).[1] In the econometric practice, the method based on the so-called **outer product of the gradients** (OPG), that is using $\widehat{\boldsymbol{\Upsilon}}_N$, is often favored due to computational considerations. In fact, statistical softwares routinely compute scores in order to perform MLE, and the OPG adds little computational cost to the problem of estimating the sample variance-covariance.

[1] In the treatment developed in Lecture 6, the two alternative options are expressed through the notation $\widehat{\mathbf{H}}_N$ and $\widehat{\mathbf{J}}_N$, respectively.

Example 11.11. MLE and Regression, continued. Continue the analysis of the MLE estimator of a linear regression model with normal disturbances from Example 11.5. To perform statistical inference, it is necessary to estimate the variance of the estimates. In this context, the observation-specific *score* – the individual contribution to the log-likelihood function – is:

$$\mathbf{s}_i\left(y_i, \mathbf{x}_i; \boldsymbol{\theta}\right) = \begin{bmatrix} \mathbf{x}_i \left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right) \sigma^{-2} \\ -\frac{1}{2}\sigma^{-2} + \frac{1}{2}\left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 \sigma^{-4} \end{bmatrix}$$

hence the *individual Hessian matrix*, for some given value of $\boldsymbol{\theta}$, is:

$$\mathbf{H}_i\left(y_i, \mathbf{x}_i; \boldsymbol{\theta}\right) = \begin{bmatrix} -\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\sigma^{-2} & -\mathbf{x}_i \left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right) \sigma^{-4} \\ -\mathbf{x}_i^{\mathrm{T}} \left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right) \sigma^{-4} & \frac{1}{2}\sigma^{-4} - \left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 \sigma^{-6} \end{bmatrix}$$

which is symmetric. By plugging the MLE estimates into the above matrix, summing it over all the observations and taking the inverse of the result one obtains a consistent estimate of the *opposite* of the information matrix:

$$\frac{\widehat{\mathbf{Q}}_N^{-1}}{N} = \left[\sum_{i=1}^{N} \mathbf{H}_i\left(y_i, \mathbf{x}_i; \widehat{\boldsymbol{\theta}}_{MLE}\right)\right]^{-1} = -\begin{bmatrix} \left(\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1} \widehat{\sigma}^2_{MLE} & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & \frac{2}{N}\widehat{\sigma}^4_{MLE} \end{bmatrix}$$

where the border elements (except the lower bottom one) are equal to zero because they are proportional to the first K elements of the sample score – that is, the K normal equations.[2] An equivalent way to obtain the estimator of interest is to calculate the outer product of the gradients. By the above expression for $\mathbf{s}_i\left(y_i, \mathbf{x}_i; \boldsymbol{\theta}\right)$, it is easy to verify that:

$$\begin{aligned} \frac{\widehat{\boldsymbol{\Upsilon}}_N^{-1}}{N} &= \left[\sum_{i=1}^{N} \mathbf{s}_i\left(y_i, \mathbf{x}_i; \widehat{\boldsymbol{\theta}}_{MLE}\right) \mathbf{s}_i^{\mathrm{T}}\left(y_i, \mathbf{x}_i; \widehat{\boldsymbol{\theta}}_{MLE}\right)\right]^{-1} \\ &= \begin{bmatrix} \left(\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1} \widehat{\sigma}^2_{MLE} & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & \frac{2}{N}\widehat{\sigma}^4_{MLE} \end{bmatrix} \end{aligned}$$

[2] The calculations to obtain the *opposite* of the bottom right element of this matrix (that is, the asymptotic variance of $\widehat{\sigma}^2_{MLE}$) are as follows.

$$\begin{aligned} \widehat{\mathbb{A}\mathrm{var}}\left(\widehat{\sigma}^2_{MLE}\right) &= -\left(\frac{N}{2}\widehat{\sigma}^{-4}_{MLE} - \sum_{i=1}^{N}\left(y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{MLE}\right)^2 \widehat{\sigma}^{-6}_{MLE}\right)^{-1} \\ &= -\left(\frac{N}{2}\widehat{\sigma}^{-4}_{MLE} - N\widehat{\sigma}^{-4}_{MLE}\right)^{-1} = -\left(-\frac{N}{2}\widehat{\sigma}^{-4}_{MLE}\right)^{-1} = \frac{2}{N}\widehat{\sigma}^4_{MLE} \end{aligned}$$

For the outer product of the gradients the calculations are similar.

that is, $\widehat{\boldsymbol{\Upsilon}}_N^{-1} = -\widehat{\mathbf{Q}}_N^{-1}$ as predicted by the information matrix equality. This result highlights more clearly that the OLS estimator of the variance of $\boldsymbol{\beta}$ in small samples (under the homoscedasticity assumption) differs from the Cramér-Rao bound by a multiplicative factor of $\frac{N-K}{N}$. ■

Unfortunately, the desirable properties of MLE break down if the i.i.d. hypothesis cannot be defended, since the *information matrix equality fails*. To illustrate, let again $\boldsymbol{x}_i = (\boldsymbol{y}_i, \boldsymbol{z}_i)$ be the collection of all endogenous and exogenous variables in the model, and allow for group dependence between observations. In this case, the likelihood function can be factored between clusters:

$$\mathcal{L}(\boldsymbol{\theta}|\, \mathbf{x}_1, \ldots, \mathbf{x}_N) = \prod_{c=1}^{C} f_{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_c}}(\mathbf{x}_{1c}, \ldots, \mathbf{x}_{N_c c}; \boldsymbol{\theta})$$

but not further; the information matrix equality no longer applies. A similar argument applies to more general cases of spatial and time dependence, as well as to the i.n.i.d. case where the observations are independent but not identically distributed (for example, when the homoscedasticity assumption which is implicit in the CMLE model described in Example 11.5 cannot be maintained, even if the error terms are always conditionally normal). In all these cases, MLE retains the sandwiched limiting variance of M-Estimators as per (11.13), and $\boldsymbol{\Upsilon}_0$ must be estimated according to the working assumptions – for example, by formula (11.20) under group dependence.

As it has been already observed, the almost ideal asymptotic properties of MLE break down even if the data are generated from a random (i.i.d.) sample, but the likelihood function is *misspecified*, that is it does not match the "true" data generation process in the population under examination. It is interesting to investigate the *consequences of misspecifation*, that is of estimating a model via MLE while assuming a wrong underlying distribution, since this can occur frequently in practice. In such cases, the estimator of interest is called the **Quasi-Maximum Likelihood Estimator** $\widehat{\boldsymbol{\theta}}_{QMLE}$, and it is useful to characterize its probability limit, which is commonly called the **pseudo-true value** $\boldsymbol{\theta}^*$:

$$\widehat{\boldsymbol{\theta}}_{QMLE} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\max}\, \mathcal{L}(\boldsymbol{\theta}|\, \mathbf{x}_1, \ldots, \mathbf{x}_N) \xrightarrow{p} \boldsymbol{\theta}^* \tag{11.30}$$

where the probability limit is evaluated with respect to the *true* distribution. A relevant question is whether the QMLE is consistent, that is $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$.

In fact, in the main example of MLE examined so far – the linear model under normality assumptions – it can be observed that the ML estimator of $\boldsymbol{\beta}_0$ coincides with the standard OLS estimator, so it is consistent if the standard conditional mean assumption for linear models $\mathbb{E}[\varepsilon_i|\, \boldsymbol{x}_i] = 0$ holds,

even if the true underlying distribution is not normal. This observation is actually more general than that: if the assumed distribution is $f_{Y,\boldsymbol{x}}(y_i, \mathbf{x}_i; \boldsymbol{\theta})$ for some scalar endogenous variable Y_i, and in addition it belongs to the **exponential** macro-**family** of distributions, that is it can be decomposed in terms of some primitive scalar functions $a[\cdot]$, $b[\cdot]$ and $c[\cdot]$ as:

$$f_{Y,\boldsymbol{x}}(y_i, \mathbf{x}_i; \boldsymbol{\theta}) = \exp\left\{a\left[\mu_{Y|\boldsymbol{x}}(\mathbf{x}_i; \boldsymbol{\theta})\right] + b[y_i] + c\left[\mu_{Y|\boldsymbol{x}}(\mathbf{x}_i; \boldsymbol{\theta})\right] y_i\right\}$$

where $\mu_{Y|\boldsymbol{x}}(\boldsymbol{x}_i; \boldsymbol{\theta}) \equiv \mathbb{E}[Y_i|\, \boldsymbol{x}_i; \boldsymbol{\theta}]$ is a parametric specification of the CEF of Y_i given $\boldsymbol{x}_i$, then MLE consistently estimates $\mu_{Y|\boldsymbol{x}}(\boldsymbol{x}_i; \boldsymbol{\theta})$ if the CEF is *correctly specified in the model*. For linear models this is quite a familiar condition, but the importance of this general statistical result goes beyond linear models, since it applies to all non-linear models that are estimated via MLE and are based on a distribution belonging to the exponential macro-family. In fact, quite a relevant number of distributions that are commonly employed in fully parametric econometric models belongs to the exponential family, as it is already observed in Lecture 5.

Example 11.12. Poisson Regression. A **count data model** is a model suited for explaining some variable of interest Y_i that only assumes non-negative integer values $Y_i = 0, 1, 2, \ldots$ and where smaller values occur with higher frequency than larger ones. A simple example of a count data model is the **Poisson regression**, which is based on the Poisson distribution:

$$\mathbb{P}(Y_i|\, \boldsymbol{x}_i) = \frac{\lambda_i(\boldsymbol{x}_i)^{Y_i} \exp(-\lambda_i(\boldsymbol{x}_i))}{Y_i!}$$

where the count Y_i for each observation $i = 1, \ldots, N$ of an N-dimensional sample is assumed to be Poisson-distributed each with a distinct Poisson parameter $\lambda_i(\boldsymbol{x}_i)$, treated as a function of the individual characteristics $\boldsymbol{x}_i$. By the properties of the Poisson distribution:

$$\lambda_i(\boldsymbol{x}_i) = \mathbb{E}[Y_i|\, \boldsymbol{x}_i] = \mathbb{V}\mathrm{ar}[Y_i|\, \boldsymbol{x}_i]$$

that is, $\lambda_i(\boldsymbol{x}_i)$ equals both the conditional mean and the conditional variance of Y_i given $\boldsymbol{x}_i$. The most common choice is $\lambda_i(\boldsymbol{x}_i) = \exp\left(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0\right)$; this results in the following conditional distribution.

$$\mathbb{P}(Y_i|\, \boldsymbol{x}_i) = \frac{\exp\left(Y_i \cdot \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0\right) \exp\left[-\exp\left(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0\right)\right]}{Y_i!}$$

An implication of this assumption is that:

$$\frac{\partial \exp\left(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0\right)}{\partial \boldsymbol{x}_i} = \exp\left(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0\right) \boldsymbol{\beta}_0 \;\Rightarrow\; \boldsymbol{\beta}_0 = \frac{\partial \mathbb{E}[Y_i|\, \boldsymbol{x}_i]}{\partial \boldsymbol{x}_i} \frac{1}{\mathbb{E}[Y_i|\, \boldsymbol{x}_i]}$$

hence, $\boldsymbol{\beta}_0$ can be interpreted as a **semi-elasticiy** just like in a *log-lin* model. A Poisson regression may be more convenient than a simple linear regression of $\log Y_i$ on $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ as it allows for the frequent observations $Y_i = 0$.

Given a random sample $\{(y_i, \mathbf{x}_i)\}_{i=1}^N$, the log-likelihood function of the Poisson Regression model is the following.

$$\log \mathcal{L}\left(\boldsymbol{\beta} | \left\{(y_i, \mathbf{x}_i)\right\}_{i=1}^N\right) = \sum_{i=1}^N \left[y_i \cdot \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \exp\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right) - \log y_i!\right]$$

The First Order Conditions of the MLE problem, expressed as the sum of the individual scores, are:

$$\sum_{i=1}^N \mathbf{s}_i\left(y_i, \mathbf{x}_i; \widehat{\boldsymbol{\beta}}_{MLE}\right) = \sum_{i=1}^N \mathbf{x}_i\left[y_i - \exp\left(\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{MLE}\right)\right] = \mathbf{0}$$

and they lack a closed form solution; consequently, the estimator in question must be obtained by numerical methods. The *empirical* Hessian matrix is:

$$\widehat{\mathbf{Q}}_N = \frac{1}{N}\sum_{i=1}^N \mathbf{H}_i\left(y_i, \mathbf{x}_i; \widehat{\boldsymbol{\beta}}_{MLE}\right) = -\frac{1}{N}\sum_{i=1}^N \exp\left(\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{MLE}\right) \cdot \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}$$

the opposite of which – divided by N – is an appropriate estimator of the information matrix. Since the Poisson distribution belongs to the exponential family, even if the likelihood is misspecified the CEF of Y_i given $\boldsymbol{x}_i$ is consistently estimated if it is itself well specified; together with the "exponential" specification for $\lambda_i(\boldsymbol{x}_i)$ this implies that the MLE can be interpreted in such a model in terms of semi-elasticities. ■

The Poisson regression is quite an extreme example where MLE works even if the likelihood is misspecified. In the case of the conditionally normal linear model, even if the estimates of $\boldsymbol{\beta}_0$ survive the misspecification, the MLE estimate of σ^2 is rendered meaningless even by the smallest deviation from the parametric assumptions – for example, if the errors are still normal, but heteroscedastic – which affects the estimate of the variance of $\widehat{\boldsymbol{\beta}}_{MLE}$. This is a consequence of the more general fact that under misspecification of the likelihood function, the information matrix equality fails, and it can be shown that the limiting distribution of the QMLE is:

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_{QMLE} - \boldsymbol{\theta}^*\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left[\mathbf{Q}^*\right]^{-1}\boldsymbol{\Upsilon}^*\left[\mathbf{Q}^*\right]^{-1}\right) \tag{11.31}$$

where $\mathbf{Q}^*$ and $\boldsymbol{\Upsilon}^*$ are the analogues of $\mathbf{Q}_0$ and $\boldsymbol{\Upsilon}_0$ respectively, but evaluated at $\boldsymbol{\theta}^*$ instead of $\boldsymbol{\theta}_0$. Hence, even if the QMLE is consistent, it is safest to

estimate its covariance matrix as if the true asymptotic variance-covariance took the standard sandwiched formula (11.14) of M-Estimators, even if the observations are indeed both independent and identically distributed. This is analogous to cautiously estimating the variance of OLS estimates by the heteroscedasticity-robust formula when homoscedasticity is not too certain.

In more general cases where the assumed distribution does not belong to the exponential family, the QMLE is actually inconsistent: $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}_0$. Yet not all is lost! It turns out that the QMLE can be still given an interpretation in terms of "best approximation" to the true distribution, similarly as OLS can be interpreted as the best approximation to the true CEF in the population. The theory underlying this approach relies on the analysis of the so-called **Kullback-Leibler Information Criterion** (KLIC), defined as:

$$\begin{aligned}\mathcal{K}_{\boldsymbol{x}}(\boldsymbol{\theta}) &\equiv \mathbb{E}_g\left[\log\left(\frac{g_{\boldsymbol{x}}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N;\boldsymbol{\theta}_0)}{f_{\boldsymbol{x}}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N;\boldsymbol{\theta})}\right)\right] \\ &= \int_{\mathbb{X}} \log\left(\frac{g_{\boldsymbol{x}}(\mathbf{x}_1,\ldots,\mathbf{x}_N;\boldsymbol{\theta}_0)}{f_{\boldsymbol{x}}(\mathbf{x}_1,\ldots,\mathbf{x}_N;\boldsymbol{\theta})}\right) g_{\boldsymbol{x}}(\mathbf{x}_1,\ldots,\mathbf{x}_N;\boldsymbol{\theta}_0)\,\mathrm{d}\mathbf{x}_1\ldots\mathrm{d}\mathbf{x}_N\end{aligned}$$

where $f_{\boldsymbol{x}}(\cdot)$ is the *assumed* joint mass or density function generating the data, while $g_{\boldsymbol{x}}(\cdot)$ is the *true* function, which is taken as given; the expectation is taken with respect to $g_{\boldsymbol{x}}(\cdot)$. Note that by construction, $\mathcal{K}_{\boldsymbol{x}}(\boldsymbol{\theta}) \geq 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$; but if the distribution is correctly specified, it is $f_{\boldsymbol{x}}(\cdot) = g_{\boldsymbol{x}}(\cdot)$, and thus $\mathcal{K}_{\boldsymbol{x}}(\boldsymbol{\theta}_0) = 0$: the KLIC would attain its minimum. In addition:

$$\begin{aligned}\mathcal{K}_{\boldsymbol{x}}(\boldsymbol{\theta}) &= \mathbb{E}_g\left[\log g_{\boldsymbol{x}}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N;\boldsymbol{\theta}_0) - \log f_{\boldsymbol{x}}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N;\boldsymbol{\theta})\right] \\ &= \mathbb{E}_g\left[\log g_{\boldsymbol{x}}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N;\boldsymbol{\theta}_0)\right] - \log \mathcal{L}_0^g(\boldsymbol{\theta}|\,\mathbf{x}_1,\ldots,\mathbf{x}_N)\end{aligned}$$

where $\log \mathcal{L}_0^g(\boldsymbol{\theta}|\,\mathbf{x}_1,\ldots,\mathbf{x}_N)$ is the pseudo-population likelihood that results if the assumed distribution is $f_{\boldsymbol{x}}(\cdot)$, but the true one is $g_{\boldsymbol{x}}(\cdot)$. Clearly, under general assumptions:

$$\widehat{\boldsymbol{\theta}}_{QMLE} \xrightarrow{p} \boldsymbol{\theta}^* = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \log \mathcal{L}_0^g(\boldsymbol{\theta}|\,\mathbf{x}_1,\ldots,\mathbf{x}_N) = \min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{K}_{\boldsymbol{x}}(\boldsymbol{\theta})$$

that is, the QMLE converges in probability to the pseudo-true value, which is at the same time the maximizer of the pseudo-population likelihood and the minimizer of the KLIC! In this sense, the probability limit of the QMLE minimizes a well-defined criterion of *distance* between the assumed and the true density, and as such it is thinkable as some "best approximation" of sort. Mirroring the analogous discussion of Least Squares as best approximation of the CEF, this is not an excuse for disregarding the problem of correctly specifying the likelihood function! However, it motivates the applied practice of enriching MLE models with flexible parametric specifications (say, polynomials of $\boldsymbol{x}_i$) in order to best approximate the true distribution.

11.5 Introduction to Binary Outcome Models

A case where the choice of MLE over semi-parametric methods is well motivated is when some dependent variable Y_i is **limited**, meaning that it only takes a discrete set of values. In the most extreme, as well as most common cases of **limited dependent variable** (LDV) models, Y_i is a dummy or **binary** variable: $Y_i \in \{0, 1\}$. The reason is that economists are constantly interested in the determinants of binary outcomes, that are usually framed as microeconomic problems of choice over two alternatives. Examples are:

- What are the determinants of individual enrollment in college?
- Which factors influence on the probability of default of a firm?
- What are the causes of the eruption of civil war in a country?

in all these cases, either outcome takes one value (say *college*, *default* and *civil war* take $Y_i = 1$) while the alternative takes the other ($Y_i = 0$).

Other LDV models take multiple outcomes:

- What means of transportations do individuals choose for commuting?
- Which characteristic of one country determine its political regime?
- What type of insurance contract is preferred by different individuals?
- Which individual characteristics predict people's responses in surveys?

and the choice of LDV models often depends on whether alternatives can be *nested* in groups (e.g. public vs. private transportation; democratic vs. authoritarian regimes), and on whether they can be ranked or *ordered* (e.g. insurance contracts from minimal to maximal coverage; "strongly disagree" to "strongly agree" types of answer in surveys). The objective here to either review all types of LDV models or to summarize the immense literature about their econometric estimation. Instead, the aim is to provide a *minimal* introduction to the most common **binary outcome models**, in order to make a case for M-Estimation and specifically MLE in concrete economic settings. Moreover, this is useful towards the eventual discussion of other applications of MLE in the next section.

Let us consider a problem with binary outcomes $Y_i \in \{0, 1\}$. Treating the latter as random (Bernoulli) events, it is natural to think about their realization probability as a conditional function of some variables $\boldsymbol{x}_i$:

$$\mathbb{P}\left(Y_i = 1 \middle| \boldsymbol{x}_i\right) = G\left(\boldsymbol{x}_i, \boldsymbol{\beta}_0\right)$$
$$\mathbb{P}\left(Y_i = 0 \middle| \boldsymbol{x}_i\right) = 1 - G\left(\boldsymbol{x}_i, \boldsymbol{\beta}_0\right)$$

where $G(\cdot)$ is some function of the variables $\boldsymbol{x}_i$ parametrized by vector $\boldsymbol{\beta}_0$. Notice that as the problem is binary, the probability of either outcome can be treated residually with respect to the other's. In fact, one can write the conditional expectation of Y_i given $\boldsymbol{x}_i$ as

$$\begin{aligned}\mathbb{E}[Y_i|\,\boldsymbol{x}_i] &= 1\cdot[G(\boldsymbol{x}_i,\boldsymbol{\beta}_0)]+0\cdot[1-G(\boldsymbol{x}_i,\boldsymbol{\beta}_0)]\\ &= G(\boldsymbol{x}_i,\boldsymbol{\beta}_0)\end{aligned}$$

as it happens with any Bernoulli distribution.

A natural question is whether this model is estimable via linear regression: this is called the **linear probability model** (LPM):

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0+\epsilon_i,\quad y_i\in\{0,1\}$$

which is equivalent to assuming $G(\boldsymbol{x}_i,\boldsymbol{\beta}_0)=\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0$. Note that by definition of regression it is $\epsilon_i = Y_i-\mathbb{E}[Y_i|\,\boldsymbol{x}_i] = Y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0$; and so it follows that:

$$\begin{aligned}\mathbb{P}\left(\epsilon_i = 1-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\middle|\,\boldsymbol{x}_i\right) &= \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\\ \mathbb{P}\left(\epsilon_i = -\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\middle|\,\boldsymbol{x}_i\right) &= 1-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\end{aligned}$$

yet this instance of "natural heteroscedasticity" also implies:

$$\mathbb{E}[\epsilon_i|\,\boldsymbol{x}_i] = \left(1-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right)\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\left(1-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right)=0$$

hence OLS applied to this model would still produce unbiased and consistent estimates of $\boldsymbol{\beta}_0$ even if the problem is naturally heteroscedastic (something that is normally addressed either via "robust" standard errors or, in small samples, via FGLS). The main issues of the LPM depend on the fact that the linear conditional expectation $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0$ cannot be constrained to lie within the $(0,1)$ interval. This implies that:

1. the conditional variance of the error term might take negative values;

$$\mathbb{V}\mathrm{ar}[\epsilon_i|\,\boldsymbol{x}_i] = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\left(1-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right)\gtreqless 0$$

2. the predicted probabilities $\widehat{\mathbb{E}}[Y_i|\,\boldsymbol{x}_i=\mathbf{x}_i] = \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{LPM} = \widehat{y}_i$ might take values outside the $[0,1]$ interval.

Both are facts that make no probabilistic sense. Unsurprisingly, the LPM is used in practice only in a few limited circumstances, that is for the sake of comparison with other LDV models or in well-defined quasi-experimental settings where interest falls, in particular, on the transparent estimation of the causal effect of some variable X_i upon a binary outcome Y_i.

In general, however, econometricians tends to prefer non-linear models that produce consistent predictions of the predicted outcomes probabilities. For any given realization $\mathbf{x}_i$, an obvious choice is:

$$G\left(\mathbf{x}_i, \boldsymbol{\beta}_0\right) = F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right) = F_{\boldsymbol{x}}\left(\lambda_i\right) \tag{11.32}$$

where $F_{\boldsymbol{x}}\left(\cdot\right)$ is a **probability distribution function** with a single "free" parameter (typically a location parameter) $\lambda_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0$. Furthermore, most applications make use of a **symmetric** distribution function, that is one for which $F\left(\lambda_i\right) = 1 - F\left(-\lambda_i\right)$. It is apparent that this choice solves the problem of the predicted probabilities, given that by the very definition of probability distribution function, conditionally on any realization $\mathbf{x}_i$:

$$\begin{aligned}
\lim_{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 \to +\infty} \mathbb{P}\left(Y_i = 1 \middle| \mathbf{x}_i\right) &= \lim_{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 \to +\infty} F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right) = 1 \\
\lim_{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 \to -\infty} \mathbb{P}\left(Y_i = 1 \middle| \mathbf{x}_i\right) &= \lim_{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 \to -\infty} F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right) = 0
\end{aligned}$$

and symmetrically for $Y_i = 0$. In addition, this model has a clear advantage that appeals econometricians: it can be motivated on a "structural" model of individual choice, which describes the (micro-)economics of the problem.

Specifically, in its simplest form a **latent variable model** for a binary outcome is a model that reads like

$$y_i^* = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + \varepsilon_i \tag{11.33}$$

$$y_i = \begin{cases} 1 & \text{if } y_i^* > \alpha_0 \\ 0 & \text{if } y_i^* \leq \alpha_0 \end{cases} \tag{11.34}$$

where Y_i^* is a **latent variable** that represents the cost-benefit evaluation of the binary choice by the i-th individual. This latent variable is a theoretical construct that cannot be observed by the econometrician – much like any error term – but that is assumed to determine the choice of the outcome according to a simple rule. In particular, if Y_i^* is larger than some unknown "threshold" parameter α_0, then $Y_i = 1$ is chosen; otherwise $Y_i = 0$ is opted for. Notably, the latent variable is a linear function of the individual characteristics $\mathbf{x}_i$ and a specific error term ε_i, which is distributed according to $F_{\boldsymbol{x}}\left(\cdot\right)$. In such a model, conditionally on any realization $\mathbf{x}_i$:

$$\mathbb{P}\left(Y_i = 1 \middle| \mathbf{x}_i\right) = \mathbb{P}\left(Y_i^* > \alpha_0 \middle| \mathbf{x}_i\right) = \mathbb{P}\left(\varepsilon_i > -\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + \alpha_0 \middle| \mathbf{x}_i\right) \tag{11.35}$$

where intuitively, if $\mathbf{x}_i$ contains a constant element, its associated "intercept" parameter and α_0 are not separately identified (they would have the same

implication as the conditional probability of choosing $Y_i = 1$ over $Y_i = 0$ for some given $\mathbf{x}_i = 0$). Hence, the **normalization** $\alpha_0 = 0$ is typically imposed (this has no practical implications as the "constant probability" is included in $\boldsymbol{\beta}_0$). Moreover, if $F_{\boldsymbol{x}}(\cdot)$ is a symmetric distribution, (11.35) reshapes as:

$$\begin{aligned}
\mathbb{P}(Y_i = 1|\,\mathbf{x}_i) &= \mathbb{P}(Y_i^* > 0|\,\mathbf{x}_i) \\
&= \mathbb{P}\left(\varepsilon_i > -\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\big|\,\mathbf{x}_i\right) \\
&= 1 - \mathbb{P}\left(\varepsilon_i \leq -\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\big|\,\mathbf{x}_i\right) \\
&= \mathbb{P}\left(\varepsilon_i < \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\big|\,\mathbf{x}_i\right) \\
&= F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right)
\end{aligned} \tag{11.36}$$

where the fourth line exploits the symmetry of $F_{\boldsymbol{x}}(\cdot)$. This fact reconciles the latent variable model with our specification of the conditional probability for the outcome Y_i.

Before getting to practical aspects and the MLE estimation of models with binary outcomes, two observations need to be made.

- Latent variable models are not specific of binary outcomes: multinomial LDV models are usually motivated by more complex versions, which are outside the scope of this overview. Latent variable models are also used in the structural analysis of empirical strategic games.

- Derivation (11.36) above shows why $F_{\boldsymbol{x}}(\cdot)$ should not contain a variable scale parameter, such as the variance (if $F_{\boldsymbol{x}}(\cdot)$ is, say, normal, its variance should be known or normalized, e.g. $\sigma^2 = 1$); otherwise the K parameters in $\boldsymbol{\beta}_0$ and the scale parameter would not be separately identified. To see intuitively why, consider the case where $\alpha_0 = 0$ and $F_{\boldsymbol{x}}(\cdot)$ features some scale parameter, call it σ. Here, the two equations:

$$\begin{aligned}
y_i^* &= \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + \varepsilon_i \\
\sigma y_i^* &= \sigma\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + \varepsilon_i\right)
\end{aligned}$$

  are observationally equivalent, that is $F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right) = F_{\boldsymbol{x}}\left(\sigma \cdot \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right)$. The intuition behind this is that one can only observe whether the latent variable takes values above ($Y_i = 1$) or below ($Y_i = 0$) its hypothesized threshold, and not its variation as a function of the variation of $\boldsymbol{x}_i$. For a similar reason, the scale parameter *could* be identified *if* $\alpha_0 = 0$ *and* $\boldsymbol{x}_i$ did not include any constant term. In such a case the "scale" parameter would be identified by changes in the average value of Y_i that are not explained by $\boldsymbol{x}_i$. However, the basic fact that scale parameters are not independently identified remains. In general, there is seldom a reason to include a scale parameter instead of a constant location parameter.

Given the choice of a symmetric probability distribution $F_{\boldsymbol{x}}(\cdot)$ and a sample $\{(y_i, \mathbf{x}_i)\}_{i=1}^N$, the likelihood function of a binary choice model is:

$$\mathcal{L}\left(\boldsymbol{\beta} | \{(y_i, \mathbf{x}_i)\}_{i=1}^N\right) = \prod_{i=1}^N \left[F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right]^{y_i} \left[1 - F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right]^{1-y_i}$$

which is just a generalization of the likelihood function for a Bernoulli sample. The corresponding log-likelihood function is:

$$\log \mathcal{L}\left(\boldsymbol{\beta} | \{(y_i, \mathbf{x}_i)\}_{i=1}^N\right) = \\ = \sum_{i=1}^N \left\{y_i \log F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right) + (1 - y_i) \log\left[1 - F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right]\right\}$$

where the First Order Conditions (the sum of the individual scores, evaluated at the estimates) are:

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log \mathcal{L}\left(\widehat{\boldsymbol{\beta}}_{MLE} \middle| \{(y_i, \mathbf{x}_i)\}_{i=1}^N\right) = \sum_{i=1}^N \mathbf{s}_i\left(y_i, \mathbf{x}_i; \widehat{\boldsymbol{\beta}}_{MLE}\right) = \\ = \sum_{i=1}^N \left[\frac{y_i f_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{MLE}\right)}{F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{MLE}\right)} - \frac{(1 - y_i) f_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{MLE}\right)}{1 - F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{MLE}\right)}\right] \mathbf{x}_i = \mathbf{0}$$

where $f_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)$ is the probability density function associated with the – implicity continuous – distribution $F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)$. Like in the Poisson regression, there is no closed form solution, thus the estimator must be calculated via numerical methods; its variance-covariance is more conveniently estimated via the OPG.

Clearly, the exact solution depends on the assumptions made on $F_{\boldsymbol{x}}(\cdot)$. Even if other possibilities exist, the most common choices are:

- the **probit** model, in which $F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right) = \Phi\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right)$ where function $\Phi(\cdot)$ is a **cumulative standard normal distribution**:

$$\mathbb{P}\left(Y_i = 1 | \mathbf{x}_i\right) = \Phi\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right) = \int_{-\infty}^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{t^2}{2}\right) dt$$

- the **logit** model, in which $F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right) = \Lambda\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right)$ where function $\Lambda(\cdot)$ is a **scale-normalized cumulative logistic distribution**:

$$\mathbb{P}\left(Y_i = 1 | \mathbf{x}_i\right) = \Lambda\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right) = \frac{\exp\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right)}{1 + \exp\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right)}$$

  which is much easier to manipulate and handle computationally-wise.

In practice, the two distributions are very similar (they are both bell-shaped, although the logistic has fatter tails) and the two models usually produce similar sets of results which are easily compared against one another.

After having estimated a probit or a logit model, one must be careful at interpreting the estimates of $\boldsymbol{\beta}$! In fact, while the linear specification of the latent variable might induce some confusion, a coefficient β_k is neither the *causal effect* of X_{ik} on Y_i nor the predicted change in the *probability* to get $Y_i = 1$ following some unitary increase in variable X_{ik}. The best way to interpret the estimated parameters is by calculating the **marginal effects**. For all the explanatory variables in $\mathbf{x}_i$, these are characterized as:

$$\begin{aligned}\frac{\partial \mathbb{P}\left(Y_i = 1 | \mathbf{x}_i\right)}{\partial \mathbf{x}_i} &= \frac{\partial \mathbb{E}\left[Y_i | \mathbf{x}_i\right]}{\partial \mathbf{x}_i} \\ &= \frac{\partial F_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right)}{\partial \mathbf{x}_i} \\ &= f_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right) \boldsymbol{\beta}\end{aligned}$$

and they are a *function of the data* for any value of $\boldsymbol{\beta}$. There are two ways to calculate marginal effects that meaningful for interpretation's sake:

- to evaluate $f_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right) \boldsymbol{\beta}$ at $\widehat{\boldsymbol{\beta}}_{MLE}$ and at $\mathbf{x} = \bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$, the average characteristics in the sample;
- to evaluate $f_{\boldsymbol{x}}\left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right) \boldsymbol{\beta}$ at $\widehat{\boldsymbol{\beta}}_{MLE}$ and at $\mathbf{x}_i$ for every observation, and then average the resulting individual marginal effects over all observations.

It can be shown that these two approaches are asymptotically equivalent.

11.6 Simulated Maximum Estimation

There are instances such that the numerical evaluation of the M-Estimation criterion function (11.2) is so complicated as to make practical applications of the estimators discussed so far unfeasible. In such cases, theoretical and applied econometricians alike advocate the use of estimators that make use of **simulation methods** to approximate the evaluations in question. The leading techniques that make use of simulation techniques lie in the domain of MLE, and are adopted extensively in a subset of LDV models – those with so-called random coefficients – that are especially popular in some fields of economics such as Industrial Organization. For the sake of exposition, the following discussion starts from such particular cases of MLE and is later generalized to all M-Estimators.

Suppose that the probability mass or density function of all observable variables $\boldsymbol{x}_i$ can be written, given the model parameters $\boldsymbol{\theta}$, as follows:

$$f_{\boldsymbol{x}}\left(\mathbf{x}_i|\,\boldsymbol{\theta}\right)=\int_{\mathbb{U}} f_{\boldsymbol{x}|\boldsymbol{u}}\left(\mathbf{x}_i|\,\mathbf{u}_i;\boldsymbol{\theta}\right)dH_{\boldsymbol{u}}\left(\mathbf{u}_i\right) \tag{11.37}$$

where $\boldsymbol{u}_i$ is a random vector with cumulative distribution $H_{\boldsymbol{u}}\left(\mathbf{u}_i\right)$ that is integrated out over its support $\mathbb{U}$. Now, suppose that there is no closed form solution for the integral expressing (11.37), even if the conditional mass or density function $f_{\boldsymbol{x}|\boldsymbol{u}}\left(\mathbf{x}_i|\,\mathbf{u}_i;\boldsymbol{\theta}\right)$ is by itself tractable. It is obvious that the MLE problem based on (11.37) cannot be easily solved.

Example 11.13. Random coefficients logit. Consider the following bivariate logit model:

$$\mathbb{P}\left[Y_i=1|\,X_i\right]=\Lambda\left(\beta_0+\beta_{1i}X_i\right) \tag{11.38}$$

which has a look of a simple version – with one binary dependent variable $Y_i\in\{0,1\}$ and one possibly continuous independent variable X_i – of one of the LDV models discussed previously. Notice, however, that the parameter β_{1i} is *observation specific*: it is obvious that if there are only N observations available an econometrician cannot identify (let alone estimate) all the $N+1$ parameters implicitly expressed in (11.38). Yet, there are many real world applications where it is sensible to allow for variation in the individual of Y_i to X_i, that is to allow for *individual heterogeneity* in the regression slope. Such models are called **random coefficients** models; in particular, (11.38) is a (bivariate) random coefficients *logit*.

Random coefficients models are typically handled by assuming that the individual parameters themselves follow a probability distribution whose parameters can be estimated – so that it is possible to evaluate the extent of individual endogeneity. In the context of the current example, a typical assumption is for example that of normality:

$$\beta_{1i}\sim\mathcal{N}\left(\beta_1,\sigma^2\right) \tag{11.39}$$

where $\boldsymbol{\theta}=\left(\beta_0,\beta_1,\sigma^2\right)$ is the parameter set of interest. Notice that (11.38) and (11.39) together imply that the conditional probability mass function of Y_i is, by defining $u_i\equiv\left(\beta_{1i}-\beta_1\right)/\sigma$ and with $\phi\left(\cdot\right)$ being the probability density function of the normal distribution:

$$\begin{aligned}&f_{Y_i|X_i}\left(y_i|\,x_i;\beta_0,\beta_1,\sigma^2\right)=\\&=\int_{\mathbb{R}}\Lambda\left[\beta_0+\left(\beta_1+\sigma u_i\right)x_i\right]^{y_i}\left\{1-\Lambda\left[\beta_0+\left(\beta_1+\sigma u_i\right)x_i\right]\right\}^{1-y_i}\phi\left(u_i\right)du_i\end{aligned} \tag{11.40}$$

and that the resulting integral has no closed form solution. ■

Random coefficients LDV models exist in more complicated forms (like multiple regressors, multinomial outcomes, distributions of the latent error other than the logistic) than the one exposed in the previous example. All such models pose the econometric problem of how to evaluate the likelihood function resulting from integrals with no closed form solutions.[3] One brute force approach can be that of using numerical methods to evaluate the integrals for each value of the parameters as it is necessary in the optimization. However, this can be computationally overwhelming in practical settings. The alternative, which is more common, is that of *simulating* the values of (11.37) that are used to compute the likelihood function. Specifically, the typical method called **Direct Monte Carlo Sampling** consists of taking a sample $\{\mathbf{u}_s\}_{s=1}^{S}$ of S random draws of $\boldsymbol{u}_i$ from $H_{\boldsymbol{u}}(\mathbf{u}_i)$,[4] and constructing for each observation a **simulator** as the Monte Carlo estimate

$$\widehat{f}_{\boldsymbol{x},S}(\mathbf{x}_i|\,\boldsymbol{\theta}) = \frac{1}{S}\sum_{s=1}^{S}\widetilde{f}_{\boldsymbol{x}|\boldsymbol{u}}(\mathbf{x}_i|\,\mathbf{u}_s;\boldsymbol{\theta}) \tag{11.41}$$

where function $\widetilde{f}_{\boldsymbol{x}|\boldsymbol{u}}(\mathbf{x}_i|\,\mathbf{u}_s;\boldsymbol{\theta})$ is called a **subsimulator**. If the latter is an unbiased predictor of the true density of interest, that is:

$$\mathbb{E}\left[\widetilde{f}_{\boldsymbol{x}|\boldsymbol{u}}(\mathbf{x}_i|\,\mathbf{u}_s;\boldsymbol{\theta})\right] = f_{\boldsymbol{x}}(\mathbf{x}_i|\,\boldsymbol{\theta}) \tag{11.42}$$

where the expectation is taken over the support of $\boldsymbol{u}_i$, then by some suitable Law of Large Numbers:

$$\widehat{f}_{\boldsymbol{x}}(\mathbf{x}_i|\,\boldsymbol{\theta}) \xrightarrow{p} f_{\boldsymbol{x}}(\mathbf{x}_i|\,\boldsymbol{\theta}) \tag{11.43}$$

that is, the simulator is consistent insofar as $S \to \infty$. This approach allows to derive the **Maximum Simulated Likelihood** (MSL) estimator as:

$$\widehat{\boldsymbol{\theta}}_{MSL} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{N}\sum_{i=1}^{N}\log\widehat{f}_{\boldsymbol{x},S}(\mathbf{x}_i|\,\boldsymbol{\theta}) \tag{11.44}$$

where the summation on the right-hand side is based on simulators as in expression (11.41).

[3] Random coefficients *linear* models also exist, naturally. They are however typically easier to handle, as slope deviations like $u_i = (\beta_{1i} - \beta_1)/\sigma$ in Example 11.13 are subsumed into the error term, and multiple approaches to handle this case exist. In the MLE case, the inability to evaluate the likelihood function is a more fundamental problem.

[4] More precisely, computational techniques of this sort are based on "pseudo-random" draws. In the univariate case, sequences are typically drawn from the standard uniform distribution and projected back onto the support of interest through the quantile function of $H_{\boldsymbol{u}}(\mathbf{u}_i)$, where $|\boldsymbol{u}| = 1$. The logic is easily extended to the multivariate case.

Some practical considerations apply to the given MSL estimator. First, it is clearly easier to compute if the simulator is differentiable with respect to the parameter vector $\boldsymbol{\theta}$. It is easy to check that this is verified in, say, the MSL estimator arising from the problem of Example 11.13 where, given S draws $\{u_s\}_{s=1}^S$ from the standard normal, the subsimulator reads as follows.

$$\begin{aligned}\widetilde{f}_{Y_i,X_i|U_s}\left(y_i, x_i|\, u_s; \beta_0, \beta_1, \sigma^2\right) = \\ = \Lambda\left[\beta_0 + (\beta_1 + \sigma u_s)\, x_i\right]^{y_i} \left\{1 - \Lambda\left[\beta_0 + (\beta_1 + \sigma u_s)\, x_i\right]\right\}^{1-y_i}\end{aligned}$$

Second, it is more convenient to calculate every element of the summation on the right-hand side of (11.44) using the same draw $\{\mathbf{u}_s\}_{s=1}^S$: if the simulator is consistent, this allows to confine the statistical uncertainty of the simulator to noise in the observable, rather than simulated, variables. More generally, the asymptotic properties of the estimator clearly depend on the number of simulation draws S. In this regard, the following is a key result, originally provided by Gouriéroux and Monfort (1991), and adapted here for better consistency with the exposition from this and previous Lectures.

Theorem 11.4. Asymptotic Efficiency of Maximum Simulated Likelihood. *Suppose that the mass or density function $f_{\boldsymbol{x}}(\mathbf{x}_i|\,\boldsymbol{\theta})$ describing the model's data generation process meets the requirements of Theorem 6.18, and thus the corresponding "theoretical" MLE has a limiting distribution as per the statement of that Theorem. A SML estimator based on an unbiased subsimulator as in (11.42) is asymptotically equivalent to the "theoretical" MLE and it has the same limiting distribution:*

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_{SML} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left[\boldsymbol{I}\left(\boldsymbol{\theta}_0\right)\right]^{-1}\right)$$

if $S, N \to \infty$ (a condition that is sufficient for consistency) and $\sqrt{N}/S \to 0$.

Proof. (Outline.) Gouriéroux and Monfort (1991) work out the standard Taylor expansion of the First Order Conditions of (11.44) and detail on how it depends on two sources of noise: the one coming from the data $\{\mathbf{x}_i\}_{i=1}^N$ and the one due to the simulation draws $\{\mathbf{u}_s\}_{s=1}^S$. They show that the latter vanishes asymptotically if S grows at a rate higher than that of N. □

While this is an important result, it is not a panacea. In fact, for large datasets it implies that S be set at very large values, which at times could be unfeasible due to the high computational costs that this implies. If S is as good as finite then the MSL is inconsistent, because even if the simulator is unbiased, its logarithm is not, that is

$$\mathbb{E}\left[\log \widehat{f}_{\boldsymbol{x}|S}\left(\mathbf{x}_i|\,\boldsymbol{\theta}\right)\right] \neq \log f_{\boldsymbol{x}}\left(\mathbf{x}_i|\,\boldsymbol{\theta}\right)$$

by the (non-)properties of expectations. Following the work by Gouriéroux and Monfort (1991), it has been suggested to leverage a second-order Taylor expansion of $\log \widehat{f}_{\boldsymbol{x}|S}(\mathbf{x}_i|\,\boldsymbol{\theta})$ around $\log f_{\boldsymbol{x}}(\mathbf{x}_i|\,\boldsymbol{\theta})$, which writes as:

$$\log f_{\boldsymbol{x}}(\mathbf{x}_i|\,\boldsymbol{\theta}) \simeq \mathbb{E}\left[\log \widehat{f}_{\boldsymbol{x}|S}(\mathbf{x}_i|\,\boldsymbol{\theta})\right] + \frac{\mathbb{V}\mathrm{ar}\left[\left(\widehat{f}_{\boldsymbol{x}|S}(\mathbf{x}_i|\,\boldsymbol{\theta}) - f_{\boldsymbol{x}}(\mathbf{x}_i|\,\boldsymbol{\theta})\right)^2\right]}{2\left[f_{\boldsymbol{x}}(\mathbf{x}_i|\,\boldsymbol{\theta})\right]^2}$$

(where both moments are taken over the support of $\boldsymbol{u}_i$) in order to provide an approximated correction for the asymptotic bias which is due to a finite S. This lets define the **first-order asymptotic bias-corrected MSL** as:

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{BCMSL} &= \\ &= \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max} \sum_{i=1}^{N}\left[\log \widehat{f}_{\boldsymbol{x},S}(\mathbf{x}_i|\,\boldsymbol{\theta}) + \sum_{s=1}^{S}\frac{\left[\widetilde{f}_{\boldsymbol{x}|\boldsymbol{u}}(\mathbf{x}_i|\,\mathbf{u}_s;\boldsymbol{\theta}) - \widehat{f}_{\boldsymbol{x},S}(\mathbf{x}_i|\,\boldsymbol{\theta})\right]^2}{2S\left[\widehat{f}_{\boldsymbol{x},S}(\mathbf{x}_i|\,\boldsymbol{\theta})\right]^2}\right]\end{aligned} \tag{11.45}$$

given that the inner summation inside the brackets on the right-hand side is easily motivated as a consistent estimator of the second-order term of the above Taylor expansion for each observation $i = 1, \ldots, N$. Researchers shall consider this extended estimator if they are concerned about the size of S relative to N in a practical environment.

The theory of simulated M-Estimators extends beyond Maximum Likelihood: if a generic M-Estimator is defined in terms of an observation-specific criterion $q(\boldsymbol{x}_i;\boldsymbol{\theta})$ that is based upon integrals without closed form solution, a simulation approach is rendered necessary. A **Simulated M-Estimator** (SM), of which MSL is a special case, is defined as:

$$\widehat{\boldsymbol{\theta}}_{SM} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max}\,\frac{1}{N}\sum_{i=1}^{N}\widehat{q}_S(\mathbf{x}_i;\boldsymbol{\theta}) \tag{11.46}$$

where $\widehat{q}_S(\mathbf{x}_i;\boldsymbol{\theta}) = \frac{1}{S}\sum_{s=1}^{S}\widetilde{q}_s(\mathbf{x}_i,\mathbf{u}_s;\boldsymbol{\theta})$ is typically an average of subsimulators that are written as $\widetilde{q}_s(\mathbf{x}_i,\mathbf{u}_s;\boldsymbol{\theta})$ and are based upon pseudo-random draws of $\boldsymbol{u}_i$ in analogy with the SML case. All previous considerations about SML extend to this more general case too: the SM estimator is consistent if $S, N \to \infty$; furthermore it is as efficient as the corresponding non-simulated M-Estimator if $\sqrt{N}/S \to 0$; if S is too small, an approximated bias correction may be necessary. Clearly, all this applies so long as the conditions underpinning consistency and asymptotic normality of M-Estimators hold for the simulated estimator as they would in the standard case.

To conclude this brief overview of simulation-based M-Estimators, it is useful to make some remarks on how the asymptotic components $\boldsymbol{\Upsilon}_0$ and $\mathbf{Q}_0$ variance-covariance matrices are estimated. Clearly, standard formulae such as (11.17) and (11.19) are unfeasible because the elements of the summations involved cannot be evaluated. The solution is, unsurprisingly, that of simulating them. Specifically, a consistent estimator for $\boldsymbol{\Upsilon}_0$ is:

$$\widehat{\boldsymbol{\Upsilon}}_{M,S} = \frac{1}{N}\sum_{i=1}^{N} \widehat{\mathbf{s}}_{Si}\left(\mathbf{x}_i;\widehat{\boldsymbol{\theta}}_M\right)\widehat{\mathbf{s}}_{Si}^{\mathrm{T}}\left(\mathbf{x}_i,\widehat{\boldsymbol{\theta}}_M\right) \xrightarrow{p} \boldsymbol{\Upsilon}_0 \tag{11.47}$$

where $\widehat{\mathbf{s}}_{Si}\left(\mathbf{x}_i;\boldsymbol{\theta}\right) = \frac{\partial \widehat{q}_S(\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, while a consistent estimator for $\mathbf{Q}_0$ is:

$$\widehat{\mathbf{Q}}_N \equiv \frac{1}{N}\sum_{i=1}^{N} \widehat{\mathbf{H}}_{Si}\left(\mathbf{x}_i;\widehat{\boldsymbol{\theta}}_M\right) \xrightarrow{p} \mathbf{Q}_0 \tag{11.48}$$

where $\widehat{\mathbf{H}}_{Si}\left(\mathbf{x}_i;\boldsymbol{\theta}\right) = \frac{\widehat{\mathbf{s}}_{Si}(\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}} = \frac{\partial \widehat{q}_S(\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^{\mathrm{T}}}$. Expressions that extend (11.47) to the CCE or HAC cases can be derived. Notice that in the MSL case, the estimator of the information matrix (under i.i.d. observations) is typically obtained through the outer product of the gradients as:

$$\sum_{i=1}^{N}\left[\begin{array}{c} \dfrac{\sum_{s=1}^{S}\dfrac{\partial \widetilde{f}_{\boldsymbol{x}|\boldsymbol{u}}\left(\mathbf{x}_i|\,\mathbf{u}_s;\widehat{\boldsymbol{\theta}}_{SML}\right)}{\partial \boldsymbol{\theta}}}{\sum_{s=1}^{S}\widetilde{f}_{\boldsymbol{x}|\boldsymbol{u}}\left(\mathbf{x}_i|\,\mathbf{u}_s;\widehat{\boldsymbol{\theta}}_{SML}\right)} \dfrac{\sum_{s=1}^{S}\dfrac{\partial \widetilde{f}_{\boldsymbol{x}|\boldsymbol{u}}\left(\mathbf{x}_i|\,\mathbf{u}_s;\widehat{\boldsymbol{\theta}}_{SML}\right)}{\partial \boldsymbol{\theta}^{\mathrm{T}}}}{\sum_{s=1}^{S}\widetilde{f}_{\boldsymbol{x}|\boldsymbol{u}}\left(\mathbf{x}_i|\,\mathbf{u}_s;\widehat{\boldsymbol{\theta}}_{SML}\right)} \end{array}\right] \xrightarrow{p} \boldsymbol{I}\left(\boldsymbol{\theta}_0\right)$$

because of the particular mathematical properties of likelihood functions; a similar expression can also be derived in the Hessian's case.

11.7 Applications of Maximum Estimation

Maximum Estimators, and in particular Maximum Likelihood Estimators, are applied in all fields of economics. Especially when they are adapted from a specific structural economic model, these estimators are often unique, and it is difficult to provide a taxonomy (although some models – like certain LDV ones – can be applied in more diverse contexts). In order to illustrate this diversity, this section briefly delineates a few different applications of M-Estimation. This overview starts by describing the NLLS estimator for CES production functions, and subsequently provides a succinct summary of two different models based on MLE: Heckman's sample selection model and Bresnahan's test for collusion in oligopolistic industries.

Estimation of the CES Production Function

Lecture 7 briefly describes how the Cobb-Douglas production function, written as (7.42) in its simplest form (that is, with only two inputs: capital and labor), can be easily transformed to a *log-log* model, endowed with an error term, and estimated via OLS. This is quite a clean example of a structural econometric model! It turns out, however, that the Cobb-Douglas production function is a limiting case of the more general *Constant Elasticity of Substitution* (CES) production function, a fundamental ingredient of many economic models. In its simplest form this function writes as:

$$Y_i = [\alpha_K K_i + \alpha_L L_i]^{\frac{1}{\rho}} + \varepsilon_i \tag{11.49}$$

where Y_i is output, K_i and L_i are capital and labor, $\alpha_K > 0$ and $\alpha_L > 0$ are the respective so-called *saliency* parameters that determine the relative importance of each input, $\rho > 0$ is a parameter related to the *elasticity of substitution* between inputs, which as the model's name goes in this model is constant and writes $\sigma = (1+\rho)^{-1} \in (0,1)$, while ε_i is an error term. It can be shown that as $\rho \to 0$, (11.49) becomes a Cobb-Douglas production function like (7.42) where $\alpha_K = \beta_K$ and $\alpha_L = \beta_L$.

This model must obviously be estimated by NLLS via numerical methods, and even the simplest case (11.49) is known to entail complications. A typical estimation algorithm involves splitting the problem as follows:

$$(\widehat{\rho}, \widehat{\alpha}_K, \widehat{\alpha}_L)_{NLLS} \in \underset{\rho \in \mathbb{R}_{++}}{\arg\min} \left[\underset{(\alpha_K, \alpha_L) \in \mathbb{R}^2_{++}}{\arg\min} \sum_{i=1}^{N} \left(y_i - [\alpha_K k_i + \alpha_L l_i]^{\frac{1}{\rho}} \right)^2 \right]$$

where (y_i, k_i, l_i) denote observations of (Y_i, K_i, L_i). In words, numerical algorithms feature an inner maximizer of (α_K, α_L) given a value of ρ, and an outer maximizer for ρ; the parameter combination that minimizes the sum of squared residuals is ultimately selected. Unfortunately, the applied practice has shown that the solution is quite unstable and very dependent on the value of ρ, for this reason, practitioners often prefer to estimate a linear approximation of the model via OLS, see Kmenta (1967). The problem amplifies further for more complicated versions of the model (multiple inputs, nested CES structures) which explains why CES production functions are seldom encountered in the empirical practice.

The Heckit Model of Sample Selection

The determinants of individual *labor supply* decisions have always interested economists. In particular, due to women's lesser labor market participation

rates, *female labor supply* is of particular interest from a policy perspective. Consider an equation that describes the **intensity** of a woman's participation as a function of her individual characteristics $\mathbf{x}_i$:

$$h_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i \tag{11.50}$$

where h_i represents worked hours over a given time frame or other measures of labor supply (for example days or weeks in the case of seasonal jobs).

A problem is that h_i is only observed for women who *do* actually work:

$$z_i^* = \mathbf{w}_i^{\mathrm{T}} \boldsymbol{\gamma} + \upsilon_i \tag{11.51}$$

$$h_i \begin{cases} > 0 & \text{if } z_i^* > 0 \\ = 0 & \text{if } z_i^* \leq 0 \end{cases} \tag{11.52}$$

where there is some latent variable z_i^*, depending on a possibly different set of characteristics $\mathbf{w}_i$, which represents the individual cost-benefit evaluation on whether to work or not. The difference with binary outcome models is that in this case if an individual does participate to the labor market, as specified by the **participation** equation (11.51) and by the assignment rule (11.52), the intensity of her work is observed as a continuous variable. In fact, the ultimate objective of the researcher is to estimate a model such as (11.50) for the determinants of the intensity variable h_i, and not merely a binary outcome model for participation *per se*. Notice that this model could be alternatively specified for other intensity variables h_i such as the market wage for women; in other variations of this model interest may lie in both the quantity (hours) and the price (wage) variables.

Unfortunately, OLS cannot estimate (11.50) consistently. Denoting by H_i the random variable whence the observations of h_i are drawn, it is:

$$\begin{aligned} \mathbb{E}\left[H_i \middle| \mathbf{x}_i, h_i > 0 \right] &= \mathbb{E}\left[H_i \middle| \mathbf{x}_i, z_i^* > 0 \right] \\ &= \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \mathbb{E}\left[\varepsilon_i \middle| \mathbf{x}_i, z_i^* > 0 \right] \\ &= \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \mathbb{E}\left[\varepsilon_i \middle| \mathbf{x}_i, \upsilon_i > -\mathbf{w}_i^{\mathrm{T}} \boldsymbol{\gamma} \right] \end{aligned}$$

where $\lambda(\mathbf{w}_i) \equiv \mathbb{E}\left[\varepsilon_i \middle| \upsilon_i > -\mathbf{w}_i^{\mathrm{T}} \boldsymbol{\gamma} \right] \neq 0$ as long as the two error terms are correlated. The quantity $\lambda(\mathbf{w}_i)$ is in all effects an omitted variable of the equation, and represents the fact that individuals who are more inclined to work – or otherwise more favored by the circumstances to be able to work – will likely participate to the labor market at a higher intensity. Since this quantity is a function of $\mathbf{w}_i$, if some of the elements (variables) of vectors $\mathbf{x}_i$ and $\mathbf{w}_i$ are the same, we are in presence of an omitted variable bias type of problem: one that takes the well-known name of **sample selection bias**.

For example, a woman with a wealthy husband who is happy supporting her will be both less inclined to work *and* to work many hours if she works at all (the husband's income is an element of both $\mathbf{x}_i$ and $\mathbf{w}_i$). If, in addition to this, the natural inclinations of the woman in question are correlated for both the participation (υ_i) and the intensity (ε_i) decisions – a natural state of things – all the conditions for a sample selection bias are present.

Heckman (1977) devised a solution to this problem that was worth him the Nobel Prize in Economics. This solution, which is also known with the name of **heckit** in analogy with probit, logit and other models with LDV components, is based on a parametric assumption about the two error terms $(\varepsilon_i, \upsilon_i)$ such as a bivariate normal distribution:

$$\begin{pmatrix} \varepsilon_i \\ \upsilon_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

where ρ is the correlation coefficient between the two errors, σ^2 is the variance of the error of the intensity equation while the corresponding variance for the participation equation is normalized to 1 since it is not identified in binary LDV models. Thus, all the parameters of the model could be in principle estimated via MLE by specifying an appropriate likelihood function that accounts for the common dependence of the two equations.

However, Heckman also proposed an alternative procedure that is much easier to implement, while still requiring the bivariate normal assumption:

1. run a probit on the participation equation (11.52) and obtain $\widehat{\boldsymbol{\gamma}}_{MLE}$;
2. for each observation, calculate the **inverse Mills ratio**:

$$\lambda_i = \left[\frac{\phi\left(\mathbf{w}_i^{\mathrm{T}} \widehat{\boldsymbol{\gamma}}_{MLE}\right)}{\Phi\left(\mathbf{w}_i^{\mathrm{T}} \widehat{\boldsymbol{\gamma}}_{MLE}\right)} \right]$$

   where $\phi(\cdot)$ and $\Phi(\cdot)$ are, respectively, the density and cumulative functions of the standard normal distribution;
3. run OLS on a modified intensity equation:

$$h_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \rho \lambda_i + \varepsilon_i \tag{11.53}$$

   where the correlation parameter ρ is the OLS coefficient for λ_i.

Under the assumptions of the heckit model this procedure produces consistent estimates of $\boldsymbol{\beta}$. A disadvantage of this approach is that the standard errors of the OLS step are inconsistently estimated, because they do not account for the joint distribution of $(\varepsilon_i, \upsilon_i)$. However, "resampling" techniques such as the bootstrap can address this.

Detecting Collusion in Oligopolies

For decades, Industrial Organization has developed as a mainly theoretical field that analyzed, on the basis of strategic microeconomic analysis (mostly game theory) the implications of imperfectly competitive markets such as monopolies and oligopolies. Until 20-30 years ago there was a lamentable dearth of empirical studies in IO, as the econometric methods available were too simplistic with respect to the theoretical models developed in the field. With a new generation of structural models being introduced since the late '80s, the field has been revolutionized to the point that now it is mostly an empirical one, the "structural" field *par excellence*.

Before then, though, important empirical questions were left unsolved: a famous example was the sudden 45% increase in the US automobile production and sales (together with a corresponding prices fall) on 1955, with a rebound the following year. As demand was not that strong in 1955, most economists suspected the existence of a secret collusive agreement that for some reason broke apart in 1955 and was eventually resumed in 1956. For a long time they did not possess, however, any methodological tool to prove this hypothesis. In fact, Paul Samuelson had allegedly once said that he:

> *"would flunk any econometrics paper that claimed to provide an explanation of 1955 auto sales"*

a sentence that scared any economist who would think about actually trying to test the suspicion. Quoting Samuelson's words in the introduction of his paper, Bresnahan (1987) developed a methodology for detecting collusion in oligopolies that back then was quite innovative, becoming a starting point of the "empirical revolution" in IO. In fact, he was able to show statistically that hypotheses other than a momentary price war were unlikely.

Bresnahan models the automobile industry as one of N types of cars, each with quality $X_i = X(\boldsymbol{z}_i, \boldsymbol{\beta})$ being a function of one car's characteristics $\boldsymbol{z}_i$ given parameters $\boldsymbol{\beta}$. Qualities can be ordered from best to worst: without loss of generality, $X_i > X_h$ if $i > h$. He provides microfoundations for the **demand** functions of each car, defined for each year $t = 1, \ldots, T$ as

$$Q_{it}^D = D(P_{ht}, P_{it}, P_{jt}, X_{ht}, X_{it}, X_{jt}, \boldsymbol{\gamma}) \tag{11.54}$$

where Q_{it} is the quantity of product i, P_{it} its price, h, i, j are three **consecutive** products in the order of qualities and $\boldsymbol{\gamma}$ are some parameters. This specification makes prices and quantities only dependent, in equilibrium, on those of the "neighbors" of one product in the product space, and follows from a particular specification of consumers' utility.

As for **supply**, Bresnahan develops a standard framework of a profit-maximizing firm, whose profits from the sale of product i are:

$$\pi_{it} = P_{it}Q_{it} - c\left(X_{it}\right)Q_{it}$$

with $c\left(X_{it}\right) = \mu \exp\left(X_{it}\right)$; and he distinguishes the following two scenarios.

1. **Competition**: in this case each firm sets its own price P_{it} by taking the price of neighbors h and j as given, with First Order Conditions:

$$\frac{\partial \pi_{it}}{\partial P_{it}} = Q_{it} + \left(P_{it} - c\left(X_{it}\right)\right)\frac{\partial Q_{it}\left(\cdot\right)}{\partial P_{it}} = 0$$

   as Q_{it} is a function of P_{it} as per (11.54).

2. **Cooperation**: in this case the firm(s) selling two products, say, i and j would set prices P_{it} and P_{jt} so to maximize the joint profits, with First Order Conditions for the i-th price:

$$\frac{\partial\left[\pi_{it} + \pi_{jt}\right]}{\partial P_{it}} = Q_{it} + \left(P_{it} - c\left(X_{it}\right)\right)\frac{\partial Q_{it}\left(\cdot\right)}{\partial P_{it}} + \left(P_{jt} - c\left(X_{jt}\right)\right)\frac{\partial Q_{jt}\left(\cdot\right)}{\partial P_{it}} = 0$$

   and symmetrically for the j-th price.

Bresnahan then defines several matrices $\mathbf{H}_t$ such that, in each year,

$$h_{(ij)t} = \begin{cases} 1 & \text{cooperation between products } i \text{ and } j \\ 0 & \text{competition between products } i \text{ and } j \end{cases}$$

and characterizes several hypothetical scenarios for 1955 and surrounding years to which correspond associated sets of matrices $\mathbf{H}_t$. Thus, for a **given choice** of matrix $\mathbf{H}_t$ the supply function can be written as

$$q_{it}^S = S\left(P_{ht}, P_{it}, P_{jt}, X_{ht}, X_{it}, X_{jt}, \mathbf{H}_t, \boldsymbol{\gamma}, \mu\right) \tag{11.55}$$

where the demand function parameters $\boldsymbol{\gamma}$ enter via the derivative of the demand functions implied in the First Order Conditions.

By setting the equilibrium condition $Q_{it}^D = Q_{it}^S = Q_{it}^*$ (and similarly for prices) for each product $i = 1, \ldots, N$ in every year $t = 1, \ldots, T$, it is possible to obtain the **reduced form** of this model:

$$P_{it} = P^*\left(X_{ht}, X_{it}, X_{jt}, \mathbf{H}_t, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mu\right) \tag{11.56}$$

$$Q_{it} = Q^*\left(X_{ht}, X_{it}, X_{jt}, \mathbf{H}_t, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mu\right) \tag{11.57}$$

which is more easily obtained by solving for both the demand functions and the supply side First Order Conditions simultaneously. The last assumption is that the actual prices and quantities differ from their theoretical, reduced form values by a pair of normally distributed error terms:

$$\begin{pmatrix} P_{it} - P^* \\ Q_{it} - Q^* \end{pmatrix} = \begin{pmatrix} \xi_{it}^P \\ \xi_{it}^Q \end{pmatrix} = \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_P^2 & 0 \\ 0 & \sigma_Q^2 \end{pmatrix} \right)$$

where the variances of the two error terms reflect heteroscedasticity. Hence, the likelihood function can be written in terms of data realizations as:

$$\mathcal{L}\left(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mu | \, \mathbf{H}_t, \{p_{it}, q_{it}, \mathbf{z}_i\}_{i=1}^N\right) = \prod_{t=1}^{\mathrm{T}} \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_P^2}} \exp\left(-\frac{\left(\xi_{it}^P\right)^2}{2\sigma_P^2} \right) \times$$
$$\times \prod_{t=1}^{\mathrm{T}} \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_Q^2}} \exp\left(-\frac{\left(\xi_{it}^Q\right)^2}{2\sigma_Q^2} \right)$$

which can be estimated just by observing products' characteristics, quantities and prices in every year for a **given value** of $\mathbf{H}_t$.

Notice that in this model the structural parameters might not be separately identified, but this is not the objective the analysis: which is, in fact, to evaluate the performance of the model for two alternative choices of the "cooperation" matrix $\mathbf{H}_t$. Suppose that the hypothesis to be tested is:

$$\mathrm{H}_0 : \ \mathbf{H}_{t0} \text{ for competition} \qquad \mathrm{H}_1 : \ \mathbf{H}_{t1} \text{ for collusion}$$

where "competition" and " collusion" change over a set of products selected by the researcher. The test performed by Bresnahan to evaluate each scenario is of the likelihood ratio type:

$$\mathrm{C}_{\mathbf{H}} = 2\left[\log \mathcal{L}\left(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}}, \widetilde{\mu} \middle| \, \mathbf{H}_{t1}, \mathbf{z}_1, \ldots, \mathbf{z}_N\right) - \log \mathcal{L}\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \widehat{\mu} \middle| \, \mathbf{H}_{t0}, \mathbf{z}_1, \ldots, \mathbf{z}_N\right)\right]$$

the associated test statistic would reject the null if $\mathbf{H}_{t1}$ fits the data significantly better than $\mathbf{H}_{t0}$. Thanks to this procedure, Bresnahan has statistically shown that some car producers have been colluding in the US market in all years but 1955: Paul Samuelson must have not been happy.

It must be remarked that while the Bresnahan model is still a nice example of a structural model in IO that simultaneously incorporates both the demand and supply side within an elegant MLE framework, by today's standards it certainly feels antiquated and "mechanical." The current practice in Industrial Organization favors the use of random coefficients multinomial LDV models that incorporate the supply side while attempting to correct for endogeneity of prices and product characteristics through instrumental variables – all within a larger Generalized Method of Moments framework.

Lecture 12

Generalized Method of Moments

This lecture introduces the Generalized Method of Moments (GMM): an encompassing framework for estimating semi-parametric econometric models. To illustrate and motivate its applicability, GMM is shown to be a generalization of of standard IV estimators for linear models, and consequently how 2SLS and 3SLS are particular cases of this framework. Analogous considerations are then extended to non-linear models. This lecture also overviews some theoretical and practical issues such as methods for estimation of the variance-covariance and their implementation, tests for overidentification, and simulation-based approaches. Lastly, this lecture provides some examples about applications of GMM from actual economic research.

12.1 Generalizing the Method of Moments

The Method of Moments estimator introduced in Lecture 5, and motivated by the Analogy Principle, is suited to address many parameter estimation problems in Statistics. It turns out that also most econometric estimators examined thus far can be reformulated as Method of Moments estimators! Consider, for example, the following *zero moment conditions*:

$$\mathbb{E}\left[\boldsymbol{z}_i\left(Y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right)\right] = \mathbf{0} \tag{12.1}$$

where both $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$ are two random vectors of equal dimension (say K) and Y_i is another random variable; these conditions follow naturally from the exogeneity assumption in an IV setting, that is $\mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{z}_i\right] = 0$. From the above conditions one retrieve the parameter vector $\boldsymbol{\beta}_0$ as:

$$\boldsymbol{\beta}_0 = \mathbb{E}\left[\boldsymbol{z}_i\boldsymbol{x}_i^{\mathrm{T}}\right]^{-1}\mathbb{E}\left[\boldsymbol{z}_i Y_i\right] \tag{12.2}$$

whose sample analogue is precisely the IV estimator. By replacing $\boldsymbol{z}_i$ with $\boldsymbol{x}_i$ one obtains the standard OLS estimator instead. All M-Estimators can

be similarly formulated as Method of Moments estimators where the motivating zero moment conditions are the First Order Conditions of the population criterion maximization (11.1). In fact, the asymptotic properties of all these estimators are derived following an approach which mirrors that of Theorems 6.8 and 6.17, but extending it to possibly non i.i.d. data.[1]

Other econometric estimators, however, cannot be phrased as Method of Moments estimators: this framework, in fact, allows for a number of zero moment conditions equal to the dimensionality of the problem (the number of parameters K). While, clearly, fewer moment conditions than parameters makes for an unsolvable problem (that is, an unidentified model), some econometric estimators are *overidentified*: one can posit more moment conditions than parameters. These are situations in which "redundant information" is available to the econometrician for estimation purposes. This is, for example, the case of the 2SLS estimator that emerges if, say, the random vector $\boldsymbol{z}_i$ in (11.1) has dimension $J > K$. It turns out that the Method of Moments can be **generalized** to allow for **overidentification**. Moreover, the "restrictions" in excess[2] can be tested to "evaluate" their contribution to parameter identification.

To elaborate, suppose that a researcher postulates the validity of $J \geq K$ zero moment conditions described by a vector-valued function $\boldsymbol{g}(\cdot)$:

$$\mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right)\right] = \mathbf{0} \tag{12.3}$$

where $\boldsymbol{x}_i = (\boldsymbol{y}_i, \boldsymbol{z}_i)$ is the collection of all the variables of the model (both exogenous and endogenous) while $\boldsymbol{\theta}_0$ is the K-dimensional vector collecting the *true* values of the parameters. The sample analog of (12.3), motivated by the Analogy Principle, is the following J-dimensional vector.

$$\overline{\mathbf{g}}_N\left(\boldsymbol{\theta}\right) \equiv \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{g}\left(\mathbf{x}_i; \boldsymbol{\theta}\right) \tag{12.4}$$

The **Generalized Method of Moments** (GMM) estimator is defined as the minimizer of a quadratic form $\widehat{\mathcal{G}}_N\left(\boldsymbol{\theta}\right)$ based on these empirical moments:

$$\widehat{\boldsymbol{\theta}}_{GMM} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\min}\, \widehat{\mathcal{G}}_N\left(\boldsymbol{\theta}\right) = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\min}\, \overline{\mathbf{g}}_N^{\mathrm{T}}\left(\boldsymbol{\theta}\right) \mathbf{A}_N \overline{\mathbf{g}}_N\left(\boldsymbol{\theta}\right) \tag{12.5}$$

for *any full rank positive semi-definite* J-dimensional square matrix $\mathbf{A}_N$.

[1] This was already observed in Lecture 8, footnote 2 with reference to OLS. Recall that Theorems 6.8 and 6.17 are derived under the assumption of a random sample.

[2] In econometrics, the expression "restriction" is sometimes used to indicate a single moment conditions like (12.1). The reason for it is that in a SEM setting, such a condition is typically associated with an "exclusion restriction" of Z_i on Y_i. In principle, however, a "restriction" and a "moment condition" are two conceptually distinct notions.

Intuitively, the GMM estimator picks the value $\widehat{\boldsymbol{\theta}}_{GMM}$ that minimizes the distance of all the empirical moments from their expected "true" value (zero). Such a distance is measured as a quadratic form that employs matrix $\mathbf{A}_N$ in order to "weigh" the relative importance of different moments, as it is clarified later. Some important observations are in order.

1. The First Order Conditions of the problem are as follows:

$$\frac{\partial}{\partial \boldsymbol{\theta}} \overline{\mathbf{g}}_N^{\mathrm{T}} \left(\widehat{\boldsymbol{\theta}}_{GMM} \right) \cdot \mathbf{A}_N \cdot \overline{\mathbf{g}}_N \left(\widehat{\boldsymbol{\theta}}_{GMM} \right) = \mathbf{0} \tag{12.6}$$

   note that the term the pre-multiplies $\mathbf{A}_N$, the transposed Jacobian of the empirical moment conditions evaluated at the solution, is a $K \times J$ matrix. Given that an analytic solution is generally not available, the GMM estimator is typically obtained numerically.

2. The GMM estimator resembles – indeed, is – an M-Estimator. Define:

$$\mathcal{G}_0 (\boldsymbol{\theta}) \equiv \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left[\boldsymbol{g} (\boldsymbol{x}_i; \boldsymbol{\theta}) \right]^{\mathrm{T}} \mathbf{A}_0 \, \mathbb{E} \left[\boldsymbol{g} (\boldsymbol{x}_i; \boldsymbol{\theta}) \right] \geq 0 \tag{12.7}$$

   for some full rank positive semi-definite J-dimensional matrix $\mathbf{A}_0$ such that $\mathbf{A}_N \xrightarrow{p} \mathbf{A}_0$. One can easily see that the GMM objective function $\widehat{\mathcal{G}}_N (\boldsymbol{\theta})$ converges in probability to $\mathcal{G}_0 (\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$:

$$\widehat{\mathcal{G}}_N (\boldsymbol{\theta}) \xrightarrow{p} \mathcal{G}_0 (\boldsymbol{\theta})$$

   as in M-Estimators for $\widehat{\mathcal{G}}_N (\boldsymbol{\theta}) = -\widehat{\mathcal{Q}}_N (\boldsymbol{\theta})$ and $\mathcal{G}_0 (\boldsymbol{\theta}) = -\mathcal{Q}_0 (\boldsymbol{\theta})$.

3. The model is identified if the population criterion $\mathcal{G}_0 (\boldsymbol{\theta})$ has a unique (local) minimum, which must be equal to the true parameter $\boldsymbol{\theta}_0$, since $\mathcal{G}_0 (\boldsymbol{\theta}) \geq 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\mathcal{G}_0 (\boldsymbol{\theta}_0) = 0$ by (12.3) and (12.7) – that is, by construction. The minimization of (12.7) implies the following First Order Conditions:

$$\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left[\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{g}^{\mathrm{T}} (\boldsymbol{x}_i; \boldsymbol{\theta}_0) \right] \cdot \mathbf{A}_0 \cdot \mathbb{E} \left[\boldsymbol{g} (\boldsymbol{x}_i; \boldsymbol{\theta}_0) \right] = \mathbf{0} \tag{12.8}$$

   and one can see that a unique solution is obtained if the $J \times K$ matrix $\mathbf{G}_0$ which is defined as:

$$\mathbf{G}_0 \equiv \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left[\frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \boldsymbol{g} (\boldsymbol{x}_i; \boldsymbol{\theta}_0) \right] \tag{12.9}$$

   has full column rank K, as otherwise many combination of parameters are equally capable of minimizing $\mathcal{G}_0 (\boldsymbol{\theta})$. Note that under *identically distributed* observations, it is $\mathbf{G}_0 = \mathbb{E} \left[\frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \boldsymbol{g} (\boldsymbol{x}_i; \boldsymbol{\theta}_0) \right]$.

Under some fairly general conditions, GMM estimators are consistent and asymptotically normal. While these results could be adapted from the analysis of M-Estimators, it is worth to show them – especially asymptotic normality – via an alternative route. This allows to better highlight certain peculiar aspects of GMM while simultaneously circumventing issues about the calculation of Hessian matrices.

Theorem 12.1. Asymptotic Properties of GMM. *If a GMM estimator based on some zero moment conditions like (12.4) is identified and meets the uniform convergence requirements of M-Estimators from Theorem 11.2, it is consistent.*

$$\widehat{\boldsymbol{\theta}}_{GMM} = \min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \widehat{\mathcal{G}}_N(\boldsymbol{\theta}) \xrightarrow{p} \min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{G}_0(\boldsymbol{\theta}) = \boldsymbol{\theta}_0 \tag{12.10}$$

Furthermore, if conditions analogous to those from Theorem 11.3 are met, the GMM estimator is also asymptotically normal and its limiting variance presents the following sandwiched expression:

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left(\mathbf{G}_0^{\mathrm{T}}\mathbf{A}_0\mathbf{G}_0\right)^{-1}\mathbf{G}_0^{\mathrm{T}}\mathbf{A}_0\boldsymbol{\Omega}_0\mathbf{A}_0\mathbf{G}_0\left(\mathbf{G}_0^{\mathrm{T}}\mathbf{A}_0\mathbf{G}_0\right)^{-1}\right) \tag{12.11}$$

where $\boldsymbol{\Omega}_0$ is the following $J \times J$ limiting matrix.

$$\boldsymbol{\Omega}_0 \equiv \lim_{N\to\infty} \mathbb{V}\mathrm{ar}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \boldsymbol{g}(\boldsymbol{x}_i; \boldsymbol{\theta}_0)\right] \tag{12.12}$$

Note. Before proceeding with a "sketched" proof, it is useful to observe that like in similar cases, when the observations are *independent* it is:

$$\boldsymbol{\Omega}_0 = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\boldsymbol{g}(\boldsymbol{x}_i; \boldsymbol{\theta}_0)\,\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{x}_i; \boldsymbol{\theta}_0)\right] \tag{12.13}$$

simplifying further to $\boldsymbol{\Omega}_0 = \mathbb{E}\left[\boldsymbol{g}(\boldsymbol{x}_i; \boldsymbol{\theta}_0)\,\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{x}_i; \boldsymbol{\theta}_0)\right]$ when the observations are also *identically distributed*.

Proof. (Sketched.) A heuristic argument is useful here to show consistency. By some Weak Law of Large Numbers it must be that:

$$\frac{\partial}{\partial\boldsymbol{\theta}}\widehat{\mathcal{G}}_N(\boldsymbol{\theta}_0) = \frac{\partial}{\partial\boldsymbol{\theta}}\overline{\mathbf{g}}_N^{\mathrm{T}}(\boldsymbol{\theta}_0)\cdot\mathbf{A}_N\cdot\overline{\mathbf{g}}_N(\boldsymbol{\theta}_0) \xrightarrow{p} \mathbf{G}_0\mathbf{A}_0\cdot\mathbb{E}\left[\boldsymbol{g}(\boldsymbol{x}_i; \boldsymbol{\theta}_0)\right] = \mathbf{0}$$

since $\mathbb{E}\left[\boldsymbol{g}(\boldsymbol{x}_i; \boldsymbol{\theta}_0)\right] = \mathbf{0}$ by assumption; by the First Order Conditions (12.6) this implies that:

$$\widehat{\mathcal{G}}_N\left(\widehat{\boldsymbol{\theta}}_{GMM}\right) \xrightarrow{p} \widehat{\mathcal{G}}_N(\boldsymbol{\theta}_0) \tag{12.14}$$

which entails $\widehat{\boldsymbol{\theta}}_{GMM} \xrightarrow{p} \boldsymbol{\theta}_0$ if the model is identified ($\mathbf{G}_0$ is of full rank).

To show asymptotic normality, apply the Mean Value Theorem *directly* to the empirical moments (12.4); with very little manipulation:

$$\sqrt{N}\,\overline{\mathbf{g}}_N\left(\widehat{\boldsymbol{\theta}}_{GMM}\right) = \sqrt{N}\,\overline{\mathbf{g}}_N\left(\boldsymbol{\theta}_0\right) + \mathbf{G}_N\left(\widetilde{\boldsymbol{\theta}}_N\right)\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0\right) \quad (12.15)$$

where as usual $\widetilde{\boldsymbol{\theta}}_N$ is a convex combination of $\widehat{\boldsymbol{\theta}}_{GMM}$ and $\boldsymbol{\theta}_0$, while $\mathbf{G}_N\left(\boldsymbol{\theta}\right)$ is defined as:

$$\mathbf{G}_N\left(\boldsymbol{\theta}\right) \equiv \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\boldsymbol{g}\left(\mathbf{x}_i;\boldsymbol{\theta}\right)$$

in analogy with $\mathbf{G}_0$. Plugging (12.15) into the First Order Conditions (12.6) delivers the expression:

$$\mathbf{G}_N^{\mathrm{T}}\left(\widehat{\boldsymbol{\theta}}_{GMM}\right)\mathbf{A}_N\left[\sqrt{N}\,\overline{\mathbf{g}}_N\left(\boldsymbol{\theta}_0\right) + \mathbf{G}_N\left(\widetilde{\boldsymbol{\theta}}_N\right)\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0\right)\right] = \mathbf{0}$$

which can be manipulated so to return the following equation.

$$\begin{aligned}\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0\right) = -\left[\mathbf{G}_N^{\mathrm{T}}\left(\widehat{\boldsymbol{\theta}}_{GMM}\right)\mathbf{A}_N\mathbf{G}_N\left(\widetilde{\boldsymbol{\theta}}_N\right)\right]^{-1}\times \\ \times\,\mathbf{G}_N^{\mathrm{T}}\left(\widehat{\boldsymbol{\theta}}_{GMM}\right)\mathbf{A}_N\sqrt{N}\,\overline{\mathbf{g}}_N\left(\boldsymbol{\theta}_0\right)\end{aligned}$$

Since $\mathbf{G}_N\left(\widehat{\boldsymbol{\theta}}_{GMM}\right) \xrightarrow{p} \mathbf{G}_0$ and $\mathbf{G}_N\left(\widetilde{\boldsymbol{\theta}}_N\right) \xrightarrow{p} \mathbf{G}_0$ by consistency of GMM, if

$$\sqrt{N}\,\overline{\mathbf{g}}_N\left(\boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0},\boldsymbol{\Omega}_0\right)$$

that is if some Central Limit Theorem can be applied to the data at hand, these results can be combined via the Delta Method to deliver (12.11). □

As usual, the matrices in the limiting variance (12.11) are unknown and must be estimated. The asymptotic variance-covariance is calculated as:

$$\widehat{\mathbb{A}\mathrm{var}}\left(\widehat{\boldsymbol{\theta}}_{GMM}\right) = \frac{1}{N}\left(\widehat{\mathbf{G}}_N^{\mathrm{T}}\mathbf{A}_N\widehat{\mathbf{G}}_N\right)^{-1}\widehat{\mathbf{G}}_N^{\mathrm{T}}\mathbf{A}_N\widehat{\boldsymbol{\Omega}}_N\mathbf{A}_N\widehat{\mathbf{G}}_N\left(\widehat{\mathbf{G}}_N^{\mathrm{T}}\mathbf{A}_N\widehat{\mathbf{G}}_N\right)^{-1} \quad (12.16)$$

where $\widehat{\mathbf{G}}_N$ is a consistent estimator of $\mathbf{G}_0$:

$$\widehat{\mathbf{G}}_N = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\boldsymbol{g}\left(\mathbf{x}_i;\widehat{\boldsymbol{\theta}}_{GMM}\right) \xrightarrow{p} \mathbf{G}_0 \quad (12.17)$$

while the estimator of $\boldsymbol{\Omega}_0$, denoted as $\widehat{\boldsymbol{\Omega}}_N$, once again depends on the specific assumptions. For example, if the observations are independent:

$$\widehat{\boldsymbol{\Omega}}_N = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{g}\left(\mathbf{x}_i;\widehat{\boldsymbol{\theta}}_{GMM}\right)\boldsymbol{g}^{\mathrm{T}}\left(\mathbf{x}_i;\widehat{\boldsymbol{\theta}}_{GMM}\right) \xrightarrow{p} \boldsymbol{\Omega}_0 \quad (12.18)$$

whereas the following applies under clustering (HAC extensions also exist).

$$\widehat{\mathbf{\Omega}}_{CCE} = \frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{N_c}\sum_{j=1}^{N_c} \boldsymbol{g}_{ic}\left(\mathbf{x}_{ic};\widehat{\boldsymbol{\theta}}_{GMM}\right)\boldsymbol{g}_{jc}^{\mathrm{T}}\left(\mathbf{x}_{jc};\widehat{\boldsymbol{\theta}}_{GMM}\right) \xrightarrow{p} \mathbf{\Omega}_0 \quad (12.19)$$

It is apparent that the asymptotic variance of the GMM estimator $\widehat{\boldsymbol{\theta}}_{GMM}$ depends on the choice of the weighting matrix $\mathbf{A}_N$, as it converges to $\mathbf{A}_0$. A celebrated, yet *a posteriori* intuitive result in econometrics (Hansen, 1982) is the one that shows that the **most efficient** GMM estimator is the one for which the **optimal** "weighting matrix" of the moment conditions is:

$$\mathbf{A}_N = \widehat{\mathbf{\Omega}}_N^{-1} \quad (12.20)$$

where $\widehat{\mathbf{\Omega}}_N^{-1}$ is a matrix that converges in probability to the inverse of $\mathbf{\Omega}_0$.

$$\widehat{\mathbf{\Omega}}_N^{-1} \xrightarrow{p} \mathbf{\Omega}_0^{-1} = \mathbf{A}_0 \quad (12.21)$$

Under this circumstance, the limiting distribution of the GMM estimator reads in a conveniently simpler fashion.

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left(\mathbf{G}_0^{\mathrm{T}}\mathbf{\Omega}_0^{-1}\mathbf{G}_0\right)^{-1}\right) \quad (12.22)$$

One can demonstrate that the limiting variance in (12.22) is efficient by noting that the difference between the standard, (12.11) and the optimal variance-covariance matrices of GMM is given by:

$$\begin{aligned}\left(\mathbf{G}_0^{\mathrm{T}}\mathbf{A}_0\mathbf{G}_0\right)^{-1}\mathbf{G}_0^{\mathrm{T}}\mathbf{A}_0\mathbf{\Omega}_0\mathbf{A}_0\mathbf{G}_0\left(\mathbf{G}_0^{\mathrm{T}}\mathbf{A}_0\mathbf{G}_0\right)^{-1} - \left(\mathbf{G}_0^{\mathrm{T}}\mathbf{\Omega}_0^{-1}\mathbf{G}_0\right)^{-1} = \\ = \left(\mathbf{G}_0^{\mathrm{T}}\mathbf{A}_0\mathbf{G}_0\right)^{-1}\mathbf{G}_0^{\mathrm{T}}\mathbf{A}_0\mathbf{\Omega}_0^{\frac{1}{2}}\cdot\mathbf{M}_{\widetilde{\mathbf{G}}_0}\cdot\mathbf{\Omega}_0^{\frac{1}{2}}\mathbf{A}_0\mathbf{G}_0\left(\mathbf{G}_0^{\mathrm{T}}\mathbf{A}_0\mathbf{G}_0\right)^{-1}\end{aligned} \quad (12.23)$$

where, for $\widetilde{\mathbf{G}}_0 \equiv \mathbf{\Omega}_0^{-\frac{1}{2}}\mathbf{G}_0$, it is:

$$\mathbf{M}_{\widetilde{\mathbf{G}}_0} \equiv \mathbf{I} - \widetilde{\mathbf{G}}_0\left(\widetilde{\mathbf{G}}_0^{\mathrm{T}}\widetilde{\mathbf{G}}_0\right)^{-1}\widetilde{\mathbf{G}}_0^{\mathrm{T}}$$

which is a symmetric and idempotent matrix, hence the overall difference is a semi-definite positive matrix whatever $\mathbf{A}_0$ is. This observation is analogous to the proof of the Gauss-Markov Theorem for OLS, and in fact the statistical intuition is best given through a comparison with Generalized Least Squares (GLS): the most efficient linear estimator under heteroscedasticity. In GLS, observations are reweighted by the inverse of the variance of the respective error terms, as per the Weighted Least Squares formulation (8.43). Analogously, in the GMM problem the moment conditions are weighted by *the inverse* of their respective statistical variance: the *larger* the statistical variance of a single moment condition $g_j\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)$ – for $j = 1,\ldots,J$ – the *smaller* its contribution towards the GMM objective function.

A fundamental result associated with GMM is the following theorem, which was originally proved by Chamberlain (1987).

Theorem 12.2. Semi-Parametric Efficiency Bound of GMM. *If the moment conditions (12.3) hold, the GMM estimator derived through the optimal weighting matrix* $\mathbf{\Omega}_0^{-1}$ *hits the efficiency bound which applies to the class of all semi-parametric estimators of* $\boldsymbol{\theta}_0$.

Proof. (Outline.) The argument by Chamberlain proceeds as follows. Suppose that the data $\{\mathbf{x}_i\}_{i=1}^N$ are drawn from a *discrete* support of dimension D denoted as $\mathbb{X}_D = \{\boldsymbol{\chi}_1, \boldsymbol{\chi}_2, \ldots, \boldsymbol{\chi}_D\}$, where $\boldsymbol{\chi}_d$ for $d = 1, \ldots, D$ is a given point in the support. If the moment conditions (12.3) hold, it follows that:

$$\mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right)\right] = \frac{1}{D}\sum_{d=1}^{D} \boldsymbol{g}\left(\boldsymbol{\chi}_d; \boldsymbol{\theta}_0\right) p_d = \mathbf{0} \tag{12.24}$$

where p_d is the probability attached to the d-th element of $\mathbb{X}_D$. Thus, estimating $\boldsymbol{\theta}_0$ by optimal GMM is equivalent to solving a parametric maximum likelihood problem based on (12.24), meaning that the Cramér-Rao bound from mathematical probability (recall the discussions in Lectures 5, 6, 11) applies to this context, and the bound is obviously the limiting variance in (12.22) because no efficiency gains can be obtained with any other weighting matrices. In addition, Chamberlain shows that the result does not depend upon the granularity of $\mathbb{X}_D$ in a fundamental way, and that it approximately holds even when the data have a continuous support. □

Chamberlain's result is extremely powerful, since it provides a rationale for the practical use of GMM as the least-variance estimator under minimal semi-parametric working assumptions – especially as many econometric estimators can be rephrased as GMM estimators, including (as it is discussed later) 2SLS and 3SLS. An outstanding issue remains though, which is that matrix $\mathbf{\Omega}_0$ is not known *ex ante*, since it is a function of $\boldsymbol{\theta}_0$. To circumvent this, Hansen (1982) proposed the **two-step GMM estimation** procedure. The steps entailed in this method are illustrated next under the assumption of *independent observations*, although they apply more generally.

1. Obtain a first step estimate $\widehat{\boldsymbol{\theta}}_1$ with some arbitrary weighting matrix: usually the identity matrix $\mathbf{I}$, which implies that the GMM objective function reduces to $\widehat{\mathcal{G}}_1\left(\boldsymbol{\theta}\right) = \overline{\mathbf{g}}_N^{\mathrm{T}}\left(\boldsymbol{\theta}\right)\overline{\mathbf{g}}_N\left(\boldsymbol{\theta}\right)$. The resulting estimate $\widehat{\boldsymbol{\theta}}_1$ is consistent but inefficient; yet being consistent it allows to compute the following consistent estimator of $\mathbf{\Omega}_0$.

$$\widehat{\mathbf{\Omega}}_N = \frac{1}{N}\sum_{i=1}^{N}\left[\boldsymbol{g}\left(\mathbf{x}_i; \widehat{\boldsymbol{\theta}}_1\right)\boldsymbol{g}^{\mathrm{T}}\left(\mathbf{x}_i; \widehat{\boldsymbol{\theta}}_1\right)\right] \xrightarrow{p} \mathbf{\Omega}_0 \tag{12.25}$$

2. Obtain a second step, final GMM estimate $\widehat{\boldsymbol{\theta}}_{GMM} = \widehat{\boldsymbol{\theta}}_2$ by minimizing the objective function $\widehat{\mathcal{G}}_2(\boldsymbol{\theta}) = \overline{\mathbf{g}}_N^{\mathrm{T}}(\boldsymbol{\theta})\,\widehat{\boldsymbol{\Omega}}_N^{-1}\overline{\mathbf{g}}_N(\boldsymbol{\theta})$.

Finally, the estimation of the *asymptotic* variance is obtained as follows.

$$\widehat{\mathbb{A}\mathrm{var}}\left(\widehat{\boldsymbol{\theta}}_{GMM}\right) = \frac{1}{N}\left(\widehat{\mathbf{G}}_N^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}_N^{-1}\widehat{\mathbf{G}}_N\right)^{-1} \tag{12.26}$$

Possibly, $\widehat{\boldsymbol{\Omega}}_N$ can be re-evaluated at the final estimate $\widehat{\boldsymbol{\theta}}_{GMM} = \widehat{\boldsymbol{\theta}}_2$. While it remains perhaps the most popular estimation procedure to estimate GMM models, Hansen's two-step is not the only one – especially as simulations have shown that it is **biased in small samples**. The main available alternatives are the following two algorithms.

- The **iterated GMM estimation**: this is practically an "infinite steps" GMM estimation procedure. The idea is not to stop Hansen's algorithm at his second step, but to re-compute $\widehat{\boldsymbol{\Omega}}_N$ instead by making use of the second step estimates $\widehat{\boldsymbol{\theta}}_2$. Then, one would obtain a "third step" $\widehat{\boldsymbol{\theta}}_3$ vector of parameter estimates, re-compute $\widehat{\boldsymbol{\Omega}}_N$ once again and so forth – until convergence is achieved. This computationally demanding approach has been shown to be asymptotically equivalent to the two-steps procedure, although it may perform better in small samples.

- The **continuously updating GMM estimation** (**CUGMM**): the idea of this approach is to estimate the weighting matrix (which is a function of the parameters), jointly with the parameters themselves. The CUGMM estimator is computed as the minimizer of the objective function

$$\widehat{\mathcal{G}}_N(\boldsymbol{\theta}) = \left[\sum_{i=1}^{N}\boldsymbol{g}(\mathbf{x}_i;\boldsymbol{\theta})\right]^{\mathrm{T}}\left[\sum_{i=1}^{N}\boldsymbol{g}(\mathbf{x}_i;\boldsymbol{\theta})\,\boldsymbol{g}^{\mathrm{T}}(\mathbf{x}_i;\boldsymbol{\theta})\right]^{-1}\left[\sum_{i=1}^{N}\boldsymbol{g}(\mathbf{x}_i;\boldsymbol{\theta})\right]$$

  and takes its name from the fact that whenever $\widehat{\mathcal{G}}_N(\boldsymbol{\theta})$ is numerically optimized, the weight matrix changes at every iteration. This is shown in Monte Carlo simulations to perform better than the two-step procedure, and to make overidentification tests more reliable (Hansen et al., 1996). However, it is also very computationally demanding.

In both cases, the asymptotic variance of $\widehat{\boldsymbol{\theta}}_{GMM}$ is estimated via (12.26) by combining the estimate of $\boldsymbol{\Omega}_0$ obtained last together with an estimate of $\mathbf{G}_0$ as per (12.17). In practical applications, numerical optimization is typically inevitable for the implementation of all these procedures; the choice of the most appropriate optimization method is case-dependent and it is best left to the specific evaluation of the practitioner.

12.2 GMM and Instrumental Variables

In most applications, GMM estimators are based on so-called **conditional moment conditions**, which take the form:

$$\mathbb{E}\left[\boldsymbol{h}\left(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta}_0\right) \middle| \boldsymbol{z}_i\right] = \mathbf{0} \tag{12.27}$$

where $\boldsymbol{h}(\cdot)$ is a P-valued function and $\boldsymbol{z}_i$ is a vector of J **instrumental variables**. By the Law of Iterated Expectations, (12.27) delivers PJ moment conditions that are usable for estimation.

$$\mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right)\right] = \mathbb{E}\left[\boldsymbol{z}_i \otimes \boldsymbol{h}\left(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta}_0\right)\right] = \mathbf{0} \tag{12.28}$$

The discussion developed next shows that GMM estimators based on this class of moments encompass and generalize many common econometric estimators, including 2SLS, 3SLS and extensions of NLLS, such as Instrumental Variables Non-Linear Least Squares (IV-NLLS).

2SLS as a GMM Estimator

The initial motivation given for GMM is for addressing the scenario of "more instruments than parameters" (overidentification) in linear models. In such a situation, the moment conditions like (12.1) are based on linear functions $h\left(Y_i, \boldsymbol{x}_i; \boldsymbol{\theta}_0\right) = Y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ with $P = 1$ (hence the moments are J in total) and their sample analogs are as follows.

$$\overline{\mathbf{g}}_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i\left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right) = 0$$

The associated GMM estimator is:

$$\widehat{\boldsymbol{\beta}}_{GMM} = \underset{\boldsymbol{\beta}\in\mathbb{R}^K}{\arg\min} \left[\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i\left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right]^{\mathrm{T}} \mathbf{A}_N \left[\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i\left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right]$$

which clearly has an analytic solution. In fact, the First Order Conditions of the problem above are:

$$-2\left[\frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i\mathbf{z}_i^{\mathrm{T}}\right] \mathbf{A}_N \left[\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i\left(y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{GMM}\right)\right] = 0$$

therefore:

$$\widehat{\boldsymbol{\beta}}_{GMM} = \left[\left(\sum_{i=1}^{N} \mathbf{x}_i\mathbf{z}_i^{\mathrm{T}}\right)\mathbf{A}_N\left(\sum_{i=1}^{N} \mathbf{z}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]^{-1}\left(\sum_{i=1}^{N} \mathbf{x}_i\mathbf{z}_i^{\mathrm{T}}\right)\mathbf{A}_N\sum_{i=1}^{N} \mathbf{z}_i y_i \tag{12.29}$$

or, in compact matrix notation

$$\widehat{\boldsymbol{\beta}}_{GMM} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{Z}\mathbf{A}_N\mathbf{Z}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Z}\mathbf{A}_N\mathbf{Z}^{\mathrm{T}}\mathbf{y} \tag{12.30}$$

which already resembles the 2SLS estimator. Note, in fact, that if one were to choose the weighting matrix $\mathbf{A}_N$ as:

$$\widetilde{\mathbf{A}}_N = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i\mathbf{z}_i^{\mathrm{T}}\right)^{-1} = \left(\frac{1}{N}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}$$

this estimator would correspond exactly with the standard 2SLS estimator. The actual estimate of its variance would depend on the assumptions made by the researcher (standard heteroscedasticity, homoscedasticity, group dependence etc.) but would anyhow be easy relatable to the "long" expression of the GMM asymptotic variance-covariance (12.16).

The theory of GMM, however, allows for additional efficiency gains. In fact, if the weighting matrix $\mathbf{A}_N$ were chosen as the inverse of the estimated variance-covariance matrix of the moment conditions which is obtained under the assumption that *the observations are independent*:

$$\begin{aligned}\mathbf{A}_N = \widehat{\boldsymbol{\Omega}}_N^{-1} &= \left\{\widehat{\mathbb{A}\mathrm{var}}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{z}_i\left(y_i - \mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{GMM}\right)\right]\right\}^{-1} \\ &= \left(\frac{1}{N}\sum_{i=1}^{N}e_i^2\mathbf{z}_i\mathbf{z}_i^{\mathrm{T}}\right)^{-1} \\ &= \left(\frac{1}{N}\mathbf{Z}^{\mathrm{T}}\widehat{\mathbf{E}}_N\mathbf{Z}\right)^{-1}\end{aligned}$$

where $e_i \equiv y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{GMM}$ for $i = 1, \ldots, N$ and $\widehat{\mathbf{E}}_N$ is as in (10.68). The GMM estimator (12.30) would thus become:

$$\widehat{\boldsymbol{\beta}}_{GMM} = \left[\mathbf{X}^{\mathrm{T}}\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\widehat{\mathbf{E}}_N\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{X}\right]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\widehat{\mathbf{E}}_N\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{y} \tag{12.31}$$

which differs slightly from standard 2SLS. In fact, this kind of GMM estimation retrieves a generalized version (in the GLS sense) of the overidentified 2SLS estimator. To better appreciate this, consider the estimated asymptotic variance of (12.31):

$$\widehat{\mathbb{A}\mathrm{var}}\left[\widehat{\boldsymbol{\beta}}_{GMM}\right] = \frac{1}{N}\left[\mathbf{X}^{\mathrm{T}}\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\widehat{\mathbf{E}}_N\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{X}\right]^{-1}$$

which no longer takes a typical sandwiched form akin to (10.65). Thus, by a decomposition analogous to (12.23), linear GMM can be shown to be a more efficient estimator – yet likewise consistent – than standard 2SLS.

As a follow-up to these observations, one might wonder "how well" does the standard 2SLS fare, in terms of efficiency, relative to linear GMM. In the highly ideal case of *independent*, *identically distributed*, and *homoscedastic* observations, the probability limit of the optimal weighting matrix for linear GMM is as follows:

$$\mathbf{A}_N^{-1} = \widehat{\mathbf{\Omega}}_N = \frac{1}{N}\sum_{i=1}^{N} e_i^2 \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}} \xrightarrow{p} \sigma^2 \, \mathbb{E}\left[\boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}}\right]$$

where $\sigma^2 \equiv \mathbb{V}\mathrm{ar}\left[Y_i - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0\right]$ does not depend on $\boldsymbol{z}_i$; at the same time:

$$\widetilde{\mathbf{A}}_N^{-1} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}} \xrightarrow{p} \mathbb{E}\left[\boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}}\right]$$

so that the two estimators would asymptotically coincide: observe that σ^2 would simplify in the expression of the probability limit of (12.22). In less ideal scenarios, the equivalence collapses. Nevertheless, it has been observed that *for linear models*, the efficiency gains obtained thanks to the optimal variance GMM are small in comparison to the computational and empirical costs associated with its implementation – for example, if observations are dependent a different estimator of $\mathbf{\Omega}_0$ may be necessary in order to construct the optimal weighting matrix, and the resulting GMM estimator may be even less efficient than 2SLS if wrong choices are taken. It is no surprise, then, that current practice favors the use of the standard 2SLS estimator coupled with appropriate estimators of its variance – most typically, the heteroscedasticity- or the cluster-robust ones.

3SLS as a GMM Estimator

In analogy with 2SLS, also the 3SLS estimator for SEMs results from the solution of a GMM problem. I report here the SEM model that is analyzed in Lecture 10, already formulated in compact matrix notation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$$

or, more extensively:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_P \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}_P \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{10} \\ \boldsymbol{\beta}_{20} \\ \vdots \\ \boldsymbol{\beta}_{P0} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_P \end{bmatrix}$$

a model that has $N \times P$ dimension (N observations for P equations). Recall that each of the matrices $\mathbf{X}_p$ for $p = 1, \ldots, P$ combines all the endogenous *and* exogenous variables *not excluded* from the p-th equation. The moment conditions, however, single out the exogenous variables $\boldsymbol{z}_i$:

$$\mathbb{E}\left[\varepsilon_{pi} | \, \boldsymbol{z}_i\right] = 0 \Rightarrow \mathbb{E}\left[\boldsymbol{z}_i \varepsilon_{pi}\right] = \mathbf{0}$$

for all equations $p = 1, \ldots, P$. The sample analog of the moment conditions is in this case:

$$\overline{\mathbf{g}}_N\left(\boldsymbol{\beta}\right) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{z}_i \begin{pmatrix} y_{1i} - \mathbf{x}_{1i}^{\mathrm{T}} \boldsymbol{\beta}_1 \\ y_{2i} - \mathbf{x}_{2i}^{\mathrm{T}} \boldsymbol{\beta}_2 \\ \vdots \\ y_{Pi} - \mathbf{x}_{Pi}^{\mathrm{T}} \boldsymbol{\beta}_P \end{pmatrix} = \mathbf{0}$$

a PQ-dimensional vector (P equations for Q "exogenous" instruments). A more practical and elegant way to write these moment conditions is to use compact matrix notation:

$$\frac{1}{N}\left(\mathbf{I} \otimes \mathbf{Z}^{\mathrm{T}}\right)\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) = \mathbf{0}$$

where the Kronecker product operates by diagonally stacking the transpose of the exogenous variables' matrix $\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 & \mathbf{z}_2 & \ldots & \mathbf{z}_N \end{bmatrix}$ just P times.

$$\mathbf{I} \otimes \mathbf{Z}^{\mathrm{T}} = \begin{bmatrix} \mathbf{Z}^{\mathrm{T}} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}^{\mathrm{T}} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Z}^{\mathrm{T}} \end{bmatrix}$$

Notice that this Kronecker product has dimension $PQ \times PN$, where PN is also the length of the error terms vector $\boldsymbol{\varepsilon}$. In this case, the GMM problem is written as:

$$\widehat{\boldsymbol{\beta}}_{GMM} = \underset{\boldsymbol{\beta} \in \mathbb{R}^B}{\arg\min} \left[\frac{1}{N}\left(\mathbf{I} \otimes \mathbf{Z}^{\mathrm{T}}\right)\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)\right]^{\mathrm{T}} \mathbf{A}_N \left[\frac{1}{N}\left(\mathbf{I} \otimes \mathbf{Z}^{\mathrm{T}}\right)\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)\right]$$

where $B = |\boldsymbol{\beta}|$. The First Order Conditions are:

$$-2\left[\frac{1}{N}\mathbf{X}^{\mathrm{T}}\left(\mathbf{I} \otimes \mathbf{Z}\right)\right] \mathbf{A}_N \left[\frac{1}{N}\left(\mathbf{I} \otimes \mathbf{Z}^{\mathrm{T}}\right)\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{GMM}\right)\right] = \mathbf{0}$$

with the following "general" solution.

$$\widehat{\boldsymbol{\beta}}_{GMM} = \left[\mathbf{X}^{\mathrm{T}}\left(\mathbf{I} \otimes \mathbf{Z}\right) \mathbf{A}_N \left(\mathbf{I} \otimes \mathbf{Z}^{\mathrm{T}}\right) \mathbf{X}\right]^{-1} \mathbf{X}^{\mathrm{T}}\left(\mathbf{I} \otimes \mathbf{Z}\right) \mathbf{A}_N \left(\mathbf{I} \otimes \mathbf{Z}^{\mathrm{T}}\right) \mathbf{y} \tag{12.32}$$

The above estimator is equivalent to 3SLS if the weighting matrix is:

$$\begin{aligned}\widetilde{\mathbf{A}}_N &= \left[\left(\mathbf{I}\otimes\mathbf{Z}^{\mathrm{T}}\right)\left(\widehat{\boldsymbol{\Sigma}}_N\otimes\mathbf{I}\right)\left(\mathbf{I}\otimes\mathbf{Z}\right)\right]^{-1} \\ &= \left[\widehat{\boldsymbol{\Sigma}}_N\otimes\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right]^{-1} = \widehat{\boldsymbol{\Sigma}}_N^{-1}\otimes\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\end{aligned}$$

where $\widehat{\boldsymbol{\Sigma}}_N$ is still the estimate of the conditional error covariance matrix $\boldsymbol{\Sigma}$ as defined in (10.82), and it obtains from 2SLS estimates as per (10.83) in the previous discussion of the 3SLS estimator. Furthermore, consider that projection matrices are symmetric and idempotent, thus:

$$\begin{aligned}\left(\mathbf{I}\otimes\mathbf{Z}\right)\left[\widehat{\boldsymbol{\Sigma}}_N^{-1}\otimes\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\right]\left(\mathbf{I}\otimes\mathbf{Z}^{\mathrm{T}}\right) &= \widehat{\boldsymbol{\Sigma}}_N^{-1}\otimes\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}} \\ &= \left(\widehat{\boldsymbol{\Sigma}}_N^{-1}\otimes\mathbf{I}\right)\left(\mathbf{I}\otimes\mathbf{P}_{\mathbf{Z}}\right) \\ &= \left(\mathbf{I}\otimes\mathbf{P}_{\mathbf{Z}}\right)\left(\widehat{\boldsymbol{\Sigma}}_N^{-1}\otimes\mathbf{I}\right)\left(\mathbf{I}\otimes\mathbf{P}_{\mathbf{Z}}\right)\end{aligned}$$

where $\mathbf{P}_{\mathbf{Z}} = \mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}$. Therefore, plugging this weighting matrix $\widetilde{\mathbf{A}}_N$ into (12.32) yields the 3SLS estimator in (10.84) which is obtained under the hypotheses of within-equation homoscedasticity and cross-equation dependence; the expression of its variance is given in (10.85).

Further efficiency gains are obtained with an optimal weighting matrix. Under the hypothesis of *independent observations*, this is estimated as:

$$\begin{aligned}\mathbf{A}_N = \widehat{\boldsymbol{\Omega}}_N^{-1} &= \left\{\widehat{\mathbb{A}\mathrm{var}}\left[\frac{1}{\sqrt{N}}\left(\mathbf{I}\otimes\mathbf{Z}^{\mathrm{T}}\right)\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{GMM}\right)\right]\right\}^{-1} \\ &= \left(\frac{1}{N}\left(\mathbf{I}\otimes\mathbf{Z}^{\mathrm{T}}\right)\widehat{\mathbf{S}}_N\left(\mathbf{I}\otimes\mathbf{Z}\right)\right)^{-1}\end{aligned}$$

where $\widehat{\mathbf{S}}_N$ is an object analogous to $\widehat{\boldsymbol{\Sigma}}_N\otimes\mathbf{I}$, but slightly more complex:

$$\widehat{\mathbf{S}}_N = \begin{bmatrix}\widehat{\mathbf{S}}_{N,11} & \widehat{\mathbf{S}}_{N,12} & \cdots & \widehat{\mathbf{S}}_{N,1P} \\ \widehat{\mathbf{S}}_{N,21} & \widehat{\mathbf{S}}_{N,22} & \cdots & \widehat{\mathbf{S}}_{N,2P} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\mathbf{S}}_{N,P1} & \widehat{\mathbf{S}}_{N,P2} & \cdots & \widehat{\mathbf{S}}_{N,PP}\end{bmatrix}$$

and, for any $p, q = 1, \ldots, P$:

$$\widehat{\mathbf{S}}_{N,pq} = \begin{bmatrix}e_{pp1}^2 & 0 & \cdots & 0 \\ 0 & e_{pq2}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e_{pqN}^2\end{bmatrix}$$

with $e_{pqi} = \left(y_{pi}-\mathbf{x}_{pi}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{pGMM}\right)\left(y_{qi}-\mathbf{x}_{qi}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{qGMM}\right)$ for $i = 1, \ldots, N$.

Despite its algebraic complexity, this estimate of the variance-covariance of the moment conditions has a straightforward interpretation: in addition to allowing for cross-equation correlation (like standard 3SLS) it is robust to heteroscedasticity in the within-equation variances as well as in the cross-equation covariances. The resulting GMM estimator is:

$$\widehat{\boldsymbol{\beta}}_{GMM} = \left\{ \mathbf{X}^{\mathrm{T}} \left(\mathbf{I} \otimes \mathbf{Z} \right) \left[\left(\mathbf{I} \otimes \mathbf{Z}^{\mathrm{T}} \right) \widehat{\mathbf{S}}_N \left(\mathbf{I} \otimes \mathbf{Z} \right) \right]^{-1} \left(\mathbf{I} \otimes \mathbf{Z}^{\mathrm{T}} \right) \mathbf{X} \right\}^{-1} \times \\ \times \mathbf{X}^{\mathrm{T}} \left(\mathbf{I} \otimes \mathbf{Z} \right) \left[\left(\mathbf{I} \otimes \mathbf{Z}^{\mathrm{T}} \right) \widehat{\mathbf{S}}_N \left(\mathbf{I} \otimes \mathbf{Z} \right) \right]^{-1} \left(\mathbf{I} \otimes \mathbf{Z}^{\mathrm{T}} \right) \mathbf{y} \quad (12.33)$$

with estimated asymptotic variance:

$$\widehat{\mathbb{A}\mathrm{var}} \left[\widehat{\boldsymbol{\beta}}_{GMM} \right] = \frac{1}{N} \left\{ \mathbf{X}^{\mathrm{T}} \left(\mathbf{I} \otimes \mathbf{Z} \right) \left[\left(\mathbf{I} \otimes \mathbf{Z}^{\mathrm{T}} \right) \widehat{\mathbf{S}}_N \left(\mathbf{I} \otimes \mathbf{Z} \right) \right]^{-1} \left(\mathbf{I} \otimes \mathbf{Z}^{\mathrm{T}} \right) \mathbf{X} \right\}^{-1}$$

and it is asymptotically identical to standard 3SLS under homoscedasticity in both the within-equation variances and in the cross-equation covariances. Just like the GMM version of 2SLS, this estimator is seldom used; however it is of theoretical importance as the GMM generalization of the main full information semi-parametric estimator for simultaneous equations.

Instrumental Variables in Non-Linear Models

The GMM framework is well suited to extend the Instrumental Variables approach also to non-linear estimators like, for example, generalizations of the MLE score that allow for "exogenous" instruments. The general GMM problem associated with the moment conditions (12.28) is:

$$\widehat{\boldsymbol{\theta}}_{GMM} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\min} \left[\frac{1}{N} \sum_{i=1}^{N} \mathbf{z}_i \cdot \boldsymbol{h} \left(\mathbf{y}_i, \mathbf{z}_i; \boldsymbol{\theta} \right) \right]^{\mathrm{T}} \mathbf{A}_N \left[\frac{1}{N} \sum_{i=1}^{N} \mathbf{z}_i \cdot \boldsymbol{h} \left(\mathbf{y}_i, \mathbf{z}_i; \boldsymbol{\theta} \right) \right]$$

which in general lacks an explicit solution. However, an expression for both the limiting and asymptotic variances is easily obtained from (12.11) and (12.16) by noting that, in this case:

$$\mathbf{G}_0 = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left[\boldsymbol{z}_i \cdot \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{h}^{\mathrm{T}} \left(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta}_0 \right) \right]$$

and similarly for $\mathbf{G}_N \left(\boldsymbol{\theta} \right)$. A typical application of this is in single-equation non-linear models where:

$$\mathbb{E} \left[\frac{\partial}{\partial \boldsymbol{\theta}} h \left(\mathbf{x}_i; \boldsymbol{\theta}_0 \right) \cdot \left(y_i - h \left(\mathbf{x}_i; \boldsymbol{\theta}_0 \right) \right) \right] \neq \mathbf{0}$$

that is, the error term $\varepsilon_i = y_i - h(\mathbf{x}_i; \boldsymbol{\theta}_0)$ is not mean-independent of the implicit set of instruments that is defined by the Non-Linear Least Squares estimator – the K-dimensional vector of derivatives of $h(\cdot)$ with respect to the parameters $\boldsymbol{\theta}$, itself a function of the explanatory variables $\boldsymbol{x}_i$ – because of some type of endogeneity problem (just as for linear models).

Thus, the solution to the problem would be to look for a J-dimensional vector of instrumental variables $\boldsymbol{z}_i$ such that:

$$\mathbb{E}\left[\mathbf{z}_i \cdot \boldsymbol{h}\left(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta}\right)\right] = \mathbb{E}\left[\boldsymbol{z}_i \left(Y_i - h\left(\boldsymbol{x}_i; \boldsymbol{\theta}_0\right)\right)\right] = \mathbf{0} \tag{12.34}$$

and the **Non-Linear Two-Stages Least Squares** (NL2SLS) estimator which follows from the solution of this GMM problem crucially depends on the choice of $\mathbf{A}_N$. Its limiting variance is:

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_{NL2SLS} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left(\mathbf{J}_0^{\mathrm{T}}\mathbf{A}_0\mathbf{J}_0\right)^{-1}\mathbf{J}_0^{\mathrm{T}}\mathbf{A}_0\boldsymbol{\Omega}_0\mathbf{A}_0\mathbf{J}_0\left(\mathbf{J}_0^{\mathrm{T}}\mathbf{A}_0\mathbf{J}_0\right)^{-1}\right) \tag{12.35}$$

where $\mathbf{J}_0$, the analogue of $\mathbf{G}_0$, is defined as the following probability limit:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i\mathbf{h}_{0i}^{\mathrm{T}} \xrightarrow{p} \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\boldsymbol{z}_i\boldsymbol{h}_{0i}^{\mathrm{T}}\right] \equiv \mathbf{J}_0$$

(where $\mathbf{J}_0 = \mathbb{E}\left[\boldsymbol{z}_i\boldsymbol{h}_{0i}^{\mathrm{T}}\right]$ if the observations are identically distributed), $\boldsymbol{h}_{0i}$ is as in Examples 11.4 and 11.7, while $\mathbf{A}_0$ and $\boldsymbol{\Omega}_0$ are as in the standard theory of GMM. The estimated asymptotic variance of the NL2SLS estimator is:

$$\widehat{\mathbb{A}\mathrm{var}}\left[\widehat{\boldsymbol{\theta}}_{NL2SLS}\right] = \frac{1}{N}\left(\widehat{\mathbf{J}}_N^{\mathrm{T}}\mathbf{A}_N\widehat{\mathbf{J}}_N\right)^{-1}\widehat{\mathbf{J}}_N^{\mathrm{T}}\mathbf{A}_N\widehat{\boldsymbol{\Omega}}_N\mathbf{A}_N\widehat{\mathbf{J}}_N\left(\widehat{\mathbf{J}}_N^{\mathrm{T}}\mathbf{A}_N\widehat{\mathbf{J}}_N\right)^{-1}$$

where:

$$\widehat{\mathbf{J}}_N = \frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i\widehat{\mathbf{h}}_i^{\mathrm{T}}$$

and $\widehat{\mathbf{h}}_i$ is as in Example 11.7. The particular choice of the weighting matrix entails the same considerations as in the case of "linear" GMM-2SLS:

1. choosing $\mathbf{A}_N = N\left(\sum_{i=1}^{N}\mathbf{z}_i\mathbf{z}_i^{\mathrm{T}}\right)^{-1}$ delivers the traditional (standard) version of the NL2SLS estimator, which is akin to standard 2SLS;
2. with independent observations, $\widehat{\boldsymbol{\Omega}}_N = N\left(\sum_{i=1}^{N}e_i^2\mathbf{z}_i\mathbf{z}_i^{\mathrm{T}}\right)^{-1}$ where e_i is the usual residual for each i-th observation;
3. in such a case, however, the optimal NL2SLS estimator is obtained by setting $\mathbf{A}_N = \widehat{\boldsymbol{\Omega}}_N^{-1}$, and it is asymptotically equivalent to standard NL2SLS only under homoscedasticity.

Note that if points 1. and 2. above are maintained *and*, *in addition*, in the moment conditions (12.34) one specifies $\boldsymbol{z}_i = \boldsymbol{h}_{0i}$ – so that the instruments enter as $\mathbf{z}_i = \widehat{\mathbf{h}}_i$ in the estimation problem – the GMM returns the standard NLLS estimator which is examined in Lecture 11. This is clearly analogous to setting $\boldsymbol{z}_i = \boldsymbol{x}_i$ in linear models, which returns standard OLS.

Optimal Instruments

When operationalized via GMM, conditional moment conditions like (12.27) constitute a framework for the semi-parametric estimation of a wide class of econometric models: in fact, the theory for instrumental variables in non-linear models can be extended to non-linear *systems* of structural equations as well, similarly as how 3SLS generalizes 2SLS. However, conditional moment conditions are even more general, since any function of the instruments $\boldsymbol{l}(\boldsymbol{z}_i)$ which takes values upon a J'-dimensional set makes for valid moment conditions of the kind:

$$\mathbb{E}\left[\boldsymbol{l}(\boldsymbol{z}_i) \otimes \boldsymbol{h}(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta}_0)\right] = \mathbf{0}$$

so long as $PJ' \geq K$, where K is the total number of parameters. A relevant question is to what extent it is possible to construct appropriate **optimal instruments** so that the resulting GMM problem delivers the most efficient estimate available with the information enclosed in the conditional moment conditions (12.27). A result proved by several authors is that this objective is achieved through the $K \times P$ matrix $\mathbf{L}(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta}_0)$ defined as:

$$\mathbf{L}(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta}_0) = \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{h}^{\mathrm{T}}(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta}_0) \middle| \boldsymbol{z}_i\right] \left\{\mathbb{V}\mathrm{ar}\left[\boldsymbol{h}(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta}_0) \middle| \boldsymbol{z}_i\right]\right\}^{-1} \tag{12.36}$$

where the first term (the conditional expectation) is a $K \times P$ matrix, while the second term (the inverted conditional variance) is a $P \times P$ matrix. The efficient estimate of $\boldsymbol{\theta}_0$ is then obtained through the following K "optimal" moment conditions:

$$\mathbb{E}\left[\boldsymbol{g}(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta}_0)\right] = \mathbb{E}\left[\mathbf{L}(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta}_0) \cdot \boldsymbol{h}(\boldsymbol{y}_i, \boldsymbol{z}_i; \boldsymbol{\theta}_0)\right] = \mathbf{0}$$

and the corresponding estimate $\widehat{\boldsymbol{\theta}}_{MM}$ solves a simple Method of Moments sample analog system of equations.

$$\frac{1}{N}\sum_{i=1}^{N}\left[\mathbf{L}\left(\mathbf{y}_i, \mathbf{z}_i; \widehat{\boldsymbol{\theta}}_{MM}\right) \cdot \boldsymbol{h}\left(\mathbf{y}_i, \mathbf{z}_i; \widehat{\boldsymbol{\theta}}_{MM}\right)\right] = \mathbf{0}$$

Clearly, the limiting variance for this estimator assumes a "sandwiched" expression which is simpler than (12.16): as the problem is exactly identified, the weighting matrix is redundant.

Unfortunately, the practical use of the optimal instruments defined in (12.36) requires either *a priori* knowledge or the formulation of assumptions about specific moments conditional on the instruments, and any mistakes would jeopardize the entire approach. This is analogous to the problem of specifying the variance of the error term conditional on the regressors in linear models in order to make GLS "feasible." In fact, the parallel is more than just intuitive: if $\boldsymbol{z}_i = \boldsymbol{x}_i$, $P = 1$ and $\boldsymbol{h}(Y_i, \boldsymbol{x}_i; \boldsymbol{\beta}_0) = Y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0$, then:

$$\mathbf{L}_{GLS}(Y_i, \boldsymbol{x}_i; \boldsymbol{\beta}_0) = -\boldsymbol{x}_i \cdot \frac{1}{\sigma_L^2(\boldsymbol{x}_i)}$$

and the resulting Method of Moments estimator is just GLS! Similarly, if $\boldsymbol{z}_i = \boldsymbol{x}_i$, $P = 1$ but the model is non-linear: $\boldsymbol{h}(Y_i, \boldsymbol{x}_i; \boldsymbol{\theta}_0) = Y_i - h(\boldsymbol{x}_i; \boldsymbol{\theta}_0)$:

$$\mathbf{L}_{GNLLS}(Y_i, \boldsymbol{x}_i; \boldsymbol{\theta}_0) = -\boldsymbol{h}_{0i} \cdot \frac{1}{\sigma_{NL}^2(\boldsymbol{x}_i)}$$

which yields the "Generalized" version of Non-Linear Least Squares instead.[3] Because of the complications entailed in specifying the conditional moments that compose $\mathbf{L}(Y_i, \boldsymbol{x}_i; \boldsymbol{\theta}_0)$, in more general cases the current practice favors using moment conditions based on the simple vector of instruments $\boldsymbol{z}_i$.

12.3 Testing Overidentification

While discussing identification in the context of SEMs (Lecture 9), it was already mentioned how it is possible to statistically *test* the conditions that give rise to overidentification. The idea is that, when multiple restrictions allow to identify a single parameter, a researcher may think about holding one of them "constant" and verify, by performing a formal statistical test, how likely the other ones are in probabilistic terms, given the estimated parameters. Thus, if the statistical test is not favorable to the null hypothesis associated with the additional identifying conditions, the researcher may consider revising the model. It turns out that overidentification tests are well integrated in the GMM framework. In what follows the most common of such tests, the **Sargan-Hansen test**, is described briefly. The starting point is the following "Hansen J" statistic, which is a quadratic form of the estimated J moment conditions.

$$\mathrm{J}\left(\widehat{\boldsymbol{\theta}}_{GMM}\right) = N\, \overline{\mathbf{g}}_N^{\mathrm{T}}\left(\widehat{\boldsymbol{\theta}}_{GMM}\right) \widehat{\boldsymbol{\Omega}}_N^{-1} \overline{\mathbf{g}}_N\left(\widehat{\boldsymbol{\theta}}_{GMM}\right) \xrightarrow{d} \chi^2_{J-K} \tag{12.37}$$

[3] Here $\sigma_L^2(\boldsymbol{x}_i) = \mathbb{V}\mathrm{ar}\left[Y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 \middle| \boldsymbol{x}_i\right]$ and $\sigma_{NL}^2(\boldsymbol{x}_i) = \mathbb{V}\mathrm{ar}\left[Y_i - h(\boldsymbol{x}_i; \boldsymbol{\theta}_0) \middle| \boldsymbol{x}_i\right]$ are the two conditional variances of the two models' error terms. Clearly, under homoscedasticity the two MM estimators result in OLS and standard NLLS, respectively.

Observe that when the GMM estimation is performed with an optimal weighting matrix $\mathbf{A}_N$, the Hansen J-statistic corresponds to the estimated GMM objective function evaluated at $\widehat{\boldsymbol{\theta}}_{GMM}$ and multiplied by N. Under standard conditions this statistic is asymptotically χ^2_{J-K} distributed with $J-K$ degrees of freedom, where K degrees are subtracted to account for the fact that $\widehat{\boldsymbol{\theta}}_{GMM}$ has been estimated. The intuition about this test statistic is that if the moment conditions accurately describe the real world, it follows that empirically their sample analogues, when evaluated at $\widehat{\boldsymbol{\theta}}_{GMM}$, should be close to zero. The test evaluates exactly this hypothesis, taking care of normalizing the sample moment conditions by their empirically observed variance. Hence, if the measured Hansen J-statistic following some GMM estimation is too large relative to some critical value, the test may lead to rejecting the null hypothesis that all moments are zero in the population.

The Hansen J-statistic is mostly useful when there are fewer moment conditions than parameters. In such a case, if the null hypothesis is rejected, the researcher might selectively remove moment conditions and evaluate the consequent performance of the modified model. However, when the J moment conditions exceed the number of parameters K by a substantial amount, one might be interested in testing a subset of moment conditions *in block*. This task is performed by the so-called "incremental" Sargan test. In order to characterize this test, suppose one can divide the moment conditions in two subsets:

$$\mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right]=\mathbb{E}\begin{bmatrix}\boldsymbol{g}_1\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\\ \boldsymbol{g}_2\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\end{bmatrix}=0$$

where $\left|\boldsymbol{g}_1\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right|=J_1>K$, $\left|\boldsymbol{g}_2\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right|=J_2$, and $J_1+J_2=J$. Suppose that $\mathbb{E}\left[\boldsymbol{g}_1\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right]=0$ can be confidently believed, because of either prior knowledge or previous testing. Instead, one is interested about testing the null hypothesis $\mathrm{H}_0:\ \mathbb{E}\left[\boldsymbol{g}_2\left(\boldsymbol{x}_i;\boldsymbol{\theta}_0\right)\right]=0$.

The "incremental" **Sargan statistic** is defined as follows:

$$\mathrm{J}_S\left(\widehat{\boldsymbol{\theta}}_{GMM},\widetilde{\boldsymbol{\theta}}\right)=\mathrm{J}\left(\widehat{\boldsymbol{\theta}}_{GMM}\right)-\widetilde{\mathrm{J}}\left(\widetilde{\boldsymbol{\theta}}\right)\xrightarrow{d}\chi^2_{J_2} \tag{12.38}$$

for:

$$\widetilde{\mathrm{J}}\left(\widetilde{\boldsymbol{\theta}}\right)=N\overline{\mathbf{g}}^{\mathrm{T}}_{N1}\left(\widetilde{\boldsymbol{\theta}}\right)\widehat{\boldsymbol{\Omega}}^{-1}_{N_1}\overline{\mathbf{g}}_{N1}\left(\widetilde{\boldsymbol{\theta}}\right)\xrightarrow{d}\chi^2_{J_1-K}$$

and:

$$\widetilde{\boldsymbol{\theta}}=\underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\min}\,\overline{\mathbf{g}}^{\mathrm{T}}_{N1}\left(\boldsymbol{\theta}\right)\widehat{\boldsymbol{\Omega}}^{-1}_{N_1}\overline{\mathbf{g}}_{N1}\left(\boldsymbol{\theta}\right)$$

where here $\overline{\mathbf{g}}_{N1}\left(\boldsymbol{\theta}\right)$ is the sample mean of $\boldsymbol{g}_1\left(\mathbf{x}_i;\boldsymbol{\theta}\right)$, $\widehat{\boldsymbol{\Omega}}_{N_1}$ is some consistent estimate of its variance-covariance matrix,[4] and $\mathrm{J}\left(\cdot\right)$ is Hansen's J-statistic.

[4]It may well be the J_1-dimensional upper left square block of $\widehat{\boldsymbol{\Omega}}_N$.

In practice, the incremental Sargan test results from subtracting from the original Hansen J-statistics another Hansen J-statistic, where the latter is obtained from a "reduced" GMM model of J_1 "certain" moment conditions. This second Hansen J-statistic is always smaller by construction, because there are fewer moment conditions to match the zero vector. Therefore, the incremental Sargan test measures how much the other J_2 conditions deviate from zero once the J_1 "certain" ones are held constant. It is apparent how the intuition behind this test presents many analogies with the Distance or "Likelihood Ratio" test from the Trinity of statistical tests.

Example 12.1. Overidentified Mincer Equation. Return once more to the recurring example about the Mincer Equation:

$$\log W_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 S_i + \alpha_i + \epsilon_i$$

however, suppose that there are now three potential instruments available. The first is Z_i, the already mentioned "distance from college" instrument by Card. The second is G_i, representing past eligibility to some "fellowship grant" for attending higher education programs. For example, G_i might be motivated on some random (*exogenous*) past allocation of scholarship grants by the government authorities. Clearly, in this case the eligible individuals had obtained an advantage at the time of deciding whether or not to enroll in college. This, however, did not likely affect their future wages other than via better education (*exclusion restriction*). The last instrument is F_i, the average education of one individual's friends or close social network. One may argue that one's friends might have affected the individual decision on whether to enroll in college, but not his or her wages: the latter statement however is about *exogeneity*, and it is dubious at best.

The resulting set of moment conditions is

$$\mathbb{E}\left[\begin{pmatrix} 1 \\ X_i \\ X_i^2 \\ Z_i \\ G_i \\ F_i \end{pmatrix} \underbrace{\left(\log W_i - \beta_0 - \beta_1 X_i - \beta_2 X_i^2 - \beta_3 S_i\right)}_{=\alpha_i+\epsilon_i}\right] = 0 \tag{12.39}$$

six conditions for four parameters of the Mincer Equations. As we have discussed, this model can be estimated via 2SLS-GMM by setting $Y_i = \log W_i$, $\boldsymbol{x}_i^{\mathrm{T}} = (1, X_i, X_i^2, S_i)$ and $\boldsymbol{z}_i^{\mathrm{T}} = (1, X_i, X_i^2, Z_i, G_i, F_i)$. However, having three instruments for the education variable provides the interesting opportunity to *test* the exogeneity conditions implied, respectively, by the fourth, fifth

and sixth rows of (12.39). In this context, the Hansen's J-statistic reads as

$$\mathrm{J}\left(\widehat{\boldsymbol{\beta}}_{2SLS}\right) = N\left[\sum_{i=1}^{N}\mathbf{z}_i e_i\right]^{\mathrm{T}}\left[\sum_{i=1}^{N}e_i^2\mathbf{z}_i\mathbf{z}_i^{\mathrm{T}}\right]^{-1}\left[\sum_{i=1}^{N}\mathbf{z}_i e_i\right] \xrightarrow{d} \chi^2_2$$

where, as usual, $e_i = y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{2SLS}$. A rejection of the associated test would cast doubts on the validity of (12.39). Which are, however, the very "false instruments" responsible for the rejection of Hansen's J test? An answer to this question can be given by sequentially removing each of the instruments Z_i, G_i and F_i, estimating a smaller 2SLS model with five moments for four parameters, and computing the associated Hansen J-statistic, which would be asymptotically χ^2 distributed with only one degree of freedom. Such a battery of incremental Sargan tests may highlight the fact that, for example, the moment condition for the "friends" instrument F_i is patently violated, arguably because the quality of one individual's connections are correlated to determinants of his or her wage in the labor market. ■

12.4 Methods of Simulated Moments

The discussion in this lecture has thus far assumed that functions $\boldsymbol{g}(\boldsymbol{x}_i;\boldsymbol{\theta})$ – or more restrictively, $\boldsymbol{z}_i \otimes \boldsymbol{h}(\boldsymbol{y}_i,\boldsymbol{z}_i;\boldsymbol{\theta})$ – that are used to define and construct GMM estimators can always be computed given values for the parameters $\boldsymbol{\theta}$ and realizations of the random variables involved. Sometimes, however, this is not the case: for example, these functions may be expressed in terms of integrals lacking a closed form solution. This is analogous to the problem discussed in Lecture 11, identified in the case of MLE and more generally of all M-Estimators, a typical solution for which is the theoretical development and practical implementation of simulation-based estimators. Solutions of this sort also extend to estimators based on moment conditions.

To better introduce such approaches to methods of moments, it is useful to rephrase some concepts and notation previously introduced in the discussion of simulated M-Estimators from Lecture 11. Suppose that theoretical analysis postulates some moment conditions $\mathbb{E}\left[\boldsymbol{g}(\boldsymbol{x}_i;\boldsymbol{\theta}_0)\right] = \mathbf{0}$, where:

$$\boldsymbol{g}(\mathbf{x}_i;\boldsymbol{\theta}) = \int_{\mathbb{U}} \boldsymbol{g}_{\boldsymbol{u}}(\mathbf{x}_i,\mathbf{u}_i;\boldsymbol{\theta})\, dH_{\boldsymbol{u}}(\mathbf{u}_i) \tag{12.40}$$

where, like in (11.37), $\boldsymbol{u}_i$ is a random vector with cumulative distribution $dH_{\boldsymbol{u}}(\mathbf{u}_i)$ that is integrated out over its support $\mathbb{U}$. The integral in (12.40) could lack a closed form solution for any given values of the parameters $\boldsymbol{\theta}$ and realizations of the observable variables $\mathbf{x}_i$: naturally, this would prevent the straightforward implementation of (G)MM.

In such cases, Direct Monte Carlo Sampling based on a sample $\{\mathbf{u}_s\}_{s=1}^{S}$ of S random draws of $\boldsymbol{u}_i$ from $H_{\boldsymbol{u}}(\mathbf{u}_i)$ allows to construct a **simulator** that converges in probability to (12.40) for each observation $i = 1, \ldots, N$:

$$\widehat{\boldsymbol{g}}_S(\mathbf{x}_i; \boldsymbol{\theta}) = \frac{1}{S} \sum_{s=1}^{S} \widetilde{\boldsymbol{g}}_{\boldsymbol{u}}(\mathbf{x}_i, \mathbf{u}_s; \boldsymbol{\theta}) \tag{12.41}$$

where $\widetilde{\boldsymbol{g}}_{\boldsymbol{u}}(\mathbf{x}_i, \mathbf{u}_s; \boldsymbol{\theta})$ is a **subsimulator** that is ideally an unbiased estimator of (12.40) so as to guarantee $\widehat{\boldsymbol{g}}_S(\mathbf{x}_i; \boldsymbol{\theta}) \xrightarrow{p} \boldsymbol{g}(\mathbf{x}_i; \boldsymbol{\theta})$ by standard asymptotic arguments. Thus, the **Method of Simulated Moments** (MSM) estimator is defined as:

$$\widehat{\boldsymbol{\theta}}_{MSM} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\min} \left[\frac{1}{N} \sum_{i=1}^{N} \widehat{\boldsymbol{g}}_S(\mathbf{x}_i; \boldsymbol{\theta}) \right]^{\mathrm{T}} \mathbf{A}_N \left[\frac{1}{N} \sum_{i=1}^{N} \widehat{\boldsymbol{g}}_S(\mathbf{x}_i; \boldsymbol{\theta}) \right] \tag{12.42}$$

where $\mathbf{A}_N$ is a $J \times J$ weighting matrix whose probability limit is $\mathbf{A}_0$, as in standard GMM. The MSM estimator expressed in (12.42) naturally accommodates overidentification; notice however that this estimator is extensively used even in just-identified cases ($J = K$) where the expression of the moment conditions call for simulation and where $\mathbf{A}_N = \mathbf{A}_0 = \mathbf{I}$. In those cases where the moment conditions are conditional upon instrumental variables as in (12.27), the simulator takes the form

$$\widehat{\boldsymbol{g}}_S(\mathbf{x}_i; \boldsymbol{\theta}) = \frac{1}{S} \sum_{s=1}^{S} \mathbf{z}_i \widetilde{\boldsymbol{h}}_{\boldsymbol{u}}(\mathbf{y}_i, \mathbf{z}_i, \mathbf{u}_s; \boldsymbol{\theta}) \tag{12.43}$$

where $\widetilde{\boldsymbol{h}}_{\boldsymbol{u}}(\mathbf{y}_i, \mathbf{z}_i, \mathbf{u}_s; \boldsymbol{\theta})$ is again a suitable subsimulator, here for $\boldsymbol{h}(\mathbf{y}_i, \mathbf{z}_i; \boldsymbol{\theta})$.

The following example illustrates the necessity for MSM estimators using a particular case, that builds upon Example 11.13, of the LDV settings that originally motivated the seminal article on MSM by McFadden (1989).

Example 12.2. Random coefficients logit with instrumental variables. Recall the random coefficients logit model from Example 11.13. The assumptions of that model naturally lead to the moment condition

$$\mathbb{E}\left[Y_i - \Lambda(\beta_0 + \beta_{1i} X_i) \middle| X_i \right] = 0 \tag{12.44}$$

where $\Lambda(\cdot)$ is the cumulative logistic distribution of the binary outcome Y_i conditional on X_i. The interpretation of (12.44) is analogous to that of the mean independence condition of linear models: conditional on the regressor X_i, the "error" $Y_i - \Lambda(\beta_0 + \beta_{1i} X_i)$ must equal zero on average, that is, the

logistic model predicts the outcome Y_i without bias. Although (12.44) lends itself naturally to Method of Moments estimation, in practice function $\Lambda(\cdot)$ might be hard to calculate under probabilistic assumptions on the random coefficients β_{i1}. Let again $\beta_{1i} \sim \mathcal{N}(\beta_1, \sigma^2)$ and $u_i = (\beta_{1i} - \beta_1)/\sigma$, then:

$$\Lambda(\beta_0 + \beta_{1i} X_i) = \int_{\mathbb{R}} \Lambda(\beta_0 + (\beta_1 + \sigma u_i) X_i) \, \phi(u_i) \, du_i \tag{12.45}$$

which is an integral without closed form solution, similarly as in (11.40). It is easy to see how a simulation allows to construct a MSM estimator with the particular expression of (12.43) here being:

$$\widehat{g}_S(y_i, x_i; \theta) = \frac{1}{N} \sum_{i=1}^{N} x_i \left[y_i - \frac{1}{S} \sum_{s=1}^{S} \widetilde{\Lambda}(\beta_0 + (\beta_1 + \sigma u_s) x_i) \right]$$

given the subsimulator $\widetilde{\Lambda}(\beta_0 + (\beta_1 + \sigma u_s) x_i)$ and $\theta = (\beta_0, \beta_1, \sigma^2)$. Notice that this would be a case of **just-identified** MSM.

Now suppose that (12.44) does not hold with equality because of some instance of endogeneity: for whatever reason, the regressor X_i is correlated with the model error $Y_i - \Lambda(\beta_0 + \beta_{1i} X_i)$, thereby invalidating the moment condition! A straightforward solution to this problem in the GMM spirit is to use $J \geq K$ **instrumental variables** $\boldsymbol{z}_i$ that satisfy

$$\mathbb{E}[Y_i - \Lambda(\beta_0 + \beta_{1i} X_i) | \, \boldsymbol{z}_i] = 0 \tag{12.46}$$

and that correlate with the main regressor X_i. Thus, using the simulator

$$\widehat{g}_S(y_i, x_i, \mathbf{z}_i; \theta) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{z}_i \left[y_i - \frac{1}{S} \sum_{s=1}^{S} \widetilde{\Lambda}(\beta_0 + (\beta_1 + \sigma u_s) x_i) \right]$$

allows to construct an **overidentified** MSM estimator suited to this setting, so long as assumption (12.46) can be maintained. This example illustrates the type of flexibility that the MSM affords for LDV models that has helped make this estimation framework popular in relatively more structural fields of economics like Industrial Organization. ■

The discussion so far has borne many similarities with that about MSL from Lecture 11. There are less analogies, however, as far as the asymptotic properties of the two estimators are concerned. To appreciate this, define

$$\widetilde{\Omega}_0 \equiv \lim_{N \to \infty} \mathbb{V}\text{ar} \left[\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \widehat{\boldsymbol{g}}(\boldsymbol{x}_i; \theta_0) \right] \tag{12.47}$$

as the $J \times J$ limiting variance-covariance of the simulators, akin to (12.12). Note that this matrix also features noise due to the simulation; indeed one can show by the Law of Total Variance that if the simulator is unbiased:

$$\widetilde{\boldsymbol{\Omega}}_0 = \boldsymbol{\Omega}_0 + \lim_{N\to\infty} \mathbb{E}_{\boldsymbol{x}} \left[\mathbb{V}\mathrm{ar}_{\boldsymbol{u}} \left[\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \widehat{\boldsymbol{g}} \left(\boldsymbol{x}_i ; \boldsymbol{\theta}_0 \right) \right] \right] \tag{12.48}$$

where with regard to the second element on the right-hand side, the outer expectation is taken with respect to the observable variables, whereas the inner variance-covariance is taken with respect to the simulation draws. It must be remarked, however, that as $S \to \infty$ this second element vanishes! Intuitively, the larger the simulation the smaller the noise associated with it. Having these considerations in mind, it is easier to discuss the following result, which extends one originally given by McFadden (1989).

Theorem 12.3. Asymptotic Efficiency of the Method of Simulated Moments estimators. *If some MSM estimator is based upon an ubiased simulator $\widehat{\boldsymbol{g}}_S \left(\boldsymbol{x}_i ; \boldsymbol{\theta} \right)$, and all conditions implicit in the statement of Theorem 12.1 are met, even with fixed S the estimator is consistent and its limiting distribution is as follows.*

$$\sqrt{N} \left(\widehat{\boldsymbol{\theta}}_{MSM} - \boldsymbol{\theta}_0 \right) \xrightarrow{d} \mathcal{N} \left(\mathbf{0}, \left(\mathbf{G}_0^{\mathrm{T}} \mathbf{A}_0 \mathbf{G}_0 \right)^{-1} \mathbf{G}_0^{\mathrm{T}} \mathbf{A}_0 \widetilde{\boldsymbol{\Omega}}_0 \mathbf{A}_0 \mathbf{G}_0 \left(\mathbf{G}_0^{\mathrm{T}} \mathbf{A}_0 \mathbf{G}_0 \right)^{-1} \right) \tag{12.49}$$

Proof. (Outline.) The proof proceeds by decomposing and reworking the First Order Conditions of (12.42) as in the proof of Theorem 12.1, relying on the unbiasedness of the simulator for simplifying some key expressions. □

The key phrase of the Theorem's statement is "even with fixed S." The key result is that a large simulation size is not necessary in order to guarantee consistency, unlike the case of MSL! The reason is that the simulation is "washed out" in the First Order Conditions (has no effect on average if the simulator is unbasied) unlike in MSL where the need for taking logarithms complicates things. This is quite a relevant advantage of MSM, even though it has been documented that the solution might be numerically unstable for small values of S. Hence, there are still practical advantages in increasing the size of the simulation wherever possible; among the others – as already mentioned – as $S \to \infty$ it follows that $\widetilde{\boldsymbol{\Omega}}_0 \xrightarrow{p} \boldsymbol{\Omega}_0$; this implies that (12.49) and (12.11) coincide at the limit, making the estimation of the asymptotic variance-covariance easier. When comparing MSM against MSL (which can be relevant in contexts like that of Example 12.2) all other considerations already made about GMM against MLE still apply: in particular, MSM is more robust to assumption failures, otherwise MLS may be more efficient.

When conducting inference about the MSM estimator, it is necessary to consistently estimate the elements of the variance-covariance of (12.49). While $\mathbf{A}_0$ is naturally estimated by $\mathbf{A}_N$ whenever necessary, $\mathbf{G}_0$ and $\mathbf{\Omega}_0$ are estimated via versions of (12.17), (12.18) and (12.19) that substitute function $\boldsymbol{g}(\mathbf{x}_i;\boldsymbol{\theta})$ evaluated at the GMM estimator with the simulator $\widehat{\boldsymbol{g}}_S(\mathbf{x}_i;\boldsymbol{\theta})$ evaluated at the MSM estimator (and similarly when dealing with the HAC case). With small S it is also necessary to estimate the component of matrix $\widetilde{\mathbf{\Omega}}_0$ that depends on the simulation; this is done by taking the appropriate sample analogue of the second element on the right-hand side of (12.48).

12.5 Applications of GMM

Beyond making the foundation for estimators of linear models such as the 2SLS and the 3SLS, what are exactly the applications of the GMM framework in economics? Actually, one could make the case that *all* estimators in econometrics are particular cases of GMM: (as it was observed in passing, the Maximum Likelihood score at the true parameter value $\boldsymbol{\theta}_0$ can be seen as a set of moment conditions). In order to better illustrate both the usefulness and flexibility of GMM, this section discusses three *particular* settings that are especially well suited to the application of this framework: generic dynamic linear models for panel data; the estimation of production functions with its associated longstanding issues, and non-linear macroeconomic models based on the rational expectation hypothesis.

Dynamic Linear Models for Panel Data

A common hypothesis that is made when dealing with panel data is that the model is **dynamic**, that is the endogenous variables are dependent on their past realizations.

$$y_{it} = \alpha + \mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta} + \gamma y_{i(t-1)} + \varepsilon_{it} \tag{12.50}$$

In a model of this kind, parameter γ measures the dependence of the dependent variable from its immediate past: a recurring theme for macroeconomic variables such as say GDP, unemployment or inflation. A typical problem is that the error term ε_{it} is typically also autoregressive, that is it also depends on its past:

$$\varepsilon_{it} = \rho\varepsilon_{i(t-1)} + \xi_{it}$$

for some $\rho \in [-1, 1]$. This fact alone clearly invalidates the use of a a simple estimator like OLS: past outcomes depend on both current and past shocks, leading to endogeneity: $\mathbb{E}[Y_{i(t-1)}\varepsilon_{it}] \neq 0$.

Another problem that is specific of dynamic panels is that the standard solutions to the issue of **unobserved heterogeneity** (observation-specific omitted factors) are not available. Suppose that $\boldsymbol{\alpha} = 0$ and that instead of being autoregressive, the error term contains a *constant* observation-specific factor α_i: a so-called *fixed effect*.

$$\varepsilon_{it} = \alpha_i + \epsilon_{it}$$

The fixed effect α_i is by construction correlated to the the lagged outcome: $\mathbb{E}[Y_{i(t-1)}\alpha_{it}] \neq 0$. In panel data models that are not "dynamic," the typical solution to "endogenous" fixed effects is a transformation like:

$$y_{it} - \bar{y}_i = \left(\mathbf{x}_{it} - \bar{\mathbf{x}}_i\right)^{\mathrm{T}} \boldsymbol{\beta} + \gamma \left(y_{i(t-1)} - \bar{y}_i\right) + \varepsilon_{it} - \bar{\varepsilon}_i$$

where the upper bar applied to some variable x_{ki} denote observation-specific averages over time t, like $\bar{x}_{ki} = T^{-1} \sum_{t=1}^{\mathrm{T}} x_{kit}$ (this particular approach has been called "within transformation" in Lecture 10). Not even this method can, unfortunately, work for dynamic linear models, because the demeaned lagged outcome $Y_{i(t-1)} - \bar{Y}_i$ is **mechanically** correlated to the demeaned shock $\varepsilon_{it} - \bar{\varepsilon}_i$. In fact, past values of the outcomes depend on **all** the past shocks, which in turn are **all** included in the average shock $\bar{\varepsilon}_i$!

$$\mathbb{E}\left[\left(Y_{i(t-1)} - \bar{Y}_i\right)\left(\varepsilon_{it} - \bar{\varepsilon}_i\right)\right] \neq 0$$

The typical solution to this problem is to estimate dynamic linear panel data models via GMM, where the moment conditions include either:

- **moment in levels**, featuring a product between the first difference of the error term and the higher lags (one period and beyond) of the dependent variable (see Arellano and Bond, 1991; Arellano and Bover, 1995), these can be expressed as:

$$\mathbb{E}\left[Y_{i(t-s)}\Delta\varepsilon_{it}\right] = 0 \text{ for } s \geq 2;$$

- **moment in differences**, featuring a product between the error term and the higher lags (one period and beyond) of the first difference of the dependent variable (Blundell and Bond, 1998), these write as:

$$\mathbb{E}\left[\Delta Y_{i(t-s)}\varepsilon_{it}\right] = 0 \text{ for } s \geq 2;$$

or both. Notice that both approaches generally result in overidentified models. A major problem with this type of moment conditions, however, is that they work in theory, while in practice they present endemic problems of the

weak instruments type (that is, weak statistical correlation between the instruments and the structural variables). The GMM framework, thanks to its power to combine many instruments, optimally weigh them by their statistical relevance, and test their validity, is therefore well suited to address these issues. Unsurprisingly, the GMM framework has become dominant in the estimation of dynamic macroeconomic models or other dynamic models for panel data. As the following discussion shows, however, approaches similar to that outlined above are also adopted to address other more "classical" kinds of endogeneity problem.

Estimation of Production Functions

Recall the *log-log* production function model (7.43) described in Lecture 7, based on the **Cobb-Douglas** functional form. That model can be adapted to panel data, writing for convenience $y_{it} \equiv \log Y_{it}$, $k_{it} \equiv \log K_{it}$, $\ell_{it} \equiv \log L_{it}$ and:

$$y_{it} = \alpha_i + \beta_K k_{it} + \beta_L \ell_{it} + \omega_{it} + \varepsilon_{it} \tag{12.51}$$

where:

$$\log A_{it} = \alpha_i + \omega_{it} + \varepsilon_{it}$$

that is, the logarithm of the productivity measure A_{it} has been separated between three components: a unit-constant factor α_i and two time-varying factors ω_{it} and ε_{it}. Why make this distinction? It is usually impossible to observe all the specific factors affecting one firm's productivity, and thus they must be treated as random shocks. The problem here is that that while some of these factors can be thought independent of firm's idiosyncratic decisions (exogenous), some others cannot. In fact, standard microeconomic analysis suggests that if the management of a firm experiences higher productivity A_{it}, at the same costs it will find more convenient to hire more workers L_{it} and invest in more capital K_{it}.

A researcher interested in recovering empirical values for the parameters β_K and β_L should think twice about estimating a standard regression model for (12.51). In fact, while the exogeneity assumption may sound plausible for some component of the "random" shocks to productivity, like ε_{it}:

$$\mathbb{E}\left[\varepsilon_{it} \middle| k_{it}, \ell_{it}\right] = 0$$

(think about lucky events), it is not so for the rest:

$$\mathbb{E}\left[\alpha_i, \omega_{it} \middle| k_{it}, \ell_{it}\right] \neq \mathbf{0}$$

thereby making OLS estimates inconsistent. While α_i might be removed by, say, applying the within transformation, one is still left with the so-called

endogenous unobserved **productivity shock** ω_{it}. To complicate things, this shock is typically assumed (on the basis of the empirical evidence) to present positive autocorrelation: $\omega_{it} = \rho\omega_{i(t-1)} + \xi_{it}$, where $\rho \in [0, 1]$.

There are several proposed solutions to the "unobserved productivity shock" problem in the estimation of production functions, and none of them is perfect. The method by Blundell and Bond (1998) devised for dynamic panel data models is one of these potential solutions, despite the fact that a production function is not strictly speaking a dynamic model. The intuition goes as follows: suppose to subtract $\rho y_{i(t-1)}$ from both sides of (12.51); the result is the transformed model:

$$y_{it} - \rho y_{i(t-1)} = \alpha_i (1 - \rho) + \beta_K \left(k_{it} - \rho k_{i(t-1)}\right) + \beta_L \left(\ell_{it} - \rho \ell_{i(t-1)}\right) + \upsilon_{it}$$

where $\upsilon_{it} \equiv \xi_{it} + \varepsilon_{it} - \rho\varepsilon_{i(t-1)}$: the "backward looking" autoregressive endogenous shock is removed, and all that is left are components of the random shocks that are arguably exogenous to appropriate lags of the first differences of the capital and labor inputs. This shows how moment conditions of the kind,

$$\mathbb{E}\left[\begin{pmatrix}\Delta k_{i(t-s)} \\ \Delta \ell_{i(t-s)}\end{pmatrix} \left(\alpha_i (1 - \rho) + \xi_{it} + \varepsilon_{it} - \rho\varepsilon_{i(t-1)}\right)\right] = 0$$

can be used to form GMM estimators for $s \geq 2$.

In practice, however, the Blundell-Bond approach applied to production functions is known to work poorly in small samples and it is generally less precise than other competing estimators, similarly to the weak instruments problem associated to GMM estimators for dynamic panel data models. For this reason, the frontier method for the estimation of production functions nowadays is based on a combination of moment conditions such as:

$$\mathbb{E}\left[\begin{pmatrix}k_{i(t-s)} \\ \ell_{i(t-s)}\end{pmatrix} \left(y_{it} - \beta_K k_{it} - \beta_L \ell_{it} - g\left(\widehat{\varphi}_{it} - \beta_K k_{i(t-1)} - \beta_L \ell_{i(t-1)}\right)\right)\right] = 0$$

for $s = 2, \ldots, t$; where $\widehat{\varphi}_{it} = \widehat{\varphi}\left(k_{i(t-s)}, k_{i(t-s)}, m_{i(t-s)}\right)$ is a non-parametric prediction function – e.g. a polynomial approximation – of $\alpha_i + \omega_{i(t-1)}$, that is obtained via suitable lags of k_{it}, ℓ_{it} and of another instrument or "shifter" variable m_{it} (for example, variable input materials); and $g(\cdot)$ is yet another non-parametric function. In this **control function** approach to production function estimation, a "first step" aimed at estimating $\widehat{\varphi}_{it}$ is necessary before proceeding to GMM estimation based on the above conditions. The implied exclusion restrictions are motivated on careful assumptions about the timing of firms' decisions; for example, if capital investment reflects changes in the

firm's economic conditions with a lag (it takes time to invest in new capital and equip it) it makes sense to motivate a moment condition akin to:

$$\mathbb{E}\left[\varepsilon_{it}, \xi_{it} \middle| k_{it}\right] = \mathbf{0}$$

as firms cannot observe ξ_{it} timely enough so as to affect their choice of k_{it}. For additional discussion about the more modern practices in the estimation of production functions, see Wooldridge (2009) and Ackerberg et al. (2015).

Estimation of Rational Expectations Models

While GMM is currently seen as a wide-encompassing framework which includes many econometric estimators, its initial popularity was largely due to its flexibility in estimating – without resorting to fully parametric assumptions – the possibly non-linear moment conditions *derived from economic theory*. While nowadays GMM is employed for empirical research in many fields of economics, it is traditionally especially relevant in macroeconomics, and in particular for modeling theories based on the hypothesis of **rational expectations**. An example of such an application of GMM, based on the model of **permanent income** by Hall (1978) is discussed next.[5]

Suppose that a consumer aims at maximizing his *lifetime utility* from the consumption of various goods and services as per the following *intertemporal utility function*:

$$\mathcal{U}_t\left(C_t, C_{t+1}, \ldots, C_T\right) = \mathbb{E}_t\left[\sum_{\tau=0}^{T-t}\left(\frac{1}{1+\delta}\right)^{\tau} U\left(C_{t+\tau}\right)\middle| \mathbb{I}_t\right] \tag{12.52}$$

subject to *intertemporal budget constraint*

$$\sum_{\tau=0}^{T-t}\left(\frac{1}{1+\delta}\right)^{\tau}\left(C_{t+\tau} - W_{t+\tau}\right) = A_t \tag{12.53}$$

where: C_s are aggregate consumption expenditures at time s; $U\left(C_s\right)$ is the associated per-period utility; W_s are the individual earnings at time s; A_s is the amount of individual asset owned by the individual at time s; r is the interest rate (assumed constant for simplicity); δ is the *discount factor*, a parameter that denotes an individual's "impatience" towards the idea of postponing consumption. Finally, $\mathbb{I}_s$ represents quite an abstract economic concept, the *information set*: a set of variables (possibly written also as $\boldsymbol{z}_s$)

[5]The treatment here adapts the one provided by William H. Greene in his leading econometric textbook.

whose value at time s affects individual expectations about future economic outcomes. The substance of the problem is that individuals cannot consume over their lifetime, in excess of what they actually earn, a sum that in net present value terms exceeds their current assets. However, they do not even know for certain the value of their future earnings: consequently also their future utility is subject to stochastic uncertainty. However, individuals can form expectations of the form $\mathbb{E}_t\left[W_{t+\tau}\middle|\,\mathbb{I}_t\right]$ about their future earnings, that are conditional of their information set.

Intertemporal utility is clearly separable as the sum of distinct period-specific utility functions: a crucial assumption. Hall's main result is having shown that in such a case, the problem's solution is conveniently expressed as the **Euler equation** of two consecutive marginal utilities:

$$\mathbb{E}_t\left[U'\left(C_{t+1}\right)\middle|\,\mathbb{I}_t\right] = \frac{1+\delta}{1+r}\,U'\left(C_t\right) \tag{12.54}$$

which can be operationalized if one is willing to make assumptions about the functional form of $U\left(\cdot\right)$. The "Constant Relative Risk Aversion" (CRRA) utility function is a popular choice; it reads as follows.

$$U\left(C_t\right) = \frac{1}{1-\alpha}\,C_t^{1-\alpha}$$

With this assumption, (12.54) becomes:

$$\mathbb{E}_t\left[\,\beta\left(1+r\right)R_{t+1}^{\lambda} - 1\middle|\,\mathbb{I}_t\right] = 0 \tag{12.55}$$

where $\beta \equiv \left(1+\delta\right)^{-1}$, $\lambda \equiv -\alpha$ and $R_{t+1} = C_{t+1}/C_t$.

A researcher might be interested about the GMM estimation of the two parameters (β, λ). A natural set of moment conditions that make the model just identified is given by the following expression.

$$\mathbb{E}_t\left[\begin{pmatrix}1\\R_t\end{pmatrix}\left(\beta\left(1+r\right)R_{t+1}^{\lambda} - 1\right)\right] = 0 \tag{12.56}$$

However, if the researcher knows about some other variables $\boldsymbol{z}_t$ that work as good predictors of future earnings, extra moment conditions in the form:

$$\mathbb{E}_t\left[\boldsymbol{z}_t\left(\beta\left(1+r\right)R_{t+1}^{\lambda} - 1\right)\right] = 0 \tag{12.57}$$

result in overidentification, and thus the GMM estimator of (12.57) follows from the previous discussion of instrumental variables in non-linear models. It must be appreciated that this GMM estimator has been constructed using just: *i.* some predictions of economic theory, *ii.* a specific mean assumption about the stream of future earnings, conditional on current information $\mathbb{I}_t$. In fact, no more detailed parametric assumption about the distribution of future income have resulted necessary for deriving the moment conditions.

Bibliography

Ackerberg, Daniel A., Kevin Caves, and Garth Frazer, "Identification Properties of Recent Production Function Estimators," *Econometrica*, 2015, *83* (6), 2411–2451.

Andrews, Donald W. K. and J. Christopher Monahan, "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator," *Econometrica*, 1991, *60* (4), 953–966.

Angrist, Joshua D. and Alan B. Krueger, "Empirical Strategies in Labor Economics," in O. C. Ashenfelter and D. Card, eds., *Handbook of Labor Economics*, Vol. 3, Elsevier, 1999, pp. 1277–1366.

Arellano, Manuel and Olympia Bover, "Another look at the instrumental variable estimation of error-components models," *Journal of Econometrics*, 1995, *68* (1), 29–51.

_ **and Stephen Bond**, "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations," *The Review of Economic Studies*, 1991, *58* (2), 277–297.

Bartlett, Maurice S., "Periodogram Analysis and Continuous Spectra," *Biometrika*, 1950, *37* (1/2), 1–16.

Becker, Gary S., "Investment in Human Capital: A Theoretical Analysis," *Journal of Political Economy*, 1962, *70* (5), 9–49.

Berry, Steven, "Estimation of a Model of Entry in the Airline Industry," *Econometrica*, 1992, *60* (4), 889–917.

_ **and Peter Reiss**, "Empirical Models of Entry and Market Structure," in Mark Armstrong and Robert Porter, eds., *Handbook of Industrial Organization*, Vol. 3, North Holland, 2007, pp. 1845–1886.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan, "How Much Should We Trust Differences-In-Differences Estimates?," *The Quarterly Journal of Economics*, 2004, *119* (1), 249–275.

Bester, C. Alan, Timothy G. Conley, and Christian B. Hansen, "Inference with dependent data using cluster covariance estimators," *Journal of Econometrics*, 2011, *165* (2), 137–151.

Blundell, Richard and Stephen Bond, "Initial conditions and moment restrictions in dynamic panel data models," *Journal of Econometrics*, 1998, *87* (1), 115–143.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin, "Identification of peer effects through social networks," *Journal of Econometrics*, 2009, *150* (1), 41–55.

Bresnahan, Timothy F., "Competition and collusion in the American automobile industry: The 1955 price war," *The Journal of Industrial Economics*, 1987, *35* (4), 457–482.

Cameron, Colin A., Jonah G. Douglas, and Douglas L. Miller, "Robust Inference with Multiway Clustering.," *Journal of Business & Economic Statistics*, 2011, *29* (2), 238–249.

Card, David, "Using geographic variation in cullege proximity to estimate the return to schooling," in L. N. Christofides, E. K. Grant, and R. Swidinsky, eds., *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*, University of Toronto Press, 1995.

Chamberlain, Gary, "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 1987, *34* (3), 305–334.

Conley, Timothy J., "GMM estimation with cross sectional dependence," *Journal of Econometrics*, 1999, *92* (1), 1–45.

Durbin, James, "Errors in Variables," *Review of the International Statistical Institute*, 1954, *22* (1/3), 23–32.

Eicker, Friedhelm, "Limit Theorems for Regressions with Unequal and Dependent Errors," in L. LeCam and J. Nexman, eds., *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, 1967, pp. 59–82.

Gouriéroux, Christian and Alan Monfort, "Simulation Based Inference in Models with Heterogeneity," *Annales d'Économie et de Statistique*, 1991, *20/21*, 69–107.

Hall, Robert E., "Stochastic implications of the life cycle-permanent income hypothesis: theory and evidence," *Journal of Political Economy*, 1978, *86* (6), 971–987.

Hansen, Lars P., "Large sample properties of Generalized Method of Moments estimators," *Econometrica*, 1982, *50* (4), 1029–1054.

_ , John Heaton, and Amir Yaron, "Finite-sample properties of some alternative GMM estimators," *Journal of Business & Economic Statistics*, 1996, *14* (3), 262–280.

Hausman, Jerry A., "Specification Tests in Econometrics," *Econometrica*, 1978, *46* (6), 1251–1271.

Heckman, James J., "Sample selection bias as a specification error (with an application to the estimation of labor supply functions)," *Econometrica*, 1977, *47* (1), 153–161.

Kelejian, Harry H. and Ingmar R. Prucha, "A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances," *The Journal of Real Estate Finance and Economics*, 1998, *17* (1), 99–121.

_ and _ , "HAC estimation in a spatial framework," *Journal of Econometrics*, 2007, *140* (1), 131–154.

Klein, Lawrence Robert, *Economic Fluctuations in the United States, 1921-1941*, John Wiley & Sons, 1950.

Kmenta, Jan, "On Estimation of the CES Production Function," *International Economic Review*, 1967, *8* (2), 180–189.

McFadden, Daniel, "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," *Econometrica*, 1989, *57* (5), 995–1026.

Mincer, Jacob A., "Investment in Human Capital and Personal Income Distribution," *Journal of Political Economy*, 1958, *66* (4), 281–302.

Moulton, Brent R., "Random Group Effects and the Precision of Regression Estimates," *Journal of Econometrics*, 1986, *32* (3), 385–397.

Newey, Whitney K. and Daniel McFadden, "Large Sample Estimation and Hypothesis Testing," in Robert Engle and Daniel McFadden, eds., *Handbook of Econometrics*, Vol. 4, North Holland, 1994, pp. 2111–2245.

Newey, Whitney K. and Kenneth D. West, "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 1987, *55* (3), 703–708.

Rothenberg, Thomas J., "Identification in parametric models," *Econometrica*, 1971, *39* (3), 577–591.

Rubin, Donald, "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 1974, *66* (5), 688–701.

Stock, James H., Jonathan H. Wright, and Motohiro Yogo, "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business & Economic Statistics*, 2002, *20* (4), 518–529.

Theil, Henry, "Repeated least squares applied to complete equation systems," Technical Report, Central Planning Bureau of The Hague, 1953.

Walsh, J. R., "Capital Concept Applied to Man," *Quarterly Journal of Economics*, 1935, *49* (2), 255–285.

White, Halbert, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 1980, *48* (4), 817–838.

Wooldridge, Jeffrey M., "Score diagnostics for linear models estimated by two stage least squares," in G. S. Maddala, P. C. B. Phillips, and T. N. Srinivasan, eds., *Advances in Econometrics and Quantitative Economics: Essays in Honor of Professor C. R. Rao*, Oxford: Blackwell, 1995.

_ , "On estimating firm-level production functions using proxy variables to control for unobservables," *Economic Letters*, 2009, *104* (3), 112–114.

Wu, De-Min, "Alternative Tests of Independence between Stochastic Regressors and Disturbances," *Econometrica*, 1973, *41* (4), 733–750.

Yitzhaki, Shlomo, "On Using Linear Regressions in Welfare Economics," *Journal of Business Economics and Statistics*, 1996, *14* (4), 478–486.