Statistics for Econometrics

Class Notes

Leonard Goff

This version: September 1, 2022

Guide to using these notes

This set of notes arose from the course ECON8070: Statistics for Econometrics at the University of Georgia in Fall 2021. I'm fixing typos as I find them and keeping this document updated on my website. Check www.leonardgoff.com for the most recent version.

These notes feature two kinds of box, to help organize the material:

Gray boxes offer section summaries.

White boxes indicate material that is optional, and understanding this material is not required for the course or exam.

Sections that have an asterisk at the end of their title can be skipped in their entirety: understanding this material is not required for the course or exam. These sections are mostly there for your interest and reference.

Chapter 1

Probability

1.1 Probability spaces

Main idea: A *probability function* ascribes a number to each of a collection of *events*, where each event is a set of *outcomes.*

This section develops the mathematical notion of probability. Probability is a function that associates a number between zero and one to *events.* Events, in turn, are sets of *outcomes.* It's easiest to think of outcomes in the context of a process that could have multiple distinct results, like flipping a coin or randomly choosing a number from a phone book.

1.1.1 Outcomes and events

We begin with a set Ω of conceivable outcomes, which is referred to as the *sample space* or *outcome space.*

Examples: When flipping a coin, the sample space is $\Omega = \{H, T\}$, corresponding to "heads" or "tails", respectively. When rolling a six-sided die, $\Omega = \{1, 2, 3, 4, 5, 6\}$. When drawing a card from a 52-card deck, the sample space can de denoted as a combination of a card-value and a suit, or $\{(n, s) : n \in \{A, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K\}, s \in \{hearts, spades, diamonds, clubs\}\}$. When using a random number generator to draw any number between 0 and 1, the sample space is $\Omega = [0, 1]$.

We denote a generic element of the sample space as as $\omega \in \Omega$. What we call *events* are simply sets of such ω, i.e. subsets of Ω. But in general, not all subsets of Ω necessarily need to be events. Rather, we consider a collection of sets F, referred to as an *event space.*

Definition 1.1. *An event space F is a collection of subsets $A \subseteq \Omega$.*

In all of the examples given above, the outcome space Ω has a finite number of elements. In such cases, it is typical to choose F to be the collection of *all* subsets of Ω. This collection is referred to as the *powerset* of Ω and is often denoted as 2^Ω. As an example, the powerset of the set $\{1, 2\}$ is $2^{\{1,2\}} = \pmb{\{}\emptyset, \{1\}, \{2\}, \{1, 2\}\pmb{\}}$. When we consider Ω that are uncountable sets (for example when Ω is a continuum), we'll need to restrict the event-space, as discussed below.

1.1.2 The probability of an event

A *probability function* P associates a positive real number to each event $A \in F$.

Definition 1.2. *A probability function $P(\cdot)$ is a function from F to $\mathbb{R}$, satisfying the following properties:*

1. *$P(A) \geq 0$ for each $A \in F$*
2. *$P(\Omega) = 1$*
3. *If $A_1, A_2 \ldots$ is a countable collection of disjoint sets (i.e. $A_j \cap A_k = \emptyset$ for any $j \neq k$), then*

$$P\left(\bigcup_j A_j\right) = \sum_j P(A_j)$$

This formulation of probability is sometimes referred to as the *Kolmogorov axioms* of probability.

These axioms imply several intuitive properties of probability. For example, if A has a countable number of elements, then the third property in Definition 1.2 implies that:

$$P(A) = \sum_{\omega \in A} P(\{\omega\})$$

provided that $\{\omega\} \in F$ for each $\omega \in A$. In particular, this result implies that for a finite set A we can simply sum up the probability of each of the outcomes in A. For example, for a six-sided die $P(even) = P(\{2\}) + P(\{4\}) + P(\{6\})$.

A few other properties of probability functions are left as exercises. As practice, I'll include a proof of the familiar property that $P(A^c) = 1 - P(A)$. To see this, note that A and A^c are disjoint sets, and that $A \cup A^c = \Omega$. Thus, by the third property of Definition 1.2 $P(\Omega) = P(A) + P(A^c)$. Then use the second property to obtain the result.

Exercise: Show that if $A \subseteq B$: $P(A) \leq P(B)$.

Exercise: Use the above to show that $P(A \cap B) \leq \min\{P(A), P(B)\}$.

Exercise: Derive the expression: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Exercise: Derive the expression: $P(A \cap B) = P(A) + P(B) - P(A \cup B)$. *Hint:* use $(A \cap B)^c = A^c \cup B^c$.

1.1.3 Which sets of outcomes get a probability?

In addition to the Kolmogorov axioms for the function P, we also place some requirements on the event space F. In particular, we require it to be a σ-algebra:

Definition 1.3. *A σ-algebra on Ω is a collection F of subsets of Ω with the following properties:*

1. $\Omega \in F$
2. *If any $A \in F$, then $A^c \in F$, where A^c is the complement of A in Ω*
3. *If $A_1, A_2 \dots$ are each in F, then $\bigcup_j A_j$ is also in F*

Recall that events $A \in F$ are those subsets of Ω that the function P must ascribe a probability (these sets A are called *measurable sets*). The first item above, that $\Omega \in F$, was already assumed by item 2. of Definition 1.2: we can always associate a probability with the whole outcome space, and that probability is one. Item 2. of Definition 1.3 says that if we are willing to give a probability to event A, then we should also be willing to give a probability to the event that A does not happen, i.e. A^c. The third property assures that given events A and B, we can always talk about the probability of A or B, which is $P(A \cup B)$.

Note that all of the properties of a σ-algebra tell us about things that must be in F, they guarantee that F is not to "small". The biggest collection of subsets of Ω is the set of all of its subsets: the powerset 2^Ω. The powerset of Ω is always a $\sigma-$algebra (exercise: check that it satisfies all three properties). However, using 2^Ω as the event space F can also be too *big* for certain applications. This is why it is necessary to introduce the idea of a σ-algebra.

Example: Consider as an outcome space the entire unit interval: $\Omega = [0, 1]$. It turns out that it is impossible to define a "uniform" probability function on this Ω, if we insist on using the whole powerset of $[0, 1]$ as our event space F. That is, there is no function $P(\cdot)$ satisfying Kolmogorov's axioms, and defined over all $A \in 2^{[0,1]}$, that satisfies our intuitive notion that moving a set around in the unit interval does not change its probability. See Proposition 1.2.6 of Rosenthal (2006) for details.

This example demonstrates that in some cases we may need to work with something smaller than 2^Ω. In particular, issues like the above arise when Ω is uncountably infinite, e.g. corresponding to a continuum of numbers. When Ω is finite or countable, it usually makes sense to consider the full powerset of Ω as our event space. When we are in the uncountable case (e.g. when Ω is a convex subset of the real line $[a, b]$), we typically appeal to the Borel σ-algebra:

Definition 1.4. *The Borel σ-algebra $\mathcal{B}$ is the collection that consists of all intervals of the forms $[a,b]$, $(a,b]$, $[a,b)$, (a,b), and all other sets in $\mathbb{R}$ that are then implied by the definition of a σ-algebra.*

Exercise: Show that for any Ω, the collection $\{\emptyset, \Omega\}$ is a $\sigma-$algebra.

Exercise: Show that $\emptyset \in F$ if F is a $\sigma-$algebra.

Exercise: Show that $\sigma-$algebras are closed under countable intersections, that is $\bigcap_j A_j$ is in F if $A_1, A_2, \dots$ are each in F.

1.1.4 Bringing it all together: a probability space

Once we have a sample space, event space, and probability function, we refer to them altogether as a probability space (sometimes called a *probability triple*).

Definition 1.5. *An probability space is a triple (Ω, F, P) in which F is a σ-algebra defined on Ω, and P is a probability function defined on F.*

1.2 Random variables

Main idea: If we associate a number to each outcome in a probability space, we have what is called a *random variable.*

1.2.1 Definition

Most data we use in econometrics is quantitative in nature, so its natural to think of probability spaces in which the outcome space is composed of numbers. Many of the examples have this feature already, for example $\Omega = \{1,2,3,4,5,6\}$ for a six-sided die. But even when the ω do not have an immediate numeric interpretation, we can define a random variable by associating a number to each outcome ω:

Definition 1.6. *Given a probability space (Ω, F, P), a random variable X is a function $X : \Omega \to \mathbb{R}$.*

Example: Suppose I randomly select a student in this class, which I represent by a probability space with $\Omega = \{\text{all students in this class}\}$, $F = 2^\Omega$, and $P(\{\omega\}) = 1/|\Omega|$ for each $\omega \in \Omega$. If we let $X(\omega)$ denote the height in inches of student ω.

A random variable X defined from a primitive probability space (Ω, F, P) allows us to define a new probability space in which the outcomes are real numbers. We can now define a new probability function P_X on sets of real numbers, using the original probability function P on Ω:

$$P_X(A) := P(\{\omega \in \Omega : X(\omega) \in A\}) \tag{1.1}$$

Technical note: observe that the above definition gives a way to associate a probability P_X with any set A of real numbers, provided that $\{\omega \in \Omega : X(\omega) \in A\} \in F$. To ensure this condition holds it is typical to restrict to sets A that belong to the Borel algebra $\mathcal{B}$ defined in Section 1.1.3, and further insist that the function X is *measurable.* X being measurable is a technical condition that just means that for any $x \in \mathbb{R}$, the set $\{\omega \in \Omega : X(\omega) \le x\} \in F$. Our new probability space can now be denoted as $(\mathbb{R}, \mathcal{B}, P_X)$.

A *realization* of random variable X is the specific value $X(\omega)$ that it ends up taking, given ω. While X is a *function*, $X(\omega)$ is a *number.* Lowercase letters x are often used to denote numbers that are possible realizations: e.g. $x = X(\omega)$ for some $\omega \in \Omega$.

1.2.2 Notation

The notation of Equation (1.1) is pretty cumbersome to work with, so the convention is to simplify it in a few ways.

Let's start with an example. If we're interested in the probability that $X(\omega)$ is less than or equal to 5, we'll typically write this as: $P(X \le 5)$, which can be interpreted as $P_X(A)$ where $A = (-\infty, 5]$, or equivalently: $P(\{\omega \in \Omega : X(\omega) \le 5\})$. What's changed in this notation? Let's go through step-by-step:

- First, we haven't bothered with the subscript X on P_X like in Equation (1.1) because it's clear from what's inside the parentheses that we're talking about random variable X.
- Second, inside the function P we're using the language of *conditions* rather than sets. That is, rather than writing out the set $A = (-\infty, 5]$ of values we're interested in, we just write this as a condition: "≤ 5".
- Third, we've made ω implicit and written X rather than $X(\omega)$. However, you often see ω left in. For example, we might write $P(X(\omega) = x)$ for the probability that X takes a value of x. In the context of Equation (1.1), this maps onto $P_X(\{x\})$, or equivalently $P(\{\omega \in \Omega : X(\omega) = x)\})$.

Given that we're using the language of conditions, we often write "and" inside probabilities, for example: $P(X \leq 5 \text{ and } X \geq 2)$. The "and" operation translates into intersection in the language of sets: $P(\{\omega \in \Omega : X(\omega) \leq 5\} \cap \{\omega \in \Omega : X(\omega) \geq 2\})$. Similarly, "or" translates into the union of sets: $P(X \leq 5 \text{ or } X \geq 2) = P(\{\omega \in \Omega : X(\omega) \leq 5\} \cup \{\omega \in \Omega : X(\omega) \geq 2\})$.

Note: We may have multiple random variables, e.g. X could be a randomly chosen state's minimum wage, while Y their unemployment rate. Mathematically, these two random variables correspond to functions $X(\cdot)$ and $Y(\cdot)$ applied to a common underlying outcome space Ω, which in this case corresponds to the set of US states. Probabilities like $P(X \leq \$10 \text{ and } Y \leq 5\%)$ are interpreted as $P(\{\omega \in \Omega : X(\omega) \leq \$10 \text{ and } Y(\omega) \leq 5\%\})$. If $P(\{\omega\}) = 1/50$ for all ω, then this probability is in turn equal to the number of states that have a minimum wage less than or equal $10 and an unemployment rate less than or equal 5%, divided by 50.

1.3 The distribution of a random variable

Main idea: The *cumulative distribution function* (CDF) provides a concise and convenient way to represent the probability function of a random variable or of multiple random variables. From the CDF we can define everything else we use to work with specific types of random variables, for example *probability density functions* and *probability mass functions.*

1.3.1 Central concept: the cumulative distribution function

We can summarize the probability function over values of a random variable X through the so-called *cumulative distribution function* or CDF of X.

Definition 1.7. *The cumulative distribution function of X is the function $F_X(x) := P(X \leq x)$.*

Note that $F_X(x)$ is a function from $\mathbb{R}$ to the unit interval $[0, 1]$, that is $F_X(x)$ is defined for all $x \in \mathbb{R}$ and $F_X(x)$ is always between zero and one. The following properties can be proven to hold for any random variable X:

- $F_X(x)$ is a weakly increasing function, that is $F_X(x') \geq F_X(x)$ if $x' > x$
- $\lim_{x \downarrow -\infty} F_X(x) = 0$
- $\lim_{x \uparrow \infty} F_X(x) = 1$
- $F_X(x)$ is right-continuous, i.e. $F_X(x) = \lim_{\epsilon \downarrow 0} F_X(x + \epsilon)$

Note on notation: when the context is clear, we often denote a CDF as $F(x)$ rather than $F_X(x)$. However, when we have multiple random variables like X and Y, we may need the notation $F_X(x)$ and $F_Y(y)$ to be clear about which variable we are referring to. When using the notation $F(x)$ for a CDF, keep in mind that this is not the same "F" as we used to denote the event space of a generic probability triple (Ω, F, P).

From the CDF, we can derive anything we'll need to know about a single random variable. When we have multiple random variables, the *joint-CDF* tells us everything we need to know about them.

Definition 1.8. *The joint-CDF of two random variables X and Y is the function*

$$F_{XY}(x, y) := P(X \leq x \text{ and } Y \leq y)$$

We'll come back to the joint-CDF of two (or more) random variables in Section 1.5.

Although the CDF $F(x)$ of a random variable is a function of a single variable x, we can use it to recover the probability that X lies in a *set*. For example, consider the set $(a, b]$, that is all numbers between a and b, including b itself.

Proposition 1.1. *For any numbers a and b such that $b \geq a$, $P(X \in (a, b]) = F(b) - F(a)$*

Proof. Given that $P(A) = 1 - P(A^c)$, (see Section 1.1.2), we have that:

$$P(X \in (a, b]) = P(a < X \leq b) = 1 - P(X \leq a \text{ or } X > b)$$

Using the third property of a probability function, we have that $P(X \leq a \text{ or } X > b) = P(X \leq a) + P(X > b)$, since the sets $\{x \in \mathbb{R} : x \leq a\}$ and $\{x \in \mathbb{R} : x > b\}$ are disjoint. Thus:

$$P(X \in (a, b]) = 1 - \{P(X \leq a) + P(X > b)\} = P(X \leq b) - P(X \leq a) = F(b) - F(a)$$

where I've used that $P(X \leq b) = 1 - P(X > b)$. □

More generally, we can from the function $F(x)$ compute the probability that $X \in A$ for any *Borel-measurable* set, that is a set A that belongs to the Borel $\sigma-$algebra. Sets that are simple intervals on the real line like $(a, b]$ are the leading example of such sets. Computing the probability associated with more complicated sets that aren't intervals is also possible using the CDF. The next section develops two functions that can be derived from the CDF, and are sometimes easier to work with for such computations.

1.3.2 Probability mass and density functions

Let X be a random variable with CDF $F(x)$. We often refer to the whole function F as the *distribution* of X. It always tells us everything we need to know about X. But there are two important special cases in which we can represent the distribution of X in an alternative way that is often more convenient.

1.3.2.1 Case 1: Discrete random variables and the probability mass function

Call $\mathcal{X}$ a *discrete set* if $\mathcal{X}$ contains a finite number of elements, or a countably infinte number of elements (e.g. $\mathcal{X} = \mathbb{N}$, the set of all integers).

Definition 1.9. *A discrete random variable X is a random variable such that $P(X \in \mathcal{X}) = 1$ for some discrete set $\mathcal{X}$.*

Example: If X is the number returned by rolling a die, then X is a discrete random variable because $P(X \in \{1, 2, 3, 4, 5, 6\}) = 1$.

For any random variable, we call the smallest set $\mathcal{X}$ for which $P(X \in \mathcal{X})$ the *support* of X. A discrete random variable has as its support a discrete set.

When X is a discrete random variable, its CDF ends up looking like a staircase: flat everywhere except at each x in its support, where it "jumps" up by an amount $P(X = x)$. For example, for a six-sided die:

Note: The open/closed dots at e.g. $x = 1$ indicate the $F(1)$ is equal to $1/6$, and not 0 (although it is equal to 0 for x arbitrarily close but to the left of 1). We see from this graph why CDFs are right-continuous but not necessarily left-continuous.

At each point in its support $\{1, 2, 3, 4, 5, 6\}$, the CDF for the die jumps by $P(X = x)$, or $1/6$. This is a general feature of discrete random variables. Thus, rather than use the CDF function $F(x)$ to represent the distribution of X, we can just keep track of where it jumps and by how much. To do this, we use *probability mass function* or *p.m.f.* of X

Definition 1.10. *The probability mass function of a random variable X is the function $\pi(x) = P(X = x)$*

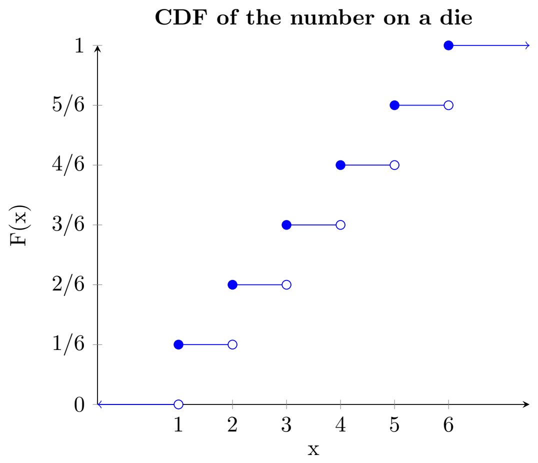


Figure 1.1: The CDF of the number returned by a fair six-sided die.

For a discrete random variable, we can express the p.m.f. alternatively as a sequence, rather than a function. Label the points in the support of X as $\{x_1, x_2, x_3, \dots\}$, in increasing order so that $x_1 < x_2 < x_3 \dots$. Let x_j denote the j^{th} value in this sequence. For any j, let $\pi_j = \pi(x_j) = P(X = x_j)$.

The sequence of probabilities $\{\pi_1, \pi_2, \pi_3, \dots\}$ coupled with the sequence of support points $\{x_1, x_2, x_3, \dots\}$ carries exactly the same information as the full CDF.

Obtaining the p.m.f. from the CDF: For a given support point x_j: $\pi_j = F(x_j) - F(x_{j-1})$, and for any x: $\pi(x) = \lim_{\epsilon \downarrow 0} F(x) - F(x - \epsilon)$. Note that $\pi(x) = 0$ for any x that is not a support point, and F is continuous $\{x_1, x_2, x_3, \dots\}$.

Obtaining the CDF from the p.m.f (only possible for a discrete random variable): $F(x) = \sum_{j: x_j \le x} \pi_j$.

Note that from this last expression, we can see that since $\lim_{x \to \infty} F(x) = 1$, we must have that $\sum_j \pi_j = 1$ – probability mass functions sum to one when the sum is taken across all support points j.

1.3.2.2 Case 2: Continuous random variables and the probability density function

For random variables that are not discrete, knowing the probability mass function isn't sufficient to recover the whole CDF. Often $P(X = x) = 0$ for all x, so the p.m.f does not even really tell us anything useful about X's distribution.

An important class of random variables that are not discrete are random variables for whom the CDF is differentiable for all x. When it is, we can define the *probability density function* or p.d.f. of X.

Definition 1.11. *The probability density function of a random variable X having a differentiable CDF $F(x)$, is $f(x) = \frac{d}{dx} F(x)$.*

We will refer to random variables that have a density function $f(x)$ as *continuous* random variables (another phrasing is that X is *continuously distributed*). Recall that for a function to be differentiable, it must be continuous; thus, the CDF of a continuous random variable must be continuous, lacking any jumps like those that characterize the CDF of a discrete random variable.

Note: you may see in various texts a few different notions of "continuity" of a random variable. For the purposes of this class, a continuous random variable is a random variable with a continuous CDF, which is basically equivalent to it being differentiable everywhere in its support. We won't worry about the distinction between these two things: e.g. random variables with CDFs that are continuous but non-differentiable.

For a continuous random variable we can use the p.d.f rather than the CDF to calculate anything we need to know. For example the probability that X lies in any interval $[a, b]$ can be obtained by integrating over the density function:

$$P(X \in [a, b]) = \int_a^b f(x)dx \tag{1.2}$$

Intuitively, this gives us the area under the curve $f(x)$ between points a and b, as depicted in Figure 1.2. Note that $\int_a^b f(x)dx = F(b) - F(a)$, because the CDF is the anti-derivative of the p.d.f.

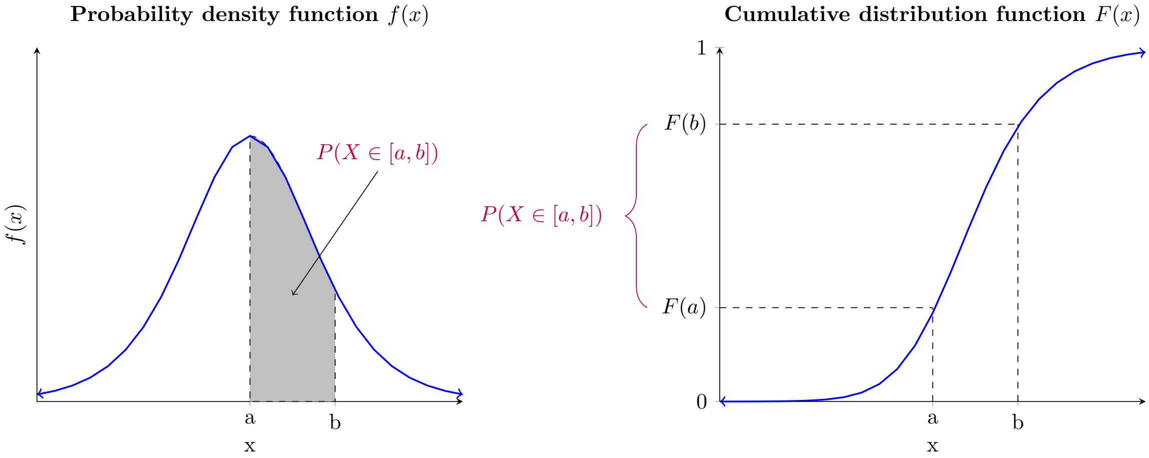


Figure 1.2: The left panel depicts an example of the p.d.f. $f(x)$ of a random variable X. The probability that $a \leq X \leq b$ is given by the area under the $f(x)$ curve between $x = a$ and $x = b$. $P(a \leq X \leq b)$ is also equal to $F(b) - F(a)$, the difference in the CDF of X evaluated at $x = b$ and at $x = a$, as depicted in the right panel.

While the probability mass function $\pi(x)$ gives us the probability that X equals x exactly, the p.d.f does not tell us the probability that $X = x$ (in fact for any x: $P(X = x) = 0$ for a continuous random variable!).

Rather $f(x)$ can be interpreted as telling us the probability that X is close to x, in the following sense. Consider a point x and some small $\epsilon > 0$. Recall the definition of $f(x)$ as the derivative of $F(x)$:

$$f(x) = \frac{d}{dx}F(x) = \lim_{\epsilon \to 0} \frac{F(x + \epsilon) - F(x)}{\epsilon} = \lim_{\epsilon \to 0} \frac{P(X \in (x, \epsilon])}{\epsilon}$$

where we've used Proposition 1.1 to replace $F(x + \epsilon) - F(x)$ with $P(X \in (x, \epsilon])$. Thus $f(x)$ is limit of the ratio of the probability that X lies in a small interval that beings at x, and the width ϵ of that interval. Note also that for small ϵ: $F(x + \epsilon) \approx F(x) + f(x) \cdot \epsilon$, which is called the first-order *Taylor approximation* to $F(x + \epsilon)$ around x.

Let us end this section with a few properties of a probably density function:

- From Eq. (1.2), we see that the density must integrate to one, when the integral is taken over the whole real line, i.e. $\int_{-\infty}^{\infty} f(x)dx = 1$.
- since $F(x)$ is increasing and $f(x)$ is its derivative, $f(x)$ is *positive* everywhere: $f(x) \geq 0$.

1.3.2.3 Case 3 (everything else): mixed distributions

Although most familiar examples of random variables are either discrete or continuous, a given random variable X need not be either. However, a powerful result known as the *Lebesque decomposition theorem* shows that we can combine the two tools we've just developed: the p.m.f. and the p.d.f., to work with any random variable.

Definition 1.12. *Given two random variables X and Y with CDFs F_X and F_Y, a third random variable Z is called a mixture of X and Y if it has a CDF that for some $p \in (0, 1)$ satisfies $F_Z(t) = p \cdot F_X(t) + (1 - p) \cdot F_Y(t)$, for all t.*

The Lebesgue decomposition theorem says that a generic random variable X can be seen as a "mixture" of a discrete random variable and a continuous one, that is

$$F(x) = p \cdot F_{discrete} + (1 - p) \cdot F_{continuous} \tag{1.3}$$

for some $p \in (0,1)$, where $F_{discrete}$ admits of a probability mass function, and $F_{continuous}$ admits of a probability density function (i.e. is differentiable everywhere). The support points of $F_{discrete}$ are often referred to as mass points of F.

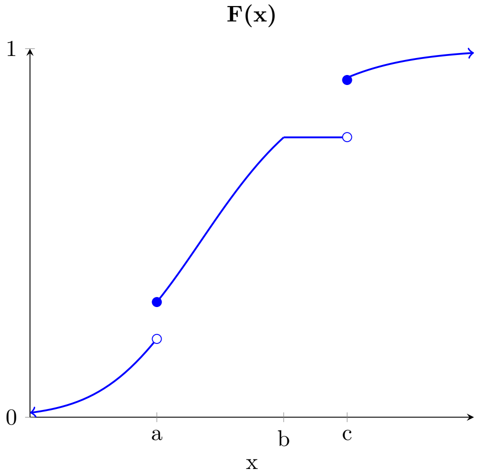


Figure 1.3: An example of the CDF of a mixed random variable. This example has mass points at a and c, where the CDF jumps discretely. It is continuous everywhere else, and is differentiable everwhere except $\{a, b, c\}$.

There are some technical aspects to stating the Lebesque decomposition theorem formally, which we won't explore here. Rather, it's easiest to think of this result visually: a generic CDF is any increasing function bounded between 0 and 1 (which is also right-continuous). The jumps in $F(x)$ define the discrete part of X (note that it can only jump up, and not down, since F is increasing). The function $F(x)$ will be differentiable almost everywhere else, defining it's continuous part.[1]

Note for the interested: to explicitly generate decomposition (1.3), first collect the locations x_j and sizes y_j of each of the jumps $j = 1, 2, \dots$ in $F(x)$. Then $\pi(x_j) = \sum_j y_j$, and $\pi_j = y_j/p$ yields a well-defined p.m.f. function. This characterizes $F_{discrete}$. For any remaining point where $F(x)$ is differentiable, we define a density $f_{continuous}(x) = \frac{1}{1-p}\frac{d}{dx}F(x)$, which characterizes $F_{continuous}$. Note that there may be points at which $F(x)$ doesn't jump, but also isn't differentiable, such as point b in Figure 1.3. We can safely ignore such points, since they are isolated and have probability zero, e.g. $P(X = \{b\}) = 0$.

1.3.3 Marginal and joint distributions

Recall that when we have two random variables X and Y, we have defined the joint CDF $F_{XY}(x, y) = P(X \leq x, Y \leq y)$ as well as the individual CDFs: $F_X(x) = P(X \leq x)$ and $F_Y(y) = P(Y \leq y)$. The functions F_X and F_Y are often referred to as the *marginal distributions* of X and Y.

The following relationships hold between marginal and joint distributions:

- $F_X(x) = F_{XY}(x, \infty) = P(X \leq x, Y \leq \infty) = P(X \leq x)$. Similarly, $F_Y(y) = F_{XY}(\infty, y)$.
- If Y is discrete: $P(X = x) = \sum_j P(X = x \text{ and } Y = y_j)$ where y_j are the support points of Y
- If X and Y are both continuously distributed: $f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)dy$, where the joint density $f_{XY}(x, y)$ is the derivative of the joint CDF with respect to both x and y: $f_{XY}(x, y) = \frac{\partial^2}{\partial x \partial y}F_{XY}(x, y)$.

Intuitively, we can obtain the marginal distribution of X from the joint distribution by summing or integrating over all values of Y, and we can similarly derive the marginal distribution of Y from the joint distribution of X and Y by summing/integrating over values of X.

The above results all follow from a fundamental identity for probabilities called the *law of total probability*:

[1] "Almost everywhere" here has a technical meaning. Any monotonic function is guaranteed to be differentiable everywhere except at isolated points: see Lebesque's theorem for the differentiability of a monotone function.

Proposition (law of total probability): Consider a countable collection of events $A_1, A_2, \dots$ that partition the sample space (this means that the A_j are disjoint and that $\bigcup_j A_j = \Omega$). Then for any event B: $P(B) = \sum_j P(B \cap A_j)$.

Proof. The proof is good practice, so I include it here. Since any event $B \subseteq \Omega$, $B = B \cap \Omega$ and thus $P(B) = P(B \cap \Omega)$. Now, since $\bigcup_j A_j = \Omega$, we have that $P(B) = P\left(B \cap (\bigcup_j A_j)\right)$. Observe that $B \cap (\bigcup_j A_j) = \bigcup_j (B \cap A_j)$, and that the events $(B \cap A_j)$ are disjoint for different values of j (since each is a subset of A_j). Thus, $P(B) = \sum_j P(B \cap A_j)$, proving the result. □

We can use the ideas of marginal and joint distributions to define the notion of *independence* between two random variables:

Definition 1.13. *We say that random variables X and Y are independent if $F_{XY}(x, y) = F_X(x) \cdot F_Y(y)$ for all x and y.*

When X and Y are independent, we denote this fact as $X \perp Y$. When they are not, we say $X \not\perp Y$.

1.3.4 Functions of a random variable

An important property of random variables is that we can apply a function to a random variable, and this results in a new random variable. For example, if we start with a random variable X, and have a function $g : \mathbb{R} \to \mathbb{R}$, then $g(X)$ is also a random variable. For example, $X + 1$ defines a new random variable that is one larger than X for all i.

The reason that we can do this is simple: the original random variable was defined from a function X defined on an underlying outcome space Ω. Evaluating $g(X(\omega))$ for any $\omega \in \Omega$ yields a new function, the so-called composition of g with X (this is often denoted as $g \circ f$).

Technical note: Recall that the function $X(\omega)$ that defines the original random variable X must be a *measurable* function. For the above logic to go through, the function $g(\cdot)$ applied to X must also be measurable, so that the function $g \circ f = g(X(\cdot))$ is also measurable. A sufficient condition for a function to be measurable is that it is piece-wise continuous, which is a very weak condition.

To work with a random variable $Y = g(X)$, we need to know it's CDF, which is:

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y)$$

This RHS expression can always be evaluated using the CDF of X. However, there are two important special cases in which deriving the distribution of Y from that of X is particularly easy:

1. If X has a discrete distribution with support points $x_1, x_2, \dots$ and p.m.f. $\pi_1, \pi_2, \dots$, then Y has the same p.m.f. $\pi_1, \pi_2, \dots$ but at new support points $g(x_1), g(x_2), \dots$.
   - Example: if X is a random variable that takes value 0 with probability p and 1 with probability $1 - p$, then the random variable $Y = X + 1$ is a random variable that takes value 1 with probability p and 2 with probability $1 - p$.
2. (homework problem) If X has a continuous distribution with density $f_X(x)$, and if the function $g(x)$ is strictly increasing and differentiable with derivavtive g', then Y has a density $f_Y(y) = \frac{f_X(g^{-1}(y))}{g'(g^{-1}(y))}$ where g^{-1} is the inverse function of g.
   - Example: if $g(x) = \log(x)$, then $f_Y(y) = f_X(e^y) \cdot e^y$, since $g^{-1}(y) = e^y$ and $g'(x) = 1/x$.

Just as a function applied to a random variable defines a new random variable, functions applied to *multiple* random variables also yield a new random variable. For example, if X and Y are each random variables, then $Z = g(X, Y)$ is also a random variable, where $g(x, y)$ is now a function that takes two arguments. Some examples would be the random variables $X + Y$, $X \cdot Y$, or $\min\{X, Y\}$. When taking a functions of two random variables, $Z = g(X, Y)$, we need the full *joint distribution* of X and Y to derive the CDF of Z. Knowing the two functions $F_X(x)$ and $F_Y(y)$ is generally not enough, rather we need to know the function $F_{XY}(x, y)$ (see Definition 1.8). This will come up later in the course.

1.4 The expected value of a random variable

Main idea: The *expected value* of a random variable is a measure of its average value across realizations. In the special case of a continuous random variable, its value can be obtained by an integral involving the density function. In the special case of a discrete random variable, its value can be obtained by a sum involving the probability mass function.

The expected value (a.k.a. *expectation value*, or simply *expectation*) of a random variable is a measure of its average value over all possible realizations. The expectation of X is denoted $\mathbb{E}[X]$.

To motivate how $\mathbb{E}[X]$ will be defined, think of task of computing the average of a list of numbers. For example, the average of the numbers 1, 2, 2, and 4 is $(1+2+2+4)/4=2$. Notice that the number 2 occurred twice in the series, so we added 2 to the sum two times. We could thus have written the averaging calculation as $\frac{1}{4}(1\cdot 1+2\cdot 2+4\cdot 1)$, where each number is mulitplied by the number of times it occurs in the list. The general formula could be written

$$\text{average of a list of numbers} = \sum_j (j^{th}\text{ distinct number })\cdot \underbrace{\frac{\#\text{ times } j^{th}\text{ distinct number occurs in the list}}{\text{length of the list}}}_{w_j}$$

where notice that "weight" w_j on the j^{th} distinct number sums to one over all j, i.e. $\sum_j w_j = 1$.

The definition of $\mathbb{E}[X]$ for a discrete random variable is exactly analogous to this formula, where we average over the values x_j that X can take, and use as "weights" the probabilities π_j:

$$\mathbb{E}[X] = \sum_j x_j \cdot \pi_j \tag{1.4}$$

where $x_1, x_2, \ldots$ are the distinct support points of the random variable and π_j is it's p.m.f. Note that the π_j sum to one, as we saw in Section 1.3.2.

In the case of a continuous random variable, the analogous expression to Eq. (1.4) replaces the sum with an integral, and the probability $\pi_j = \pi(x_j)$ is replaced by $f(x)dx$:

$$\mathbb{E}[X] = \int x \cdot f(x)dx \tag{1.5}$$

The quantity $f(x)\cdot dx$ can be interpreted as the probability that X lies in an interval $[x, x+dx]$ having a very small width dx, as discussed in Section 1.3.2.

1.4.1 General definition

We now give a general definition of the expectation of a random variable X, and see that Equations (1.4) and (1.5) emerge as simple special cases of it when X is discrete or continuous, respectively.

Definition 1.14. *The expectation of a random variable X having CDF $F(x)$ is $\mathbb{E}[X] = \int_{-\infty}^{\infty} x\cdot dF(x)$, where we define the integral $\int_{-\infty}^{\infty} x\cdot dF(x)$ as*

$$\int_{-\infty}^{\infty} x\cdot dF(x) := \lim_{a\to-\infty, b\to\infty}\lim_{N\to\infty}\sum_{n=1}^{N}\left\{a+n\cdot\frac{b-a}{N}\right\}\cdot\left\{F\left(a+n\cdot\frac{b-a}{N}\right)-F\left(a+(n-1)\cdot\frac{b-a}{N}\right)\right\}$$

The quantity $\int_{-\infty}^{\infty} x\cdot dF(x)$ in an example of a *Riemann–Stieltjes integral*, in which we "integrate'with respect to the function" $F(x)$ rather than with respect to the variable x. Let's try to unpack this long expression, with the aid of the color-coding above.

First, let's fix values of a, b, N and consider the quantity appearing inside all of the limits. For given $b > a$, imagine cutting the interval $[a, b]$ into N regions of equal size, so that they each have width $\frac{b-a}{N}$. The n^{th} such region extends from the value $a+(n-1)\cdot\frac{b-a}{N}$ to the value $a+n\cdot\frac{b-a}{N}$. Note the following:

- $F\left(a+n\cdot\frac{b-a}{N}\right)-F\left(a+(n-1)\cdot\frac{b-a}{N}\right)$ yields $P(X\in\text{ region } n)$.
- $\left\{a+n\cdot\frac{b-a}{N}\right\}$ is the location of (the right end of) region n.

- $\lim_{N\to\infty}$ takes the sum to an integral, and the a, b limit covers full support of X.

Thus, we can interpret $\mathbb{E}[X]$ as an integral of the function x over the whole real line, in which each value of x is multiplied by the probability that X is very close to x, essentially $F(x+dx) - F(x)$.

Discrete case: Now let's see how Definition 1.14 yields Eq. (1.4) in the special case that X is a discrete random variable. Let $x_1, x_2 \dots$ be the support points of X. Notice that for large enough N, only one x_j can be between $a + \frac{n-1}{N}(b-a)$ and $a + \frac{n}{N}(b-a)$. Thus: $F\left(a + \frac{n}{N}(b-a)\right) - F\left(a + \frac{n-1}{N}(b-a)\right) = \pi_j$ if x_j lies in the n^{th} region. If on the other hand no x_j lies in the n^{th} region, this quantity is equal to zero. We arrive at one term for each value x_j, and $\mathbb{E}[X] = \sum_j x_j \cdot \pi_j$.

Continuous case: When X is a continuous random variable with density $f(x)$, we can recover Eq. (1.4) by noticing that for large N:

$$F\left(a + \frac{n}{N}(b-a)\right) - F\left(a + \frac{n-1}{N}(b-a)\right) \approx f\left(a + \frac{n}{N}(b-a)\right) \cdot \frac{b-a}{N}$$

Substituting in this approximation delivers the familiar formula that $\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x) dx$.

Exercise: Consider a so-called *Bernoulli random variable* X that takes a value 1 with probability p and 0 with probability $1-p$. Show that $\mathbb{E}[X] = p$.

Exercise: Consider a *uniform* $[0,1]$ *random variable*, that is a continuous random variable with density $f(x) = x$ for all $0 \le x \le 1$, and $f(x) = 0$ everywhere else. Show that $\mathbb{E}[X] = 1/2$.

A key property of the expectation operator that is very useful is that it is *linear*. It's actually "linear" in a few distinct senses:

1. *Linearity with respect to functions of a single variable:* $\mathbb{E}[a + b \cdot X] = a + b \cdot \mathbb{E}[X]$
2. *Linearity over sums of random variables:* $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.
3. *Linearity with respect to mixtures:* if X, Y and Z are random variables such that $F_Z(t) = p \cdot F_X(t) + (1-p) \cdot F_Y(t)$, then $\mathbb{E}[Z] = p \cdot \mathbb{E}[X] + (1-p) \cdot \mathbb{E}[Y]$.

Note that because of Property 2, we can compute the expectation value of the random variable $X+Y$ knowing only the CDFs $F_X(x)$ and $F_Y(y)$, without needing the full joint-CDF $F_{XY}(x,y)$ of X and Y. This is a very special property of the expectation, which doesn't hold for most of the things we might want to know about the random variable $X+Y$ (for example $P(X+Y \le t)$).

Property 3. gives us a nice way to evaluate the expectation value of a random variable that is neither discrete nor continuous. Recalling decomposition (1.3) of a general mixed random variable, let $f^c(x)$ be the density of the continuous part $F_{continuous}$ and let x_j^d and π_j^d denote the support points and associated probabilities according to the discrete part $F_{discrete}$. Then:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot dF(x) = p \cdot \left\{ \sum_j x_j^d \cdot \pi_j^d \right\} + (1-p) \cdot \int_{-\infty}^{\infty} x \cdot f^c(x) \cdot dx$$

1.4.2 Application: variance

From the expectation operator, we can also define the *variance* of a random variable, which measures how "dispersed" it is. We'll see that the variance plays an important role in asymptotic theory.

Definition 1.15. *The variance of X is the expected value of the random variable $(X - E[X])^2$, i.e. $Var(X) := E[(X - E[X])^2]$.*

The variance of X can be interpreted as the average value of the squared distance between X and its expectation $\mathbb{E}[X]$. Note that $Var(X) \ge 0$ for any random variable, with $Var(X) = 0$ only when X takes one value with probability one (i.e. X is a so-called *degenerate random variable*).

Exercise: Use the linearity of the expectation operator to prove the following (very useful) alternative expression for the variance: $Var(X) = E[X^2] - (E[X])^2$.

Exercise: Show that for a Bernoulli random variable (defined above), the variance is equal to $p \cdot (1-p)$.

1.5 Conditional distributions and expectation

In this section we develop a final fundamental tool that we will use to analyze random variables: the idea of *conditional* distributions and *conditional* expectations.

> **Main idea:** *Conditioning* on an event allows us to examine a restricted probability space in which that event is true (but other things are still random). When this idea is applied to random variables, we can define *conditional distributions* that we can work with in all of normal ways.

1.5.1 Conditional probabilities

We begin with a concept that applies to all probability spaces, not just to random variables.

Definition 1.16. *Given an event B such that $P(B) > 0$, the conditional probability of event A given B is defined as*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

where recall that the intersection of two events $A \cap B$ can be interpreted as the event that *both* of events A and B occur. This definition is often referred to as *Bayes' rule*. You can think of Bayes' rule as a way to define a probability function using B as the whole outcome space: yielding a way to talk about the probability that $\omega \in A$, *given* that $\omega \in B$.

Extension: Given events A, B and C, we can also define the probability of A given B *and* C as $P(A|B \cap C) = P(A \cap B \cap C)/P(B \cap C)$, and so on for any number of events.

Exercise: We call events A and B *independent* if $P(A \cap B) = P(A) \cdot P(B)$. Suppose that $P(B) > 0$. Show that A and B are independent if and only if $P(A|B) = P(A)$.

1.5.2 Conditional distributions

Consider now two random variables X and Y.

Definition 1.17. *The conditional CDF of Y given $X = x$ is*

$$F_{Y|X=x}(y) := P(Y \leq y | X = x) := \lim_{\epsilon \downarrow 0} P(Y \leq y | X \in [x, x+\epsilon])$$

where the conditional probability appearing in the RHS is defined by Definition 1.16.

Notation: The conditional CDF will sometimes also be denoted as $F_{Y|X}(y|X)$.

We define $P(Y \leq y|X = x)$ using $X \in [x, x+\epsilon]$ as our conditioning event B, and then taking the limit, because the probability of $X = x$ may be zero, e.g. for a continuously distributed X. *Note:* The Hansen book uses $x \in [x-\epsilon, x+\epsilon]$ instead of $x \in [x, x+\epsilon]$, but the two definitions are equivalent.

Given the general Definition 1.17, we can consider each of our two typical special cases:

- When $P(X = x) > 0$ (e.g. for a discrete random variable with a support point at x), Definition 1.17 reduces to the simpler expression $P(Y \leq y|X = x) = \frac{P(Y \leq y \text{ and } X=x)}{P(X=x)}$. We can interpret $F_{Y|X=x}(y)$ as the CDF among the sub-population of i for which $X = x$.

- If on the other hand $f_X(x) = \frac{d}{dx}F_X(x)$ exists (e.g., for a continuous random variable), then Definition 1.17 simplifies to $P(Y \leq y|X = x) = \frac{\frac{d}{dx}P(Y \leq y, X \leq x)}{f_X(x)}$. We can interpret $F_{Y|X=x}(y)$ as the CDF among the sub-population of i for which X is "very close" to x.

Exercise: derive each of these two expressions from Definition 1.17 . For the discrete case, you may find useful the "quotient rule" that $\lim_{t \to 0} \frac{g(t)}{h(t)} = \frac{\lim_{t\to 0} g(t)}{\lim_{t \to 0} h(t)}$ when both limits exist and $\lim_{t \to 0} h(t) \neq 0$. For the continuous case, try dividing both the numerator and the denominator of $P(Y \leq y|X \in [x, x+\epsilon])$

by ϵ before taking the limit.

Exercise: Show that if X and Y are independent then $F_{Y|X=x}(y) = F_Y(y)$ and $F_{X|Y=y}(x) = F_X(x)$ for all x and y. Note: it's actually an if-and-only-if, but proving the other direction is more difficult.

1.5.3 Conditional expectation (and variance)

Consider a fixed value of x, and view the conditional CDF $F_{Y|X=x}(y)$ as a function of y. This function satisfies the four properties of a CDF mentioned in Section 1.3.1: it is weakly increasing, right-continuous, and ranges from zero to one.

Thus, we can define the expectation over this distribution in exactly the same way as we would for $\mathbb{E}[Y]$ based on Definition 1.14, except that use $F_{Y|X=x}(y)$ as the CDF rather than it's "unconditional" analog $F(y)$. We can write this using the general notation of Definition 1.14 as:

$$\mathbb{E}[Y|X=x] = \int_{-\infty}^{\infty} y \cdot dF_{Y|X=x}(y)$$

We can unpack this expression depending on what type of random variable Y is:

- If Y is continuous: $\mathbb{E}[Y|X=x] = \int_{-\infty}^{\infty} y \cdot f_{Y|X=x}(y) \cdot dy$, where $f_{Y|X=x}(y) = \frac{d}{dy}F_{Y|X=x}(y)$.
- If Y is discrete: $\mathbb{E}[Y|X=x] = \sum_j y_j \cdot \pi_{j|X=x}$, where $\pi_{j|X=x} = \lim_{\epsilon \downarrow 0} \left\{F_{Y|X=x}(y_j) - F_{Y|X=x}(y_j - \epsilon)\right\}$.

Observe that the conditional expectation $\mathbb{E}[Y|X=x]$ depends on x only, as we've averaged over various values of Y. Accordingly, we can define a function that evaluates $\mathbb{E}[Y|X=x]$ over different values of x:

Definition 1.18. *The conditional expectation function (CEF) of Y given X is $m(x) := \mathbb{E}[Y|X=x]$.*

We can also use the CEF to define a new random variable, denoted $\mathbb{E}[Y|X]$.

Definition 1.19. $\mathbb{E}[Y|X] = m(X)$*, where* $m(x) := \mathbb{E}[Y|X=x]$.

For example, if X is discrete, then $\mathbb{E}[Y|X]$ takes value $m(x_j) = \mathbb{E}[Y|X=x_j]$ with probability π_j.

The so-called *law of iterated expectations* shows that the expectation value of $\mathbb{E}[Y|X]$ recovers the (unconditional) expectation of Y:

Proposition (law of iterated expectations): $\mathbb{E}[Y] = \mathbb{E}\left[\mathbb{E}[Y|X]\right]$

Proof. We prove it for the case in which both X and Y are continuous random variables. The other cases are analagous.

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}[Y|X]\right] &= \int_{x\in\mathbb{R}:f_X(x)>0} f_X(x) \cdot \mathbb{E}[Y|X=x] \cdot dx \\
&= \int_{x\in\mathbb{R}:f_X(x)>0} f_X(x) \cdot \left\{\int_{y\in\mathbb{R}} y \cdot f_{Y|X}(y|x) \cdot dy\right\} \cdot dx \\
&= \int_{x\in\mathbb{R}:f_X(x)>0} \cancel{f_X(x)} \cdot \left\{\int_{y\in\mathbb{R}} y \cdot \frac{f_{XY}(x,y)}{\cancel{f_X(x)}} \cdot dy\right\} \cdot dx \\
&= \int_{y\in\mathbb{R}} y \cdot \underbrace{\left\{\int_{x\in\mathbb{R}:f_X(x)>0} f_{XY}(x,y) \cdot dx\right\}}_{} = f_Y(y) \cdot dy = \int_{y\in\mathbb{R}} y \cdot f_Y(y) \cdot dy = \mathbb{E}[Y]
\end{aligned}$$

□

The law of iterated expectations is useful because in many settings the quantity $\mathbb{E}[Y|X=x]$ is easier to work with than $[Y]$ is directly.

Example: Suppose that Y is individual i's height and X is an indicator for whether they are a child or an adult. Then the law of iterated expectations tells us that the average height in the population can be obtained by averaging together the mean height among children with the mean height among adults. Suppose that 75% of the population are adults. Then the law of iterated expectations reads as:

$$\mathbb{E}[height] = .75 \cdot \mathbb{E}[height|adult] + .25 \cdot \mathbb{E}[height|child]$$

Proposition (CEF minimizes mean squared prediction error): Suppose we're interested in constructing a function $g(\cdot)$ with the goal of using $g(X)$ as a prediction of Y. We can show that $m(x) := \mathbb{E}[Y|X=x]$ is the best such function, in the sense that for each value of x

$$m(x) = \text{argmin}_g E[Y - g(X))^2]$$

Proof. Here I'll use the general notation so we don't need to make any assumptions about what type of random variable X is (discrete, continuous, etc.):

$$\begin{aligned}\mathbb{E}[(Y-g(X))^2] = \mathbb{E}\left\{\mathbb{E}[(Y-g(X))^2|X]\right\} &= \int \mathbb{E}[(Y-g(X))^2|X=x]\cdot dF(x) \\ &= \int \mathbb{E}[(Y-g(x)^2|X=x]\cdot dF(x) = \int \mathbb{E}[Y^2 - 2Yg(x) + g(x)^2|X=x]\cdot dF(x) \\ &= \int \left\{\mathbb{E}[Y^2|X=x] - 2g(x)E[Y|X=x]g(x) + g(x)^2\right\}\cdot dF(x)\end{aligned}$$

For each value of x, the quantity in brackets is minimized by $g(x) = E[Y|X=x]$. To see this, note that the quantity $\mathbb{E}[Y^2|X=x] - 2gE[Y|X=x]g + g^2$ is a convex function of g, and the first-order condition for minimizing it is satisfied when $g = E[Y|X=x]$. □

We can also define a *conditional variance* function $Var(Y|X=x) = E[(Y - E[Y|X=x])^2|X=x]$ from the conditional distribution $F_{Y|X=x}$. An analog to the law of iterated expectations exists for the conditional variance, which is sometimes called the *law of total variance*.

Proposition (law of total variance): $Var(Y) = E[Var(Y|X)] + Var(E[Y|X])$.

Example: Recall the height example from the law of iterated expectations. The law of total variance reveals that the variance of heights in the population overall is *greater* than what we would get by just averaging the variances of each subgrup. That is:

$$Var(height) > .75 \cdot Var(height|adult) + .25 \cdot Var(height|child)$$

The reason is that $Var(height)$ involves making comparisons directly between the heights of children and adults, which are not captured in $Var(Y|X=x)$ for either value of x. The law of total variance tells us exactly what correction we would need to make, which is to add the second term $Var(\mathbb{E}[Y|X])$. Remarkably, the correction required just depends on the *average* height within each group $\mathbb{E}[Y|X=x]$, as well as the proportion of adults vs. children: $P(X=x)$.

1.6 Random vectors and random matrices

Main idea: *Random vectors* are vectors in which each component is a random variable, and *random matrices* are matrices where each entry is a random variable. These concepts allow us to define the expectation, variance, and covariance between random vectors, which gives us a compact notation to discuss many random variables at the same time.

1.6.1 Definition

Rather than coming up with new letters X, Y, Z for multiple random variables, sometimes a more compact notation is to think of a single "random vector" containing all three.

Definition 1.20. *A random vector X is a vector in which each component is a random variable, e.g.*

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}$$

where X_1, X_2, etc. are each random variables.

Note the following:

- A realization $\mathbf{x}$ of random vector X is a point in $\mathbb{R}^k$, i.e. $\mathbf{x} = (x_1, x_2, \ldots, x_k)'$:

$$P(X = \mathbf{x}) = P(X_1 = x_1 \text{ and } X_2 = x_2 \ldots \text{ and } \ldots X_k = x_k)$$

- For a random vector X, the function F_X denotes the joint-CDF of the random variables $X_1, X_2, \ldots X_k$:

$$F_X(\mathbf{x}) = P(X_1 \leq x_1 \text{ and } X_2 \leq x_2 \ldots \text{ and } \ldots X_k \leq x_k)$$

- The expectation of a random vector X is simply the vector of expectations of each of its components, i.e.

$$E[X] = \left[\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}\right] := \begin{pmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_k] \end{pmatrix}$$

- The law of iterated expecations $\mathbb{E}[Y] = \mathbb{E}\left[\mathbb{E}[Y|X]\right]$ still holds when X is a random vector, rather than a random variable.

.

Definition 1.21. *An $n \times k$ random matrix $\mathbf{X}$ is a matrix in which each component is a random variable, e.g.*

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1k} \\ X_{21} & X_{22} & \ldots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{nk} \end{pmatrix}$$

where X_{lm}, is a random variable for each entry lm.

Just as with a random variable, we define the expectation of a random matrix as a matrix composed of the expectation of each of it's components, i.e.

$$\mathbb{E}[\mathbf{X}] = \begin{pmatrix} \mathbb{E}[X_{11}] & \mathbb{E}[X_{12}] & \ldots & \mathbb{E}[X_{1k}] \\ \mathbb{E}[X_{21}] & \mathbb{E}[X_{22}] & \ldots & \mathbb{E}[X_{2k}] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[X_{n1}] & \mathbb{E}[X_{n2}] & \ldots & \mathbb{E}[X_{nk}] \end{pmatrix}$$

This allows us to generalize the notion of variance to random vectors.

Definition 1.22. *The variance of a random vector X is $Var(X) = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])']$*

where use the notation that for a vector $\mathbf{x}$: $\mathbf{x}'$ indicates its transpose $(x_1, x_2, \ldots x_k)$. Note that for vectors $\mathbf{x} = (x_1 \ldots x_n)'$ and $\mathbf{y} = (y_1 \ldots y_k)$, $\mathbf{x}\mathbf{y}'$ is an $n \times k$ matrix, where the lm component of $\mathbf{x}\mathbf{y}'$ is $x_l \cdot y_m$. We will also use $'$ to denote the matrix transpose, i.e. $[X']_{lm} = X_{ml}$.

Note that when X is a random vector rather than a random variable, $Var(X)$ is often referred to as the "variance-covariance matrix" of X. We'll use the variance-covariance matrix a lot, because it plays an important role in studying parametric distributions like the multivariate normal distribution, and in asymptotic theory.

To understand the name, let us first define the *covariance* between random vectors X and Y:

Definition 1.23. *The covariance of random vectors X and Y is $Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])']$*

Note the following properties of covariance:

- For random vector X: $Var(X) = Cov(X, X)$
- When X and Y are scalars (i.e. single random variables), $Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$
- For scalar X and Y, and numbers a, b: $Cov(X, a + bY) = b \cdot Cov(X, Y)$

- For a random vector X, the components of the matrix $Var(X)$ are scalar variance and covariances, hence its name:

$$Var(X) = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \dots & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Var(X_2) & \dots & Cov(X_2, X_k) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_k, X_1) & Cov(X_k, X_2) & \dots & Var(X_k, X_k) \end{pmatrix}$$

  A consequence of this expression is that $Var(X)$ is a *symmetric* matrix: $[Var(X)]_{lm} = [Var(X)]_{ml}$, because $Cov(X_l, X_m) = Cov(X_m, X_l)$.

- When X and Y are scalars, we can define the *correlation coefficient* ρ_{XY} as $\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$ (note that all quantities involved here are scalars). ρ_{XY} is always a number between -1 and $+1$ (homework problem).

Exercise: Show that $Cov(X, Y) = \mathbb{E}[XY'] - \mathbb{E}[X]\mathbb{E}[Y]'$

1.6.2 Conditional distributions with random vectors

1.6.2.1 Conditioning on a random vector

In Section 1.5 we defined the conditional distribution of one random variable Y given another random variable X. This idea extends naturally to conditioning a random variable Y on multiple random variables at the same time, e.g. $F_{Y|X=x,Z=z}(y)$. Random vectors give us a nice notation for this:

Definition 1.24. *With X a random vector, the conditional CDF of random variable Y given $X = \mathbf{x}$ is*

$$F_{Y|X}(y|\mathbf{x}) = \lim_{\substack{\epsilon_1 \downarrow 0 \\ \epsilon_2 \downarrow 0 \\ \dots \\ \epsilon_k \downarrow 0}} P(Y \leq y | X_1 \in [x_1, x_1 + \epsilon_1], X_2 \in [x_2, x_2 + \epsilon_2] \dots X_k \in [x_k, x_k + \epsilon_k])$$

where $\mathbf{x} = (x_1, x_2, \dots x_k)'$.

We can always use the above definition, even if the components of X can be a mix of continuous and discrete random variables.

For any given value of $\mathbf{x}$, $F_{Y|X=\mathbf{x}}(y) = F_{Y|X}(y|\mathbf{x})$ yields a proper CDF function for y, which means we can continue to define the conditional expectation as $\mathbb{E}[Y|X = \mathbf{x}] = \int_{\infty}^{\infty} y \cdot dF_{Y|X=\mathbf{x}}(y)$, where the meaning of this integral is as given in Definition 1.14. The conditional variance of Y given $X = \mathbf{x}$ can also be defined in the typical way from the conditional distribution $F_{Y|X=\mathbf{x}}(y)$.

The law of iterated expectations caries over unchanged when X is a random vector. That is: $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$, regardless of whether X has continuous or discretely distributed components, or a mix of the two. The law of total variance caries over too (see below).

Understanding and estimating the object $\mathbb{E}[Y|X = \mathbf{x}]$ from data, where X can be a vector, will be one of our main interests in this course, motivating the use of regression analysis. Take a deep breath, we made it!

1.6.2.2 The conditional distribution of a random vector

This section can be skipped for now, but later in the course we'll need to talk about joint-distribution of a random vector, conditional on the value of one or more other random variables.

When *both* X and Y are random vectors, we can talk about the conditional distribution of Y given X by defining a *conditional joint-CDF* of all the components of Y, conditional on $X = \mathbf{x}$.

Definition 1.25. *With X and Y random vectors, the conditional CDF of Y given $X = \mathbf{x}$ is*

$$F_{Y|X}(\mathbf{y}|\mathbf{x}) = \lim_{\substack{\epsilon_1 \downarrow 0 \\ \epsilon_2 \downarrow 0 \\ \dots \\ \epsilon_k \downarrow 0}} P(Y_1 \leq y_1, Y_2 \leq y_2, \dots | X_1 \in [x_1, x_1 + \epsilon_1], X_2 \in [x_2, x_2 + \epsilon_2] \dots X_k \in [x_k, x_k + \epsilon_k])$$

where $\mathbf{x} = (x_1, x_2, \dots x_k)'$.

An important application of the concept of a conditional joint-distribution is the idea of conditional independence.

Definition 1.26 (conditional independence). *We say that X and Y are independent conditional on Z, denoted $(X \perp Y)|Z$, if for any value z of Z: $F_{XY|Z=z}(x,y) = F_{X|Z=z}(x) \cdot F_{Y|Z=z}(y)$ for all x, y.*

This definition can be understood by using Definition 1.25 to define interpret $F_{XY|Z=z}(x,y)$ as the joint-CDF of a random vector composed of X and Y, conditional on the random vector Z. In this definition X and Y could be random variables or can each be random vectors themselves!

As another application of Definition 1.25, the *law of total covariance* provides an analog of the law of iterated expectations for covariance (and hence, as a special case, for variance):

Proposition 1.2. *For random vectors X, Y and Z: $Cov(X,Y) = \mathbb{E}[Cov(X,Y|Z)] + Cov(\mathbb{E}[X|Z], \mathbb{E}[Y|Z])]$*

Note that as a special case we have the *law of total variance*, that: $\mathbb{E}[Var(Y|X)] + Var(\mathbb{E}[Y|X])$.

Chapter 2

Empirical illustration: National Longitudinal Survey of Young Working Women

2.1 Introduction

Let's illustrate some of the concepts from the last chapter with an empirical example. I'll be working with the `nlswork` dataset, which reports a sample of young working women from the Bureau of Labor Statistics' National Longitudinal Survey in the 1970s and 1980s. Obviously this data is pretty old, but it's easy to load into R, and has a lot of interesting variables. I'll be showing code in R, but the dataset is also easy to load into Stata using the command `wenuse nlswork`.

You can install R by downloading it from https://www.r-project.org/. If you do that, I also recommend installing RStudio from https://www.rstudio.com/ for a nicer interface. You can also create a free acount at `rstudio.cloud` to work with RStudio in the cloud, for low-intensity applications like this one.

To get started and load `nlswork` dataset into RStudio, run the following code:

```
library(webuse)    #For importing the dataset
library(ggplot2)   #For plotting
library(Rmisc)     #Needed for multiplot

df<-webuse("nlswork")
```

The first three lines load R libraries that we'll need. The first allows to load the dataset directly from the web using the `webuse()` command. The second two will be useful for creating pretty plots. Note that `#` in R introduces a comment, which allows one to annotate their code with things that R ignores when you run it.

Note: The `library()` command loads a R library up, but you need to install a library before you load it up. To install the packages that are loaded above, run `install.packages("webuse,ggplot2,Rmisc")` in R. You only need to do this once, then you can just jump straight to `library(webuse)`, etc..

The last line of the code block above reads in the dataset and stores it in a dataframe called `df`. A dataframe is what R calls a dataset. I could have given it any name, but I chose "df", which is usually what I use by default. To take a look at this dataset, you can type `View(df)` into R after running the above code. If you'd like to learn more about the variables, see here: https://rdrr.io/rforge/sampleSelection/man/nlswork.html.

2.2 The "empirical distribution" of a dataset

When you look at the dataframe `df`, you'll notice that the first two columns are called `idcode` and `year`. This survey tracks workers over several years, so each value of `idcode` shows up once for each year in which that worker was surveyed. Altogether, there are $28,534$ rows covering about $4,711$ distinct workers. There are 25 variables reported for each row.

Let us index values of each of the 25 variables recorded in row i with a subscript i, e.g. $\texttt{age}_i$ denotes the age recorded in row i. This is the age of a particular person indexed by $\texttt{id\_code}_i$ in the particular year $\texttt{year}_i$.

Now consider the following probability space: we draw a single row ω from the dataset at random, with an equal probability $P(\{\omega\}) = \frac{1}{28{,}534}$ of selecting any given row (the sample space is finite: $|\Omega| = 28,534$, so we can let our event space F be the full powerset of Ω). We have 25 random variables represented in each row, which define 25 random variables $X_1(\omega), X_2(\omega), \dots X_{25}(\omega)$, with names like `idcode`, `year`, `birth_yr`, `age`, `race`, etc.

What I'll call the *empirical distribution* of the dataset is simply the joint-distribution of our random vector $X = (X_1, X_2, \dots X_{25})'$ given the probability space described above. For example, the CDF of `age` evaluated at 25 is:

$$P(\texttt{age} \leq 25) = \frac{\#\text{ rows in which } \texttt{age}_i \leq 25}{\#\text{ rows in dataset}},$$

i.e. the proportion of the 28,534 rows in which the age recorded is less than or equal to 25. This section will use the empirical distribution of this NLS dataset as an example of the various concepts covered in the last chapter.

Note: In Chapter 4 we'll make a big deal out of distinguishing our *sample* from the underlying *population* of interest, which in this case might be the population all all young working women in the U.S. in the 1970-1980s, from which our sample is drawn. Given this distinction, one could view the various distribution functions plotted in this chapter as estimates of the corresponding distributions in the population. But this is not important for the present purpose, which is simply to illustrate some properties of distributions in general.

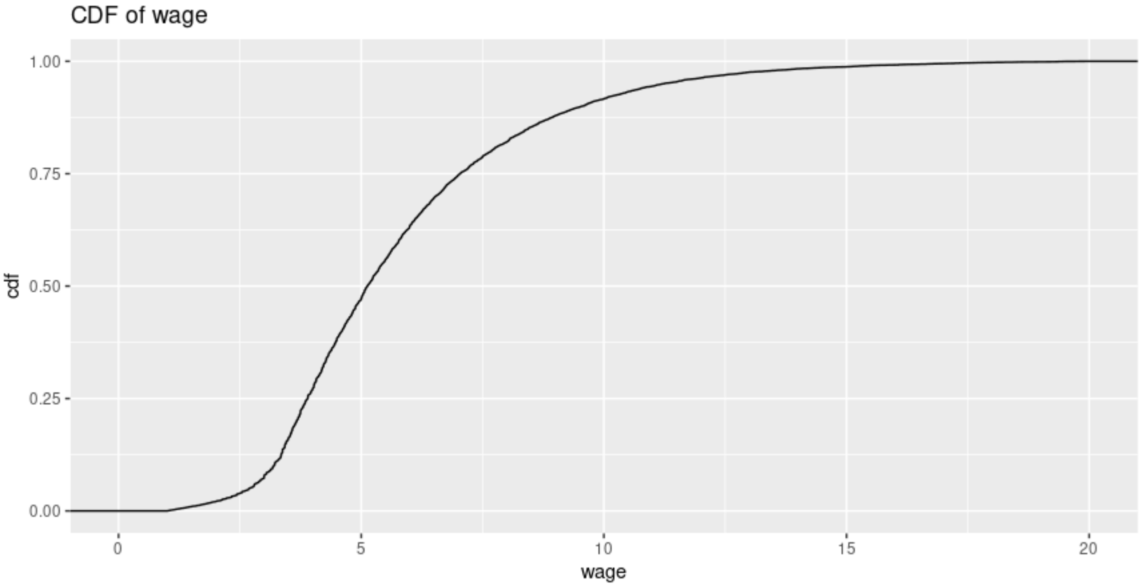
2.3 Examples of conditional distributions and the law of iterated expectations

With our dataset loaded and our probability space well-defined, we can start to consider any of the quantities we defined for random variables in Chapter 1.

Let's start with the (marginal) CDF of a single variable. We'll do this by relying on the `stat_ecdf()` command of the `ggplot` library.

```
#Define a new variable called wage:
df$wage<-exp(df$ln_wage)

#Plot unconditional CDF wage of wage:
ggplot(df, aes(wage)) + stat_ecdf(geom = "step") + labs(title="CDF
    of wage",y="cdf") + xlim(0, 20)
```

The goal of the above code block is to plot the CDF of hourly wages in the dataset. The dataset contains only the log of wages (what labor economists often focus on), so our first task above is to generate a new column in the dataset with the wage, rather than its natural logarithm. This is done with the command `df$wage<-exp(df$ln_wage)`, which adds a new column `wage` to the dataset that equals to e to the power of `ln_wage`. The second command generates the following figure:



The syntax of `ggplot` takes some getting used to, and really the best way to learn it is just to look at examples like this one and start playing with them. The important things to recognize in the above is that `ggplot(df, aes(wage))` tells `ggplot` that we're going to be looking at variable `wage` in dataframe `df`, `stat_ecdf` tells it to plot the empirical CDF, and the last two parts of the line set the plot labels and the range of the x-axis.

As you can see, the CDF of wages is a monotonically increasing function that ranges from 0 to 1. It makes sense to focus on the range \$0 to \$20 because the CDF is essentially flat for wages above \$20.

Now let's consider the conditional CDF $F_{wage|collgrad}$, where `collgrad` $= 1$ indicates that i graduated college, and `collgrad` $= 0$ indicates that they did not. Since `collgrad` is a discrete random variable, the function

$$F_{wage|collgrad=c}(w) = \frac{P(wage \leq w \text{ and } collgrad = c)}{P(collgrad = c)}$$

for any value $c \in \{0, 1\}$ and wage value w. Given the empirical distribution of our data, this is equivalent to

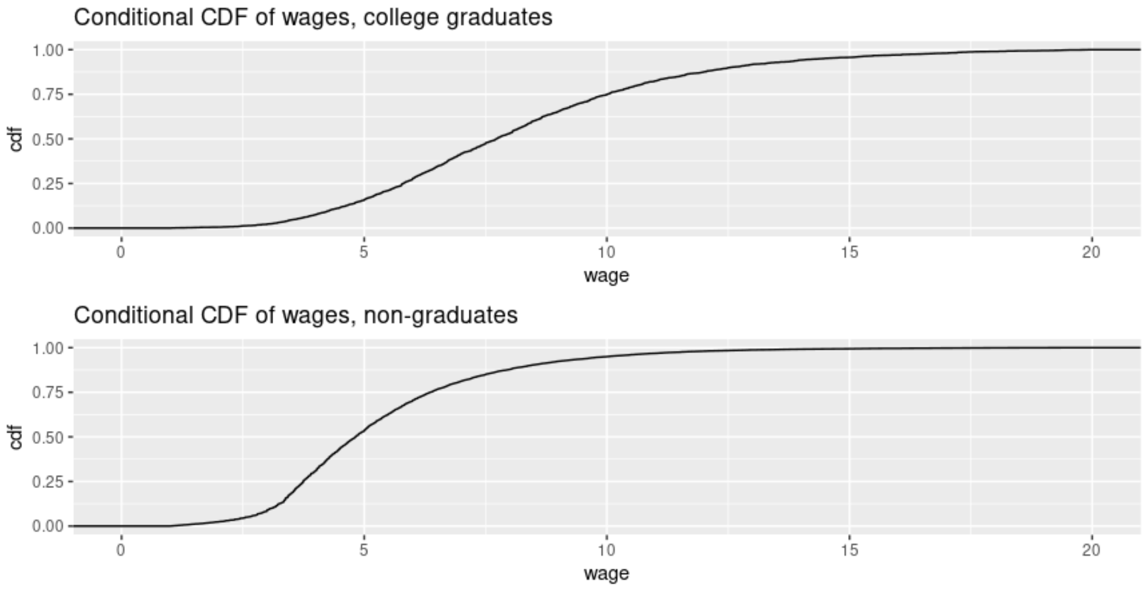
$$F_{wage|collgrad=c}(w) = \frac{\#\text{ rows in which wage} \leq w \text{ and collgrad} = c}{\#\text{ rows in which collgrad} = c} \tag{2.1}$$

Consider for example college graduates: $c = 1$. We can calculate this conditional CDF by creating a new dataset composed of just the rows from `df` in which $collgrad = 1$, and then computing the empirical CDF of wages with respect to that new dataset. The reason is that the empirical CDF with respect to this new dataset counts the number of rows in which wage $\leq w$ (which is the number of rows in the original dataset in which wage $\leq w$ *and* $collgrad = 1$) and divides by the number of rows in the new dataset (which is the number of rows in the original dataset in which $collgrad = 1$). This exactly recovers Eq. (2.1) above.

In general, conditioning on a discrete random variable in an empirical distribution is identical to simply sub-setting the data. Thus we can generate plots of our conditional CDFs for $c = 0$ and $c = 1$ as follows:

```
dfgrads<-df[df$collgrad==1,]
plot1<-ggplot(dfgrads, aes(wage)) + stat_ecdf(geom = "step") +
   labs(title="Conditional CDF of wages, college graduates",y="cdf
   ") + xlim(0, 20)

dfnongrads<-df[df$collgrad==0,]
plot2<-ggplot(dfnongrads, aes(wage)) + stat_ecdf(geom = "step") +
   labs(title="Conditional CDF of wages, non-graduates",y="cdf") +
    xlim(0, 20)
multiplot(plot1, plot2)
```

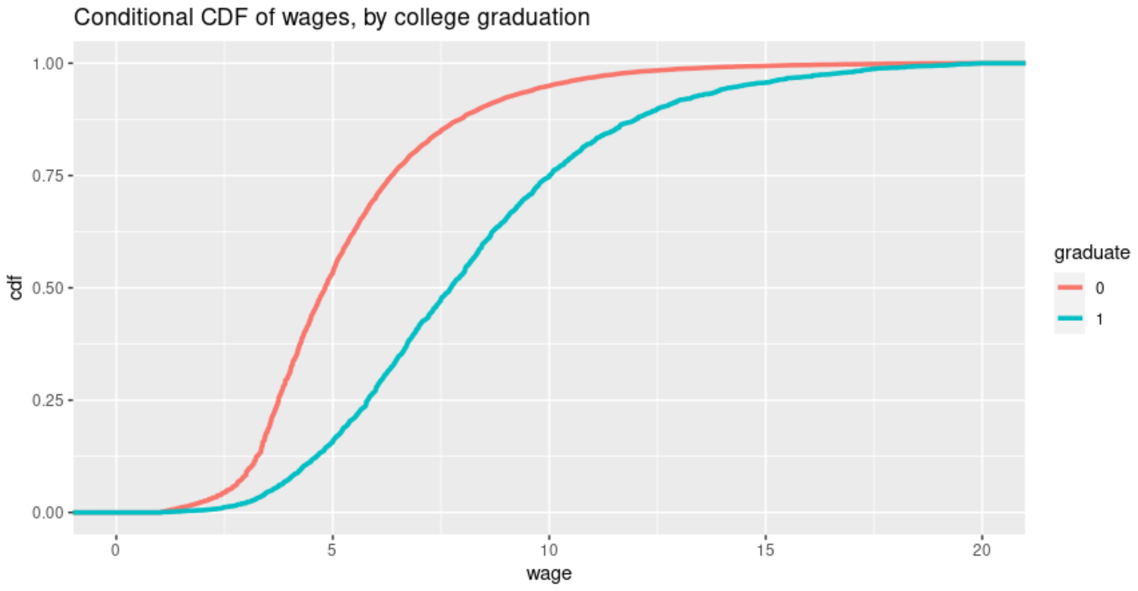
The command `dfgrads<-df[df$collgrad==1,]` for example creates a new dataframe `dfgrads` which contains only the college grduates, and the next line uses the same command as before to generate the empirical CDF of wages in `dfgrads`. Rather than displaying it write away, we save the graph object as `plot1`, so that we can display the two conditional CDFs alongside one another using the `multiplot` command. The result is the following:



`ggplot` makes it easy to automatically compute these conditional CDFs and plot them alongside one-another in the same plot. For example, the following code:

```
df$graduate = factor(df$collgrad)
ggplot(df, aes(x = wage, color = graduate)) + stat_ecdf(lwd =
    1.25, geom = "step") + labs(title="Conditional CDF of wages, by
     college graduation",y="cdf") + xlim(0, 20)
```

generates the combined plot:



This requires creating a so-called “factor variable”, which is a variable that R knows takes on discrete categorical values. The first line creates a factor version of `collgrad` and calls it `graduate`. Then the syntax `color = graduate` tells `ggplot` to break up the data and color it acording to values of `graduate`.

Notice that the CDF of college graduates is lower than the CDF of college non-graduates at every wage value. For example, about 50% of non-college graduates have a wage of $5 or less, while only about 20% of college graduates have a wage of $5 or less. This is what we should expect, if college graduates tend to be paid better than non-graduates.

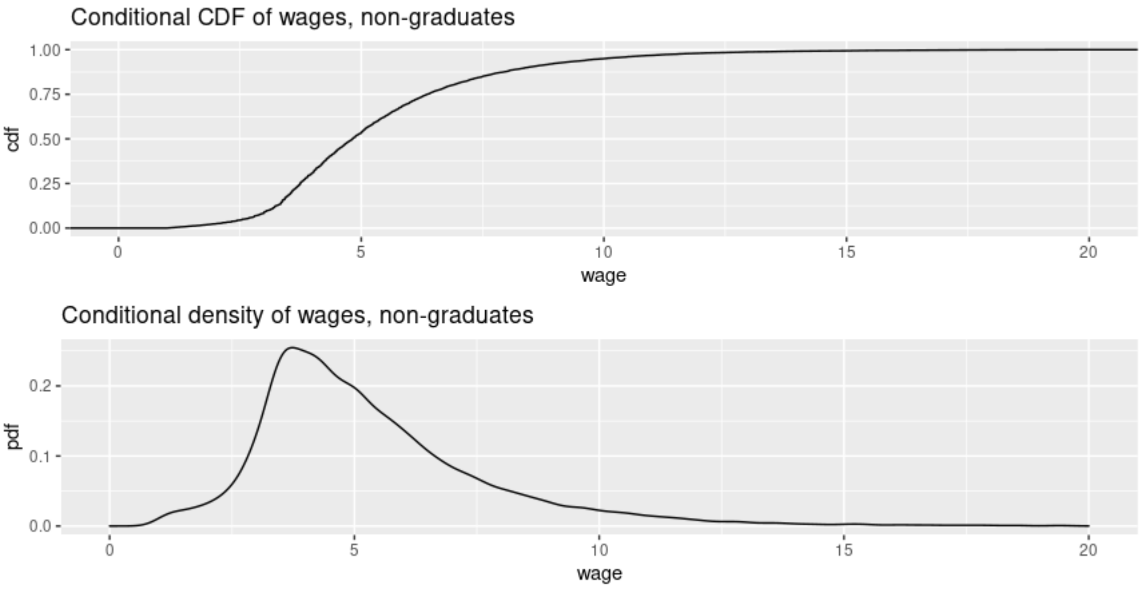
Is our wage variable a continuous or a discrete random variable? Notice that our CDF functions of wages looks smooth, like the right panel of Figure 1.2 for a continuous random variable, rather than like a staircase as in 1.1 for a discrete random variable. But when evaluating probabilities using the empirical distribution of our dataset, `wage` can't literally be a continuous random variable: in a dataset of 28,534 there can at most be 28,534 distinct values of wage represented! In fact, with a finite number of people on Earth, only a finite number of wages would be represented even if we had an idealized dataset that included everybody.

Therefore, strictly speaking, we're plotting the CDF of a discrete random variable above. But since wages can take any real number as a value, (rather than, e.g. being confined to integers or some set of categories), the distribution of wages is well-approximated by thinking of it as being a continuous random variable. The "jumps" in the above plot are so tiny in our dataset that they are imperceptible to the naked eye.

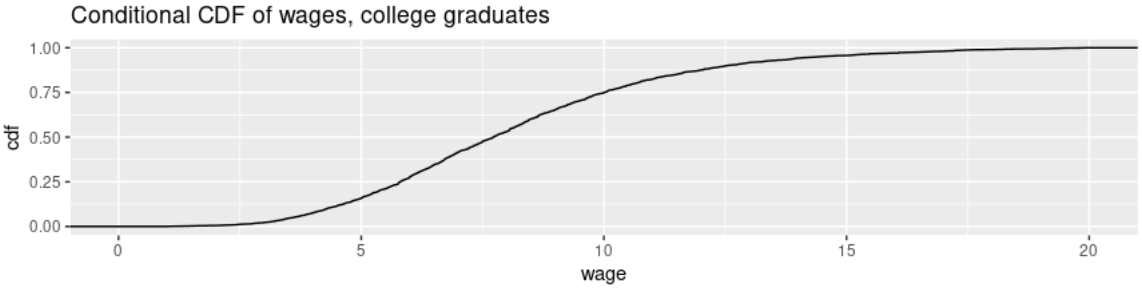
Thus, it's meaningful to talk about the conditional density of wages associated with each of the conditional CDFs above. R makes it easy to generate plots of the these conditional densities, with the following code:

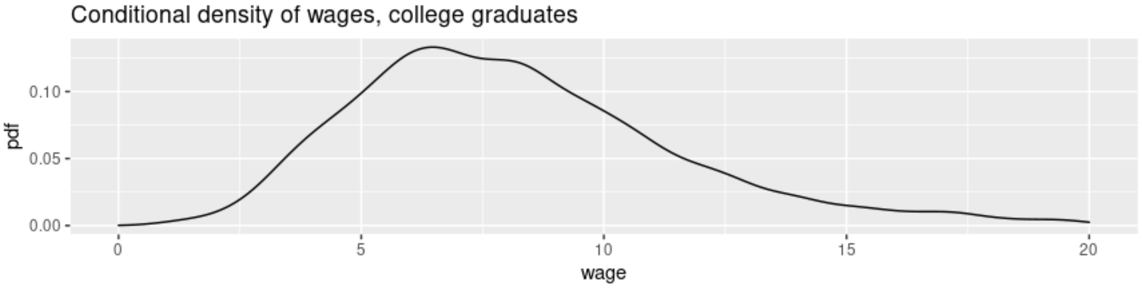
```
plot3<-ggplot(dfgrads, aes(wage)) + geom_density()+xlim(0, 20) +
    labs(title="Conditional density of wages, college graduates",y=
    "pdf")
multiplot(plot1, plot3)

plot4<-ggplot(dfnongrads, aes(wage)) + geom_density()+xlim(0, 20)
     +labs(title="Conditional density of wages, non-graduates",y="
    pdf")
multiplot(plot2, plot4)
```

The command `geom_density()` is actually doing some fairly complicated calculations in the background, which aren't important here. We just want the graphs. Here's the conditional distribution for non-graduates, represented both as a CDF and as a density function:



For college graduates we have:


Conditional CDF of wages, college graduates


Conditional density of wages, college graduates

In each case, the density is highest where the CDF is steepest, and lowest where the CDF is the flattest. This is what we should expect, since the density function is the derivative of the CDF. Notice that the p.d.f. of wages for non-graduates peaks around \$4 and hour, while for graduates it peaks around \$7 an hour.

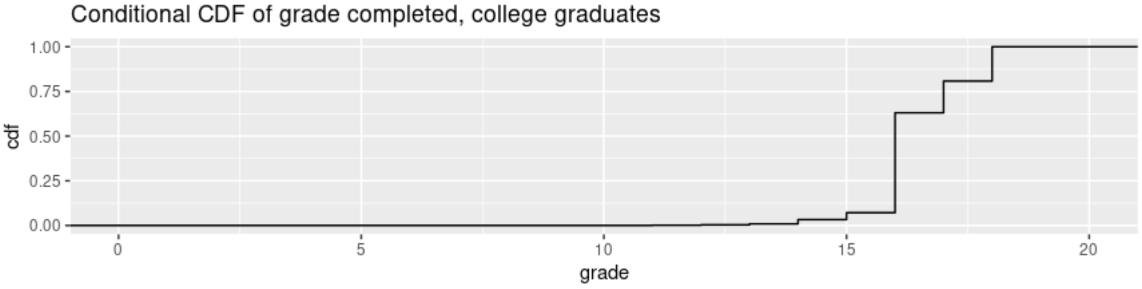
We can easily compute the expectations of these conditional distributions using a command like `mean(df[df$collgrad==1,]$wage)`. This command takes the average value of wage across all rows in which `collgrad`=1. Running this command for each value of `collgrad` yields $\mathbb{E}[wage|collgrad = 0] \approx$ \$5.52 and $\mathbb{E}[wage|collgrad = 1] \approx$ \$8.65.

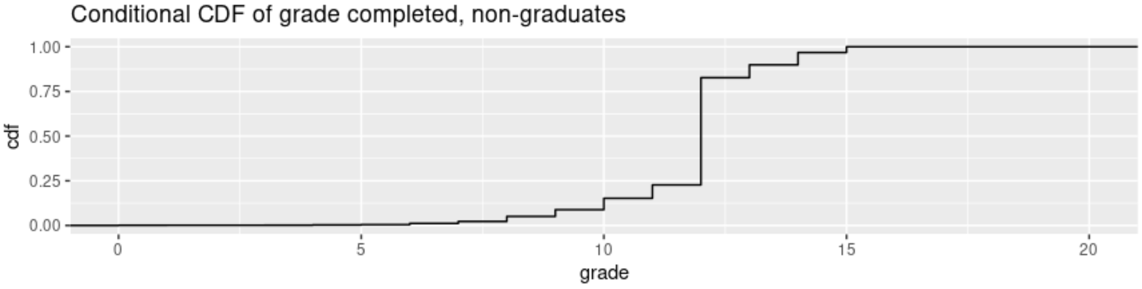
This provides us an opportuntiy to test the law of expectations, which says that

$$\begin{aligned}\mathbb{E}[wage] = \mathbb{E}\left[\mathbb{E}[wage|collgrad]\right] = P(colgrad = 0) \cdot \mathbb{E}[wage|collgrad = 0]\\ + P(colgrad = 1) \cdot \mathbb{E}[wage|collgrad = 1]\end{aligned}$$

Indeed `mean(df$wage)` yields $\mathbb{E}[wage] \approx$ \$6.04 and `mean(df$collgrad==0)` reveals that $P(collgrad = 1) \approx 0.17$, and $0.83 \cdot 5.52 + 0.17 \cdot 8.65 \approx 6.04$.

Now let's look at an example of condition distributions in which Y is discrete, rather than continuous like wage. Here are conditional CDFs of a workers highest grade of education completed, by college graduation status:

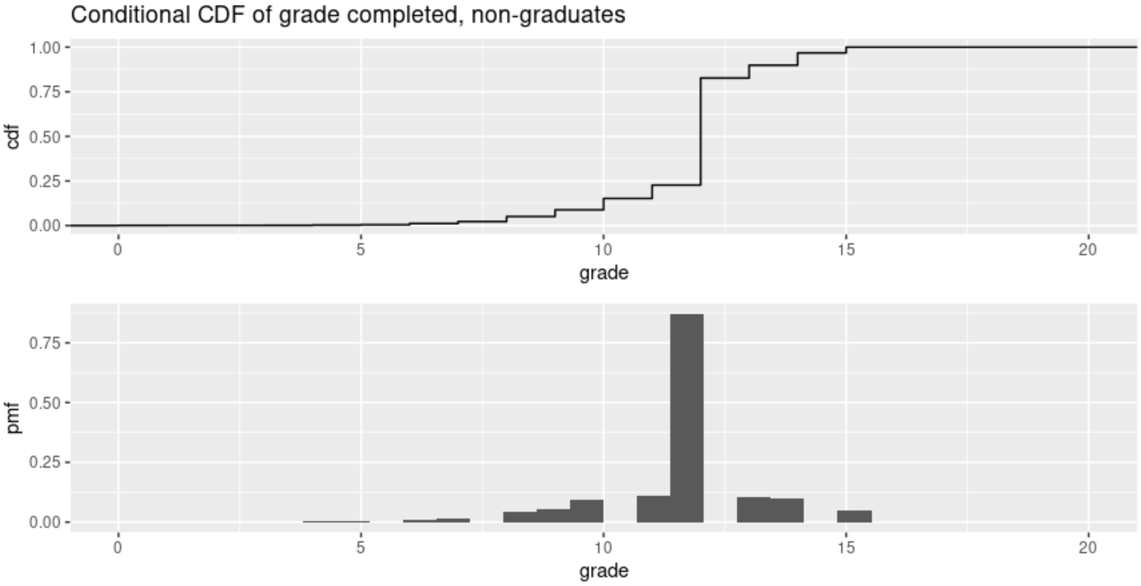

Conditional CDF of grade completed, college graduates


Conditional CDF of grade completed, non-graduates

Now the "staircase" nature of the CDF for a discrete random variable is clear. About 10% of college graduates have 15 or less years of education, yet have graduated college. Notice that the big jump in the

CDF for graduates is at grade 16 (12 years + a 4 year college degree), but there are also jumps beyond that for post-graduate degrees. By contrast, the big jump for non-graduates occurs at 12, indicating high-school completion.

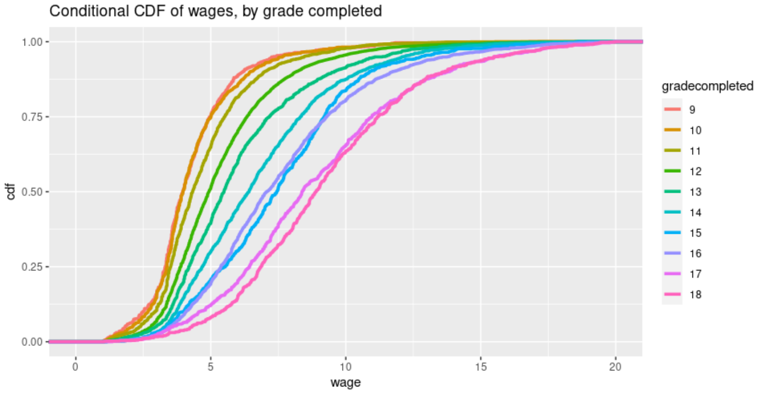
Here is the CDF for the non-graduates alongside its probability mass function, plotted as a histogram.

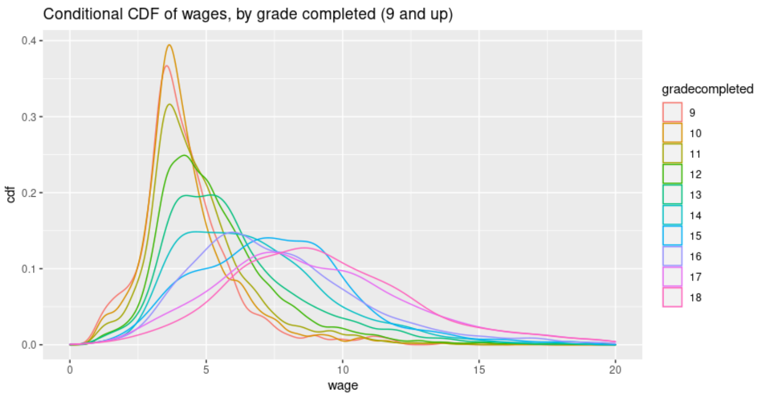


These last two figures were generated with the following code

```
plot1<-ggplot(dfgrads, aes(grade)) + stat_ecdf(geom = "step") +
   labs(title="Conditional CDF of grade completed, college
   graduates",y="cdf")+xlim(0, 20)
plot2<-ggplot(dfnongrads, aes(grade)) + stat_ecdf(geom = "step") +
    labs(title="Conditional CDF of grade completed, non-graduates"
   ,y="cdf")+xlim(0, 20)
multiplot(plot1, plot2)

plot3<-ggplot(dfnongrads, aes(grade)) + geom_histogram(aes(y=..
   density..))+xlim(0, 20) +labs(y="pmf")
multiplot(plot1, plot3)
```

So far we've visualized distributions that condition on the binary variable *collgrad*, so there were also two conditional distributions to look at. Now let's move beyond a binary conditioning variable. For example, the following figures show the conditional CDF and conditional density of wages given highest grade completed, for grades 9 and up.

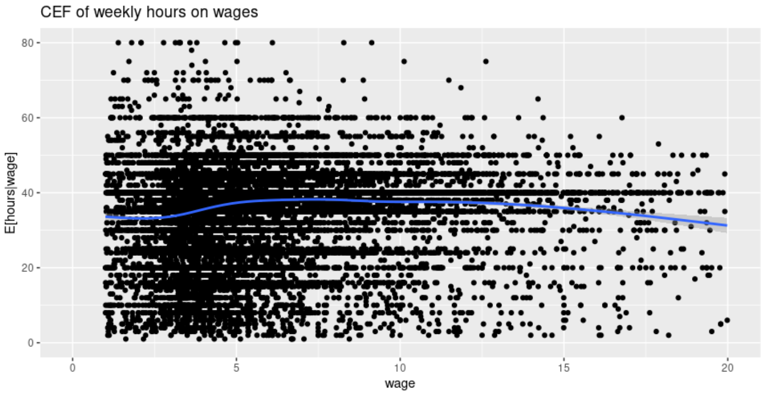




The code for these graphs is as follows (note we had to omit some rows in which `grade` is undefined using the condition `!is.na(df$grade)`).

```
df$gradecompleted = factor(df$grade)
ggplot(df[!is.na(df$grade)&df$grade>8,], aes(x = wage, color =
    gradecompleted)) + stat_ecdf(lwd = 1.25, geom = "step") + labs(
    title="Conditional CDF of wages, by grade completed",y="cdf") +
     xlim(0, 20)
ggplot(df[!is.na(df$grade)&df$grade>8,], aes(x = wage, color =
    gradecompleted)) + geom_density() + labs(title="Conditional CDF
     of wages, by grade completed (9 and up)",y="cdf") + xlim(0,
    20)
```

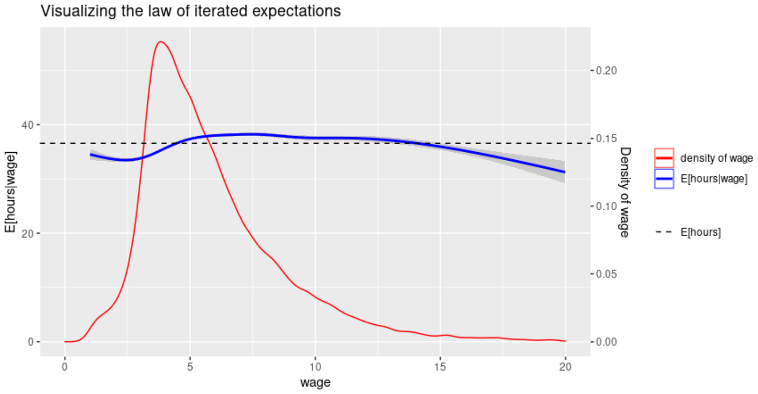
Finally, lets consider a continuous conditioning variable. We'll take Y to be usual hours worked in a week, and X to be wage. We can no longer visualize the conditional distributions $F_{Y|X=x}$ one-by-one. We can however visualize the conditional expectation function $\mathbb{E}[hours|wage = w]$ as a function of w. Below we plot it over a raw scatterplot of *hours* and *wage*, which provides a visualization of the joint distribution of the two variables.



Now lets see the law of iterated expectations in action in this setting. With wages conceived of as a continuously-distributed random variable, the law of iterated expectations says that

$$\mathbb{E}[hours] = \mathbb{E}\left[\mathbb{E}[hours|wages]\right] = \int f_{wage}(w) \cdot \mathbb{E}[hours|wages = w] \cdot dw$$

where $f_{wage}(w)$ is the density of *wages* evaluated at w. The plot below shows the function $f_{wage}(w)$ in red (scale given on the right), and the function $\mathbb{E}[hours|wages = w]$ in blue (scale given on the left). The value of $\mathbb{E}[hours]$ turns out to be about 46.56, and this is visualized by a horizontal dotted line on the same scale as the condition expectation of hours.



Observe that the dotted line cuts through the blue CEF, capturing it's "average value". This average is weighted according to the red line, the values near \$4-\$6 get the most weight, for example. It is thus sensible that the value of $\mathbb{E}[hours]$ is close to the midpoint of the function $\mathbb{E}[hours|wages = w]$ in that range.

The code used to generate these graphs and calculate $\mathbb{E}[hours]$ is:

```
#CEF of hours by wage
ggplot(df, aes(x=wage, y=hours)) + geom_point() + geom_smooth() +
   labs(title="CEF of weekly hours on wages",y="E[hours|wage]") +
   xlim(0, 20) + ylim(0, 80)
#CEF of hours by wage with density

avghours<-mean(df$hours, na.rm=TRUE)
ggplot(df) + geom_density(aes(x=wage,y=..density..*250, color="
   density of wage"))+geom_smooth(aes(x=wage, y=hours, color="E[
   hours|wage]")) + labs(title="Visualizing the law of iterated
   expectations",y="E[hours|wage]") + xlim(0, 20) + ylim(0, 80)  +
    scale_y_continuous(name = "E[hours|wage]", sec.axis = sec_axis
   ( trans=~./250, name="Density of wage")) + geom_hline(aes(
   yintercept = avghours,linetype = "E[hours]"), colour = "black")
   +theme(legend.key = element_rect(colour = NA, fill = NA),
   legend.title = element_blank())+scale_linetype_manual(name = "E
   [hours]", values = c(2), guide = guide_legend(override.aes =
   list(color = c("black"))))+scale_color_manual(name = "E[hours]"
   , values = c("red","blue"),guide = guide_legend(override.aes =
   list(fill=c("white", "white"), color = c("red", "blue"))))
```

Chapter 3

Statistical models

In Chapter 1, we've developed the idea of a random vector, which has a probability distribution that can be characterized by the joint-CDF of all of its components. This allows us to then define concepts like expectation, conditional distributions, and the conditional expectation function. Chapter 2 has shown an empirical illustration in these concepts.

The data used in Chapter 2 is an example of a *sample*. In this case, it was a collection of observations regarding $\sim$5000 young working women in the 1970s and 1980s. This chapter develops tools that let us address the following question: what are we learning about the *population* of young working women in this time period, given our sample? In doing so we move from the theory of *probability* to the theory of *statistics*, which studies what we can learn about probability distributions from data.

To do so, it is useful to start with the definition of a statistical *model*, which embodies a set of assumptions about the distribution of a set of random variables. In Section 3.3, we'll apply this idea to model how a sample of data is generated from an underlying population.

3.1 Modeling the distribution of a random vector

Let X be a random vector (let's say it has k components), having some joint-CDF function F. We'll define a *model* to be a set of such distributions, which we think that the true F might belong to.

Definition 3.1. *A statistical model is a set $\mathcal{F}$ of potential candidates for F.*

One thing we know for sure about the CDF function $F(\mathbf{x}) = P(X_{1i} \le x_1, X_{2i} \le x_2, \dots X_{ki} \le x_k)$ is that it is increasing and right-continuous with respect to each x_j and takes values between zero and one. Let $\mathfrak{F}$ be the set of all such functions, which reflect valid CDF functions for a $k-$dimensional random vector X. A statistical model thus specifies a *subset* $\mathcal{F} \subseteq \mathfrak{F}$.

Most familiar models in statistics are "parametric" models. A *parametric* statistical model is a model in which each $F \in \mathcal{F}$ can be written in terms of a vector of parameters $\theta \in \mathbb{R}^d$. What I mean by this is that each function $F(\mathbf{x})$ in $\mathfrak{F}$ can be written as $F(\mathbf{x};\theta)$: the whole *function* $F(\cdot;\theta)$ depends on the values of the parameters $\theta = (\theta_1, \theta_2, \dots \theta_d)'$:

Definition 3.2. *A parametric statistical model is the set $\mathcal{F} = \{F(\cdot;\theta) : \theta \in \Theta\}$, where each $\theta \in \Theta$ is some finite-dimensional vector (i.e. Θ is some subset of $\mathbb{R}^d$ for a finite d).*

Example: Suppose somebody hands you a coin, which could be "weighted" so that the probability of heads is different than 1/2. They don't tell you the value of $p = P(heads)$. Thus, your statistical model is that $P(heads) = p$ and $P(tails) = 1 - p$ for some $p \in [0, 1]$. This model has one parameter, p.

Notation: When a parametric model contains only continuous distributions, the distributions in $\mathcal{F}$ are often denoted by their densities $f(\cdot;\theta)$ rather than CDFs. You also sometimes see the notation $f(\cdot|\theta)$, with a "|" rather than ";".

Preview: parametric vs. non-parametric models:

Note that a parametric statistical model must be characterized by a finite number of real-valued parameters, i.e. d is finite. This distinguishes parametric models from *non-parametric* models $\mathcal{F}$, in which there's no way to come up with a finite number of parameters that can fully characterize each $F \in \mathcal{F}$.

Example: Suppose that X is a random variable, and we let $\mathcal{F}$ contain all valid CDFs such that the function $F(x)$ is concave in x (on its support, where it's increasing). This represents a *non-parametric* statistical model.

Non-parametric models play an important role in modern econometrics, as they often make weaker assumptions than parametric models, and become most practical to work with when we have big datasets. "Semi-parametric" models occur when a finite-dimensional θ pins down some—but not all—features of each $F \in \mathcal{F}$. Semi-parametric models play an important role in regression analysis, as we'll see.

We'll start by studying some important parametric models for a single random vector X. We'll then apply the idea of a model for the distribution of X to move on to our real objective: modeling the generation of a whole *dataset*, which typically consist of n realizations of a random vector X.

3.2 Two examples of parametric distributions

3.2.1 The normal distribution

Arguably the most important parametric model arises from the family of probability distributions called *normal distributions.*

The univariate normal distribution is a continuous distribution with density function:

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{3.1}$$

for some values σ and μ. We say that $X \sim N(\mu, \sigma^2)$ when it has a density function given by $f(\cdot; \mu, \sigma)$.

Exercise: Show that the normal density $f(x; \mu, \sigma)$ integrates to one. *Hint:* use $\int_{-\infty}^{\infty} e^{-t^2} dt = \sqrt{\pi}$.

The following exercises ask you to show that the parameters μ and σ are the expectation and standard deviation of X, respectively (the *standard deviation* of a random variable is defined as the square root of its variance).

Exercise: Show that if $X \sim N(\mu, \sigma^2)$ then $\mathbb{E}[X] = \mu$.

Exercise: Show that if $X \sim N(\mu, \sigma^2)$ then $Var(X) = \sigma^2$. *Hint:* use $\int_{-\infty}^{\infty} t^2 \cdot e^{-t^2} dt = \frac{\sqrt{\pi}}{2}$

Proposition 3.1. *Suppose that $X \sim N(\mu, \sigma^2)$ and we define $Y = a + bX$ for some $a, b \in \mathbb{R}$. Then Y is also normally distributed as $N(a + b\mu, b^2\sigma^2)$.*

A consequence of this is that if $X \sim N(\mu, \sigma)$, then the random variable $\left(\frac{x-\mu}{\sigma}\right) \sim N(0, 1)$. $N(0, 1)$ is called the *standard normal* distribution.

The multi-variate normal distribution generalizes the normal distribution to a random vector $X = (X_1, X_2, \dots X_k)$. The multi-variate normal density is parametrized by a a $k \times 1$ vector μ and $k \times k$ matrix $\mathbf{\Sigma}$:

$$f(\mathbf{x}; \mu, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^k \cdot det(\mathbf{\Sigma})}} e^{-\frac{1}{2}(\mathbf{x}-\mu)\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)'} \tag{3.2}$$

where $\mathbf{\Sigma}^{-1}$ denotes the matrix inverse of $\mathbf{\Sigma}$, and $det(\mathbf{\Sigma})$ its *determinant.* When $X \sim N(\mu, \mathbf{\Sigma})$, the mean of X is μ, i.e. $\mathbb{E}[X] = \mu$, and $\mathbf{\Sigma}$ is its variance-covariance matrix: $Var(X) = \mathbf{\Sigma}$.

Example: The bivariate normal distribution, the case when $k = 2$ warrants some additional attention. Let ρ be the correlation coefficient between X_1 and X_2, so that

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho \cdot \sigma_1 \cdot \sigma_2 \\ \rho \cdot \sigma_1 \cdot \sigma_2 & \sigma_2^2 \end{pmatrix} = (\sigma_1, \sigma_2) \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} (\sigma_1, \sigma_2)'$$

where $\sigma_j = \sqrt{Var(X_j)}$. In this case Eq. (3.2) simplifies to:

$$f(x_1, x_2; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right\}} \tag{3.3}$$

Exercise: Show from Eq. (3.3) that if X_1 and X_2 are jointly normally distributed, then $X_1 \perp X_2$ if and only if $Cov(X_1, X_2) = 0$.

A useful property of the family of normal distributions is that many operations keep one in the family. For example, when X is a (multivariate) normal random vector, the marginal distribution of each of it's components is a (univariate) normal random variable:

Proposition 3.2. *Let $X \sim N(\mu, \mathbf{\Sigma})$. Then the marginal distribution of each X_j is $N(\mu_j, \Sigma_{jj})$.*

A vector $\tilde{X}$ composed of any subset of the X_j is also normal.

Conditional distributions defined from a multivariate normal are also normal:

Proposition 3.3. *Let (X, Y) follow a bivariate normal distribution with parameters $(\mu_Y, \mu_X, \sigma_Y, \sigma_X, \rho)$. Then the distribution of Y given $X = x$ is a normal distribution with mean $\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$ and variance $\sigma_Y^2(1-\rho^2)$.*

Proof. The conditional density of Y given $X = x$ is

$$\begin{aligned} f_{Y|X=x}(y) = \frac{f_{XY}(x,y)}{f_X(x)} &= \frac{\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)\right\}}}{\frac{1}{\sigma_X\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2}} \\ &= \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \rho^2\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right\}} \end{aligned} \tag{3.4}$$

where we've used that $e^a/e^b = e^{a-b}$ and that $\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2 = \frac{1-\rho^2}{2(1-\rho^2)}\left(\frac{x-\mu_X}{\sigma_X}\right)^2$. Now the beautiful part: the exponent in (3.4) is a "perfect square" quantity:

$$f_{Y|X=x}(y) = \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}\left(\frac{y-\mu_Y}{\sigma_Y} - \rho\left(\frac{x-\mu_X}{\sigma_X}\right)\right)^2} = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_Y^2(1-\rho^2)}}e^{-\frac{1}{2}\left(\frac{y-\left(\mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right)}{\sigma_Y^2(1-\rho^2)}\right)^2}$$

which is exactly the formula for the density of a $N(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X), \sigma_Y^2(1-\rho^2))$ random variable. □

Later in the course, we'll see that Proposition 3.3 generalizes in a natural way beyond the bivariate case.

Sums of normal random variables are also normally distributed, provided that the two random variables being added are *jointly* normal:

Proposition 3.4. *If X and Y are jointly normal with $\begin{pmatrix} X \\ Y \end{pmatrix} = N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}\right)$*

$$\textit{Then:} \quad X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY})$$

This is helpful in establishing the following, which generalizes Proposition 3.1:

Proposition 3.5. *Let $X \sim N(\mu, \mathbf{\Sigma})$. Then for any $\mathbf{a} = (a_1, a_2, \ldots a_k)'$:*

$$\mathbf{a}'X := \sum_{j=1}^{k} a_j \cdot X_j \sim N(\mathbf{a}'\mu, \mathbf{a}'\mathbf{\Sigma}\mathbf{a}))$$

Proof. Using the fact that for X and Y normal, $X+Y$ is normal $k-1$ times establishes that $\mathbf{a}'X$ is normal (making use of the fact that a_jX_j is normal for each j). If we know that $\mathbf{a}'X$ is normal, and we know it's mean vector and variance-covariance matrix, we know its whole distribution. That $\mathbb{E}[\mathbf{a}'X] = \mathbf{a}'\mathbb{E}[X] = \mathbf{a}'\mu$ follows from linearity of the expectation. That $Var(\mathbf{a}'X) = \mathbf{a}'Var(X)\mathbf{a}$ then follows, also by linearity of expectation. □

3.2.2 The binomial distribution*

Suppose we have n random variables $Z_1 \dots Z_n$, where each Z_j is 0/1 random variable (referred to as a *Bernoulli* random variable). For each Z_j: $P(Z_j = 1) = p$ and $P(Z_j = 0) = 1-p$. We can construct a random vector $Z = (Z_1, Z_2, \dots Z_n)'$ from these n random variables.

Suppose that each of the Z_j are independent, meaning that:

$$F(\mathbf{z}) = F_{single}(z_1) \times F_{single}(z_2) \times \dots \times F_{single}(z_n)$$

where $F_{single}(z)$ is the CDF of a single Bernoulli random variable having probability p: $F_{single}(z) = (1-p)\cdot \mathbb{1}(z<1) + p \cdot \mathbb{1}(z \geq 1)$.

We now define the *binomial distribution* to be the distribution of $X := \frac{1}{n}\sum_{j=1}^n Z_j$. Often, each Z_j is interpreted as the success (1) or failure of a "trial" of some kind. The probability of success of trial j is $P(Z_j = 1) = p$. With this interpretation the random variable X simply counts the number of successes out of the n trials.

To denote that X has the binomial distribution with parameters p and n, we write $X \sim B(n,p)$. Since X can only take integer values between 0 and n, the binomial distribution is a discrete random variable. Rather than a density, it has a probability mass function. The probability that $X = k$ can be written as:

$$P(k;n,p) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

where $\binom{n}{k} := \frac{n!}{k!(n-k)!}$. The quantity $\binom{n}{k}$ simply counts the number of distinct ways that we could get k sucesses from n trials, i.e. the number of distinct vectors $\mathbf{z} \in \{0,1\}^n$ that contain k ones and $n-k$ zeroes. Each such $\mathbf{z}$ has the same probability $p^k \cdot (1-p)^{n-k}$ of occurring, leading to our final expression for $P(k;n,p)$.

We know by linearity of the expectation that the mean of X must be $\mathbb{E}[X] = \mathbb{E}\left[\sum_{j=1}^n Z_j\right] = \sum_{j=1}^n \mathbb{E}[Z_j] = n\cdot p$. We can also verify this by the explicit formula for $P(k;n,p)$:

$$\mathbb{E}[X] = \sum_{k=0}^n k \cdot \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = \sum_{k=0}^n \frac{n! \cdot k}{k!(n-k)!} p^k \cdot (1-p)^{n-k}$$

This is an intimidating sum to try to evaluate. But we can use the following "trick". Since the p.m.f. of X must sum to one, we know that $\sum_{k=1}^n P(k;n,p) = 1$ for any n and p. Notice that $\frac{k}{k!} = \frac{1}{(k-1)!}$. Thus, the factor of k that appears when we take the expectation of X has a similar effect to simply decreasing k by one, and summing over its *p.m.f.* instead. But we also have to deal with the other factors that depend on k. Note that

$$P(k;n,p) = \frac{n!}{(n-1)!} \cdot \frac{(k-1)!}{k!} \cdot p \cdot P(k-1;n-1,p)$$

if $k>0$. Thus: $\mathbb{E}[X] = \sum_{k=1}^n k \cdot P(k;n,p) = \sum_{k=1}^n np \cdot P(k-1;n-1,p) = np \cdot \sum_{k=1}^n P(k-1;n-1,p) = np \cdot \sum_{k=0}^{n-1} P(k;n-1,p) = np \cdot 1$ where in the first step we've used that $k \cdot P(k;n,p) = 0$ if $k = 0$.

Exercise: Show that if $X \sim B(n,p)$, $Var(X) = n \cdot p(1-p)$. You may find it useful that since $Z_j \perp Z_{j'}$ for each pair $j \neq j'$, $Cov(Z_j, Z_{j'}) = 0$.

3.3 Random sampling

The last section covered two parametric families of distributions, either of which could be used to form a parametric statistical model. However, the most common kind of statistical model used in practice is a non-parametric one: an *independent and identically distributed* (i.i.d.) sample.

Definition 3.3. *A collection of random vectors $\{X_1, X_2, \dots X_n\}$ are called independent and identically distributed (i.i.d.) if $X_i \perp X_j$ for $i \neq j$ and each X_i has the same marginal distribution as the others.*

When a collection of random vectors are independent of one another, as with an i.i.d. collection knowing the CDF F for each member X_i of the collection is sufficient to recover the full joint-distribution of the collection. For example, let $n = 2$ and suppose both X_1 and X_2 are *i.i.d* random variables (rather than vectors). Then $P(X_1 \leq x_1, X_2 \leq x_2) = P(X_1 \leq x_1) \cdot P(X_2 \leq x_2) = F(x_1) \cdot F(x_2)$, where $F(\cdot)$ is the marginal CDF function of each of the X_i. With an i.i.d. collection, we only need to know the CDF F that applies to each of the X_i, in order to know anything about the collection.

Note that assuming that a collection of random vectors is *i.i.d.* is a statistical model, in the sense of Section 3.1. In that section we weren't talking about collections of random vectors, but you can always think of a collection of random vectors as an even larger list of random variables. For example, in the example above Definition 3.3 restricts the joint-CDF of X_1 and X_2 to have the form $F_{12}(x_1, x_2) = F(x_1) \cdot F(x_2)$ for some valid CDF function F.

The *i.i.d.* model is typically used to describe *simple random sampling.* Simple random sampling occurs when individuals are selected at random from some underlying population I, and a set of variables $X_i = (X_{1i}, X_{2i}, \dots X_{ki})'$ are recorded for each sampled individual i. Imagine for example a telephone survey, in which enumerators have a long list I of potential individuals to contact. They use a random number generator to choose an i at random from this list, contact them, and record responses to a set of k questions. This process is then repeated n times.

Note: With a finite population I, we must allow sampling "with replacement" for the i.i.d. model to hold strictly. If individual i is removed from the list after being contacted, then the random vectors X_i may no longer be independent. For example, suppose we are randomly selecting U.S. states and recording the population of each one. Suppose California (the post populous state) has 40 million and Georgia has 11. Then for example $P(X_2 = 40m | X_1 = 40m) \neq P(X_2 = 40m | X_1 < 40m)$, since the first probability is zero and the second is $1/49$. This means that X_1 and X_2 are not independent. Simple random sampling is often referred to as *random sampling* for short, or as *i.i.d sampling.*

We'll use the terms *dataset* or *sample* to refer to an $n \times k$ matrix $\mathbf{X}$ that records characteristics $X_i = (X_{1i}, X_{2i}, \dots X_{ki})$ for each of n observational units (such as individuals) i. Data is not always generated by simple random sampling, but when it is, we can imagine $\mathbf{X}$ as being formed by randomly choosing rows from a much larger matrix that records X_i for all individuals in the population, depicted in Figure 3.1. The actual data we see in $\mathbf{X}$ is a realization of the collection of random variables $\{X_1, X_2, \dots X_n\}$.

$$\mathbf{X} = \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix} = \begin{pmatrix} (X_{11}, X_{21}, \dots X_{k1}) \\ (X_{12}, X_{22}, \dots X_{k2}) \\ \vdots \\ (X_{1n}, X_{2n}, \dots X_{kn}) \end{pmatrix}$$

The randomness of $\mathbf{X}$ comes from the random-sampling: we could have drawn a different set of individuals from the population, in which case we would have seen a different dataset $\mathbf{X}$.

Notation: Note that the entries of the sample matrix $\mathbf{X}$ are denoted X_{ji}, where i index rows (individual observations) and j index columns (variables/characteristics). This is backwards from the way we often denote entries M_{ij} of a matrix $\mathbf{M}$, where the row i comes before the column j. This is a consequence of two conventions interacting: that rows of $\mathbf{X}$ index individuals (just like when you open the dataset in `R`), but that X_{ji} indexes characteristic j of individual i (equivalently, characteristic j of the individual sampled in row i).

Note that most sampling processes in the real world occur without replacement: the same individual cannot show up twice in the data. Given the note above, this suggests that these sampling processes are not *i.i.d.*, strictly speaking. However, when the size N of the underlying population is large, such samples can still be well-approximated as being *i.i.d.*. Intuitively, that's because when N is much larger than n (often denoted as $N >> n$), the chance that you would draw the same individual twice is very low. We thus typically assume *i.i.d.*, with the idea that N is suitably large to not worry about sampling with vs. without replacement.

Sample $\mathbf{X}$

row **i**	$\omega_{\mathbf{i}}$	**age**$_i$	**married**$_i$	**college**$_i$
1	1	25	0	0
2	4	37	1	1
3	5	54	0	1

Population I

individual **i**	**age**$_i$	**married**$_i$	**college**$_i$
1	25	0	0
2	74	1	1
3	8	0	0
4	37	1	1
5	54	0	1

Figure 3.1: An example of simple random sampling, in which $n = 3$ and $N = 5$. Each row of the dataset on the left is a realization of random vector $X = (age, married, college)$, which chooses a row at random from the population matrix on the right. We can conceptualize this sampling process as a probability space with outcomes $\omega = (\omega_1, \omega_2, \omega_3)$, where ω_i yields the index of the randomly selected individual in I. The random vectors $X_i = X_i(\omega_i)$ and $X_j = X_j(\omega_j)$ are independent for $i \neq j$, but the random variables within a row are generally not independent, e.g. age_i and $college_i$ are positively correlated.

The following are some alternative methods of generating data, aside from simple random sampling:

- *Stratified random sampling*: the population is divided into groups, and then simple random sampling occurs within each group (e.g. I run my sampling algorithm separately for men and women, so that I can ensure equal representation of each).
- *Clustered random sampling*: after defining groups, we randomly select some of the groups. Then all individuals from those groups are included in the sample (e.g. I interview everybody in a household, after choosing households at random)
- *Panel data*: suppose we have observations over multiple time-periods t for each individual i, where the individuals i are drawn as a simple random sample. Then if we arrange all of i's data onto one row, we can imagine $\mathbf{X}$ as reflecting an *i.i.d.* sample. But with rows correponding to (i, t) pairs, the rows are no longer independent (in general)
- *Observing the whole popluation*: this would be the case e.g. with state-level data from all 50 U.S. states. This situation occurs increasingly frequently with individual-level data now as well, e.g. administrative data on all tax-filers in a country.

These alternative sampling methods tend to violate the *i.i.d* assumption. However, methods exist to deal with each of them.

Let us end this section with a last bit of jargon. When X_i for $i = 1 \dots n$ denotes a collection of *i.i.d* random vectors, we'll refer to the distribution F that describes the marginal distribution of each X_i as the *population distribution*. The population distribution is the distribution we get when we randomly select any individual from the population. Features of the population distribution are the ones that you naturally think of when you think about summarizing a population. For example, if I is a finite population, then

$$\mathbb{E}_F[X_i] = \frac{1}{N} \sum_{i \in I} X_i$$

where we use the notation $\mathbb{E}_F$ to make explicit that the expectation is with respect to the CDF F. The population mean is simply the mean of X_i among everybody in I. We can also talk about the population variance, the population median, and so on. Note that the *empirical distribution* introduced in Section 2.2 is an example of a population distribution in which we think of the whole population as equal to the rows of the dataset.

Another piece of terminology will be useful as we discuss samples and their population counterparts:

Definition 3.4. *A* ***statistic*** *or* ***estimator*** *is any function of the sample* $\mathbf{X} = (X_1', X_2', \dots, X_n')'$.

A generic estimator or statistic will apply some function $g(\mathbf{X}) = g(X_1, X_2, \dots X_n)$ to the collection of random vectors that constitute the sample. An example is the so-called *sample mean* $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$, which simply adds together X_i for across the sample and divides by the number of observations n. $\bar{X}_n$ is an example of a statistic. Since each of the X_i is a random variable/vector, it follows that $\bar{X}_n$ is itself a random variable/vector. This is true of statistics in general: they are random.

The reason that we also refer to statistics as "estimators" is that statistics often attempt to estimate a population quantity of some kind from data. For example, we'll see in the next Chapter that for large n, we are justified in thinking that $\bar{X}_n \approx \mu$. It is therefore reasonable to use $\bar{X}_n$ as an estimate of μ. Note that $\bar{X}_n$ is random, while μ is just a fixed number. Thus we have to be careful in what we mean by saying that $\bar{X}_n \approx \mu$, which is the topic of the next chapter.

Notation: Often estimators are depicted with a "hat" on them, e.g. $\hat{\theta} = g(\mathbf{X})$. We'll use this notation to denote a generic estimator.

A useful property of *i.i.d.* random vectors that I'll mention here is the following:

Proposition 3.6. *If $\{X_1, X_2, \dots X_n\}$ are i.i.d random vectors, then $\{h(X_1), h(X_2), \dots h(X_n)\}$ are also i.i.d for any (measurable) function h.*

An implication of Proposition 3.6 is that if we have an *i.i.d.* sample X_i, we can from it construct an *i.i.d.* sample of e.g. X_i^2.

Chapter 4

When the sample gets big: asymptotic theory

4.1 Introduction: the law of large numbers

Consider an *i.i.d.* sample $\{X_1, \dots X_n\}$ of some random variable X_i. The *sample average* of X_i in our data simply takes the arithmetic mean across these n observations:

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

The law of large numbers (LLN) states the deep and useful fact that for very large n, it becomes very unlikely that $\bar{X}_n$ is very far from $\mu = \mathbf{E}[X_i]$, the "population mean" of X_i.

Theorem 1 (law of large numbers). *If X_i are i.i.d random variables and $E[X_i]$ is finite, then for any $\epsilon > 0$:*

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$

Note: The LLN is stated above for a random variable, but the result generalizes easily to random vectors. In that case, $\lim_{n\to\infty} P(||\bar{\mathbf{X}}_n - \mu||_2 > \epsilon) = 0$ where $||\cdot||_2$ denotes the Euclidean norm, i.e.: $|\bar{\mathbf{X}}_n - \mu| = (|\bar{\mathbf{X}}_n - \mu|)'(|\bar{\mathbf{X}}_n - \mu|)$, where $\bar{\mathbf{X}}_n$ is a vector of sample means for each component of X_i, and similarly for μ.

Note: the version of the law of large numbers above is called the *weak* law of large numbers. There exists another version called the strong LLN.

Let us now prove the LLN. We will do so using a tool called *Chebyshev's inequality.* This proof assumes that $Var(X_i)$ is finite, but the LLN holds even if $Var(X_i) = \infty$. Chebyshev's inequality allows us to use the variance of a random variable to put an upper bound on the probability that the random variable is far from its mean. In particular, for any random variable Z with finite mean and variance:

$$P(|Z - \mathbb{E}[Z]| \geq \epsilon) \leq \frac{Var(Z)}{\epsilon^2}$$

To see that this holds, use the law of iterated expectations to write out the variance as

$$\begin{aligned} Var(Z) = \mathbb{E}\left[Z - \mathbb{E}[Z])^2\right] &= P(|Z - \mathbb{E}[Z]| \geq \epsilon) \cdot \mathbb{E}\left[(Z - \mathbb{E}[Z])^2 | (Z - \mathbb{E}[Z])^2 \geq \epsilon^2\right] \\ &\quad + P(|Z - \mathbb{E}[Z]| < \epsilon) \cdot \mathbb{E}\left[(Z - \mathbb{E}[Z])^2 | (Z - \mathbb{E}[Z])^2 < \epsilon^2\right] \\ \geq P(|Z - \mathbb{E}[Z]| \geq \epsilon) \cdot \epsilon^2 + P(|Z - \mathbb{E}[Z]| < \epsilon) \cdot 0, \end{aligned}$$

noting that $|Z - \mathbb{E}[Z]| \geq \epsilon$ iff $(Z - \mathbb{E}[Z])^2 \geq \epsilon^2$.

Now, we will show that as $n \to \infty$, $Var(\bar{X}_n) \to 0$. This along with Chebyshev's inequality implies the LLN, by letting $Z = \bar{X}_n$.

To see that $Var(\bar{X}_n) \overset{n}{\to} 0$, note first that

$$\mathbb{E}[\bar{X}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

The first equality is simply the definition of $\bar{X}_n$, while the second uses linearity of the expectation operator. Now consider

$$\begin{aligned}
Var(\bar{X}_n) &= \mathbb{E}\left[(\bar{X}_n - \mathbb{E}[\bar{X}_n])^2\right] = \mathbb{E}\left[(\bar{X}_n - \mu)^2\right] = \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\right)^2\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - \mu)(X_j - \mu)\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[(X_i - \mu)(X_j - \mu)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[(X_i - \mu)\right]\left[(X_i - \mu)\right] = \frac{1}{n^2}\cdot nVar(X_i) = \frac{Var(X_i)}{n}
\end{aligned}$$

where the first equality in the third line follows because when $i \neq j$, $X_i \perp X_j$ implies that $\mathbb{E}[(X_i - \mu)] \cdot \mathbb{E}[(X_j - \mu)] = 0 \cdot 0$. Thus, the only terms that remain are when $j = i$.

Another way to see that $Var(\bar{X}_n) = \frac{Var(X_i)}{n}$ is to notice that when Y and Z are independent, $Var(Y + Z) = Var(Y) + Var(Z)$. Thus:

$$Var\left(\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n +\right) = n \cdot Var\left(\frac{1}{n}X_i\right) = n \cdot \frac{1}{n^2} \cdot Var(X_i) = \frac{Var(X_i)}{n}$$

4.2 Asymptotic sequences

The law of large numbers provides a way to justify the claim that when n is large, $\bar{X}_n$ will be close to μ with high probability. The approximation $\bar{X}_n \approx \mu$ lies at the heart of our claims to be learning about an underlying population when we have a large sample.

In the next section, we'll see that there is more than one way to develop a large-n approximation to the distribution of a random variable. To talk about such approximations, it is useful to introduce the idea of a *sequence* of random variables Z_n, where $n = 1, 2, \ldots \infty$. For example, we can consider the sample mean $\bar{X}_n$—which is a random variable for any given n—across various possible sample sizes n.

4.2.1 The general problem

The primary motivation for considering such *asymptotic sequences* of random variables Z_n is when Z_n represents a statistic $\hat{\theta}$—something that depends upon my data (see Definition 3.4). Since $\hat{\theta}$ is random (it depends on the sample that I drew), I'd like to know something about its distribution. For example, how likely is it that my sample mean is far from the population mean?

Definition 4.1. *The **sampling distribution** of an statistic $\hat{\theta}$ is its CDF: $F_{\hat{\theta}}(t) = P(\hat{\theta} \leq t)$.*

When our statistic is computed as $\hat{\theta} = g(X_1, X_2, \ldots X_n)$ from an *i.i.d* sample of X_i, $F_{\hat{\theta}}$ depends upon three things: the function g, the population distribution of X_i, and the sample size n.

Knowing the sampling distribution of a statistic is typically a hard problem. We know g and n, but in a research setting we don't generally know the CDF F that describes the underlying population. However, if we view $\hat{\theta}$ as a point along a sequence of random variables Z_n, it is often possible to say something about the limiting behavior of F_{Z_n} as $n \to \infty$. *Asymptotic theory* is a set of tools for describing this limiting behavior. The law of large numbers is one such tool. If we believe that are actual sample size n is large enough that $F_{Z_n} \approx F_{Z_\infty}$, then tools like the LLN can be extremely useful. For the sample mean for example, we might, on the basis of the LLN, be prepared to believe that $\bar{X}_n$ is close to μ with very high probability.

Conceptually, we can think of what we're doing as follows. Suppose our sample size is $n = 10,576$, and we calculate a statistic $\hat{\theta} = g(X_1, X_2, \ldots X_{10,576})$ from our sample. Now imagine applying the same function g to various samples of size $1, 2, \ldots$ and so on, and defining a sequence Z_1, Z_2, Z_n of the corresponding values. Each Z along this sequence is itself a random variable: let $F_{Z_1}, F_{Z_2}, \ldots$ be their corresponding CDFs. Our statistic $\hat{\theta}$ can be seen as a specific point along this sequence: $\hat{\theta} = Z_{10,576}$ (circled in red in Figure 4.1). Since we don't know $F_{Z_{10,576}}$, but we can say something about F_{Z_∞}, we use the latter as an approximation for the former. Figure 4.1 depicts this logic.

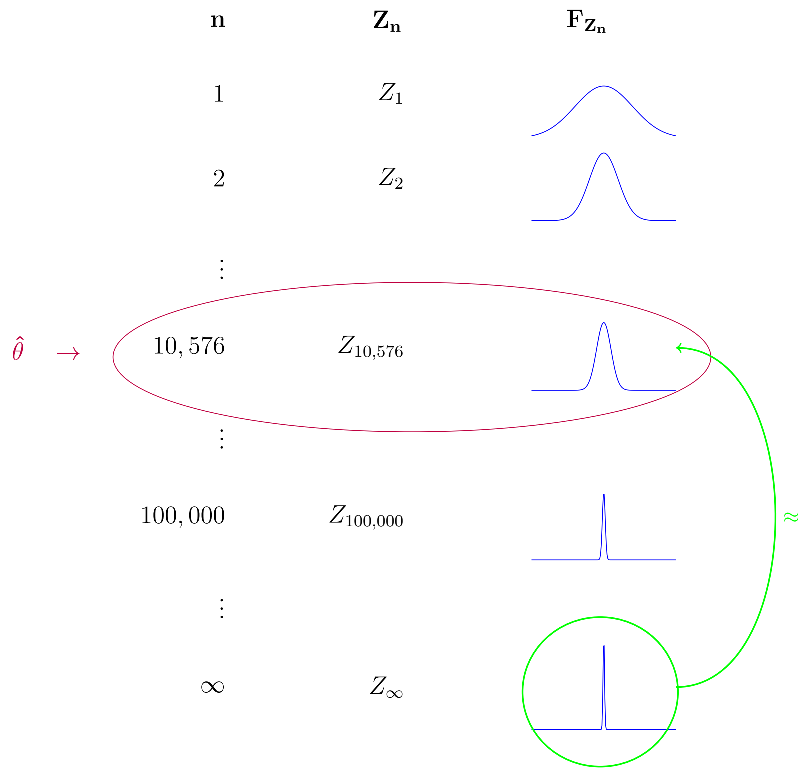


Figure 4.1: We are interested in the sampling distribution of some statistic $\hat{\theta}$, computed on our sample of $10,576$ observations. This is in general hard to compute. As a tool, we imagine a sequence of random variables $Z_1, Z_2, \ldots$ in which $\hat{\theta} = Z_{10,576}$. Asymptotic theory allows us to derive properties of F_{Z_∞}, the limiting distribution of Z_n as $n \to \infty$ (circled in green). Then we use F_{Z_∞} as an approximation to $F_{Z_{10,576}}$, which we justify by n being "large". The above figure depicts a situation in which $Z_n = \bar{X}_n$, so that the distribution of Z_n narrows to a point as $n \to \infty$ (by the LLN).

Of course, the above technique only works if we can say something definite about F_∞. The law of large numbers says that we can when our statistic is the sample mean. In Section 4.4, we'll see that the *central limit theorem* provides even more information about the limiting distribution of the sample mean: that it will become approximately normal, regardless of F.

Note: The logic of Figure 4.1 is the "classical" approach to approximating the sampling distribution of $\hat{\theta}$, but it is certainly not the only one. An increasingly popular alternative involves *bootstrap* methods. These methods still appeal to n being "large enough", but they do so in a different way. They also require computing power, because bootstrapping involves resampling new datasets from our original dataset $\mathbf{X}$. This has become increasingly feasible, and boostrap-based methods have become increasingly popular.

4.2.2 Example: LLN and the sample mean

Let's go through the logic of Figure 4.1 in more detail in the case of the the law of large numbers. The LLN tells us that when we let the sample mean $\bar{X}_n$ define our asymptotic sequence Z_n, the resulting distributions F_{Z_n} eventually cluster all of their probability mass around the point μ, the sample mean. Figure 4.2 illustrates this point, through a simulation in R. I drew 1,000 *i.i.d* samples of size n of a random variable X_i for which $P(X_i = 0) = 1/2$ and $P(X_i = 1) = 1/2$, representing a coin flip. Then, I plot a histogram of $\bar{X}_n$ across the 1,000 samples. This process is repeated for $n = 2$, $n = 10$, $n = 100$

and $n = 1,000$. You can think of this as illustrating Figure 4.1 for the specific population distribution F that describes a coin-flip. With $n = 2$, we see that we have a 50% chance of getting $\bar{X}_n$ of 0.5, which is the true "population mean" of X_i: $\mu = \mathbb{E}[X_i] = 0.5$. Then 25% of the time we get $\bar{X}_n = 0$ (two flips of tails), and 25% of the time we get $\bar{X}_n = 1$ (two flips of heads). Thus, the distribution of $\bar{X}_n$ is not very well concentrated around $\mu = 0.5$.

The red vertical lines in Figure 4.2 illustrate the law of large numbers in action. They mark the points 0.45 and 0.55, which represent a $\epsilon = .05$ in Theorem 1. We can see that by the time $n = 100$, $P(|\bar{X}_n - 1/2| > 0.05)$ starts to become reasonably small; roughly 1/3 of the mass of $\bar{X}_n$ is outside of $[0.45, 0.55]$. When $n = 1000$, there is an imperceptible chance of obtaining an $\bar{X}_n$ outside of the vertical red lines. If we continued this process for larger and larger n, we would see the mass of $\bar{X}_n$ continue to cluster closer and closer to $\mu = 1/2$. Regardless of how small a ϵ we choose, we can always find an n that fits as much of the mass as we want inside the corresponding red lines.

Note that the law of large numbers does *not* say that $P(|\bar{X}_n - \mu| > \epsilon)$ will necessarily monotonically decrease with n, for each n. For example, we can see that for $\epsilon = .05$, we have that $P(|\bar{X}_1 - \mu| > \epsilon)$ is 0.5 and $P(|\bar{X}_2 - \mu| > \epsilon)$ is about 0.25. All that the LLN says is that $P(|\bar{X}_1 - \mu| > \epsilon)$ will get (arbitrarily) small with n, for any value of ϵ.

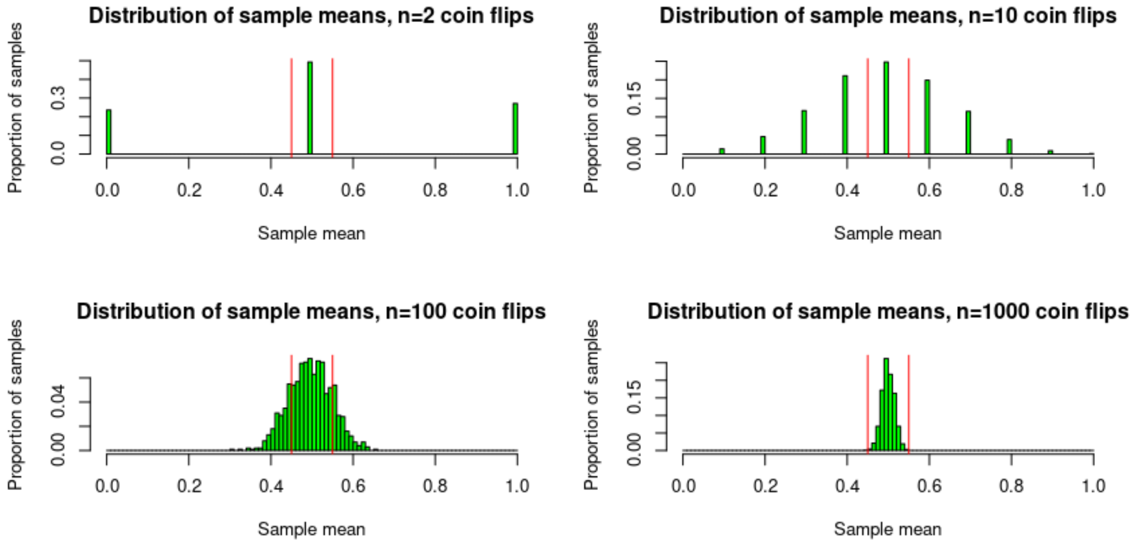


Figure 4.2: Distributions along the sequence $\bar{X}_n$ for a set of n i.i.d. coin flips. Red liness illustrate the mass of the distribution $\bar{X}_n$ that is more than .05 away from 1/2.

The following is the R code I used to generate this figure, if you'd like to copy-paste it and experiment:

```
numsims<-1000
par(mfrow=c(2,2), main="Title")
for (n in c(2,10,100,1000)){
        results<-data.frame(simulation_num=integer(), sample_mean=double())
        for (x in 1:numsims) {
                thissample<-sample(c(0, 1), size = n, replace=TRUE)
                samplemean<-mean(thissample)
                results[x,] = c(x,samplemean)
        }

        h<-hist(results$sample_mean, plot=FALSE, breaks = seq(from=0, to=1, by
            =.01))
        h$density = h$density/100
        plot(h, freq=FALSE, main=paste0("Distribution of sample means, n=",n,"
            coin flips"), xlab="Sample mean", ylab="Proportion of samples", col="
            green")
        abline(v=c(.45,.55), col=c("red", "red"))
}
```

4.3 Convergence in probability and convergence in distribution

Given a sequence of random variables or random vectors $Z_1, Z_2, \dots$, let us now define two notions of convergence of the sequence Z_n. The first is *convergence in probability*:

Definition 4.2. *We say that Z_n converges in probability to Z if for any $\epsilon > 0$:*

$$\lim_{n\to\infty} P(||Z_n - Z|| > \epsilon) = 0$$

In this definition, Z_n can be a random variable/vector. When Z_n is a random variable, then the notation $||Z_n - Z||$ jut refers to the absolute value of the difference: $|Z_n - Z|$. When Z_n is a vector, we can take $||Z_n - Z||$ to be the Euclidean norm of the difference (see Proposition 4.2 for an example).

We will often talk about Z_n converging in probability to a *constant* c. This does not require a second definition because a constant is simply an example of a random variable that has degenerate distribution $P(Z = c) = 1$. Thus we say that Z_n converges in probability to a constant c if $\lim_{n\to\infty} P(|Z_n - Z| > \epsilon) = 0$ for all $\epsilon > 0$.

Notation: When Z_n converges in probability to Z, we write this as $Z_n \xrightarrow{p} Z$, or alternatively $plim(Z_n) = Z$. We say that Z is the *probability limit* of the sequence Z_n. We use the same notation when Z is a constant.

The law of large numbers, for example, says that $\bar{X}_n \xrightarrow{n} \mu$, the sample mean converges in probability to the "population mean", or expectation, of X_i.

Exercise: This problem gives an example of a sequence that converges in probability to another random variable, rather than to a constant. Let $Z_n = Z + \bar{X}_n$, where Z is a random variable and $\bar{X}_n$ is the sample mean of i.i.d. random variables X_i having zero mean and finite variance. Suppose furthermore that Z and $\bar{X}_n$ are independent. Show that $plim(Z_n) = Z$.

Our second notion of convergence of a sequence of random vectors is *convergence in distribution.* Consider first a sequence of scalar random variables:

Definition 4.3. *We say that a random variable Z_n converges in distribution to Z if, for any z such that the CDF $F_Z(z) = P(Z \le z)$ of Z is continuous at z:*

$$\lim_{n\to\infty} P(Z_n \le z) = F_Z(z)$$

Notation: When Z_n converges in distribution to Z, we write this as $Z_n \xrightarrow{d} Z$. As with convergence in probability, Z can be a random vector or a constant.

Note: The requirement that we only consider z where $F_Z(z)$ is continuous is a technical condition, which we can often ignore because we'll be thinking about continuously distributed Z. In general, we can construct examples in which $\lim_{n\to\infty} P(Z_n \le z)$ is not right-continuous (and is thus not a valid CDF), but the valid CDF function $F_Z(z)$ nevertheless captures the limiting distribution of Z_n. In these cases we still want to say that $Z_n \xrightarrow{d} Z$.

The definition given above for convergence in distribution takes Z_n to be a random (scalar) variable to emphasize the idea, but the concept extends naturally to sequences of random vectors. We say that a sequence of random vectors Z_n converges in distribution to Z if for all $\mathbf{z}$ at which the joint CDF of the components of Z $F_Z(\mathbf{z})$ does not have a discontinuity, the limit of the CDF of Z_n evaluated at that point as $n \to \infty$ is $F_Z(\mathbf{z})$.

Convergence in distribution essentially says that the CDF of Z_n point-wise converges to the CDF of Z. By "point-wise", we mean that this occurs for each value z. When $Z_n \xrightarrow{d} Z$, we often refer to Z as the "large-sample" or "asymptotic" distribution of Z_n.

We close this section by investigating the relationship between convergence in probability and convergence in distribution. Convergence in distribution is a weaker notion of convergence (and is in fact often called "weak" convergence), in the sense that it is implied by convergence in probability.

Proposition 4.1. *If $Z_n \xrightarrow{p} Z$, then $Z_n \xrightarrow{d} Z$. In the special case that Z is a degenerate random variable taking value of c, then $Z_n \xrightarrow{d} c$ also implies $Z_n \xrightarrow{p} c$. Thus when Z is degenerate, convergence in distribution and probability are equivalent to one another.*

One manifestation of the fact that convergence in probability is stronger than convergence in distribution is that with the former, covergence of elements of a random vector implies convergence of the whole random vector:

Proposition 4.2. *If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then $\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{p} \begin{pmatrix} X \\ Y \end{pmatrix}$.*

Proof. Since for any $\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$ and $\lim_{n\to\infty} P(|Y_n - Y| > \epsilon) = 0$ holds for any $\epsilon > 0$, let's consider a value $\epsilon/\sqrt{2}$. Let $Z_n := (X_n, Y_n)'$ and $Z := (X, Y)'$. Since $||Z_n - Z|| = \sqrt{(X_n - X)^2 + (Y_n - Y)^2}$ being larger than ϵ is the same as $(Z_n - Z)^2$ being larger than ϵ^2, and since at least one of $(X_n - X)^2$ or $(Y_n - Y)^2$ must then be larger than half of ϵ^2, we have:

$$\begin{aligned} P(||Z_n - Z|| > \epsilon) &= P((X_n - X)^2 + (Y_n - Y)^2 > \epsilon^2) \le P((X_n - X)^2 > \epsilon^2/2 \text{ or } (Y_n - Y)^2 > \epsilon^2/2) \\ &\le P((X_n - X)^2 > \epsilon^2/2) + P((Y_n - Y)^2 > \epsilon^2/2) \end{aligned}$$

we have that

$$\lim_{n\to\infty} P(|Z_n - Z| > \epsilon) = \lim_{n\to\infty} P(|X_n - X| > \epsilon/\sqrt{2}) + \lim_{n\to\infty} P(|Y_n - Y| > \epsilon/\sqrt{2}) = 0 + 0 = 0$$

□

Meanwhile, the same is not true of convergence in distribution: $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$ does not in general imply that $\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ Y \end{pmatrix}$. However, one important special case in which it does is when X or Y is a degenerate random variable. This is useful for example in proving Slutsky's Theorem in Section 4.5.

The next section will introduce the most famous and useful instance of convergence in distribution: the *central limit theorem* (CLT). After introducing the CLT, we will return in Section 4.5 to some further properties of convergence in probability and convergence in distribution, that will be useful in the analysis of large samples.

Optional: There is an even stronger notion of convergence than convergence in probability, referred to as *almost-sure convergence.* We say that Z_n converges almost surely to Z, or, $Z_n \xrightarrow{a.s.} Z$, if

$$P\left(\lim_{n\to\infty} Z_n = Z\right) = 1$$

To make sense of this expression we have to place a probability distribution over entire sequences $\{Z_n\}$ (something we didn't need to do for convergence in probability or convergence in distribution). That is, we imagine a probability space in which each outcome ω yields to a realization of all of the random variables: Z, Z_1, Z_2, Z_3, and so on. Then, the above expression says that $P(\{\omega \in \Omega : \lim_{n\to\infty} Z_n(\omega) = Z(\omega)\}) = 1$. In words: the probability of getting a sequence of Z_n that does not converge to Z with n is zero.

Almost sure convergence is stronger than convergence in probability, i.e. $Z_n \xrightarrow{a.s.} Z$ implies that $Z_n \xrightarrow{p} Z$ (which of course in turn implies that $Z_n \xrightarrow{d} Z$). The *strong law of large numbers* states that the sample mean in fact converges almost surely to the population mean, that is $\bar{X}_n \xrightarrow{a.s.} \mu$.

4.4 The central limit theorem

The central limit theorem (CLT) tells us that if we construct from the sample mean $\bar{X}_n$ the a random variable $Z_n = \sqrt{n}(\bar{X}_n - \mu)$, then the sequence Z_n converges in distribution to that of a normal random variable.

Theorem 2 (central limit theorem). *If X_i are i.i.d random vectors and $\mathbb{E}[X_i'X_i] < \infty$, then*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(\mathbf{0}, \Sigma)$$

where $\Sigma = Var(X_i)$, $\mu = \mathbb{E}[X_i]$, and $\mathbf{0}$ is a vector of zeros for each component of X_i.

The central limit theorem is quite remarkable. It says that *whatever* the distribution of X_i is, the limiting distribution of $\bar{X}_n$ (recentered by μ and rescaled by $\sqrt{n}$) will be a normal distribution. This striking result will pave the way for us to perform inference on the expectation of a random variable, without knowing its full distribution˙

Why the CLT is useful:

The practical value of the CLT is that it delivers an approximation to the distribution of $\bar{X}_n$. For large n, we know that $\sqrt{n}(\bar{X}_n - \mu)$ has approximately the distribution $N(0, \Sigma)$. Using properties of the normal distribution, we can re-arrange this to say that $\bar{X}_n \sim N(\mu, \Sigma/n)$, approximately. To get a good guess of the distribution of $\bar{X}_n$, we only need to have estimates of μ and Σ, which is much easier than estimating the full CDF of X_i from data.

Example: Suppose for simplicity that we had reason to believe that $\Sigma = 1$, i.e. we have a random variable X_i with a variance of 1. However, we don't know μ. We do know then, by the CLT, that for large n, $\bar{X}_n$ is approximately normally distributed around μ with a variance of $1/n$. This is extremely useful, because we can now evaluate candidate values of μ, based on how unlikely we would be to see a value of $\bar{X}_n$ like the one that we calculate, if that value of μ was true. Suppose for example that $n = 100$, and in our sample we observed that $\bar{X}_n = 0.31$. You want to evaluate the possibility that $\mu = 0$. Well, if this were the true value of μ, then given the asymptotic approximation that $\bar{X}_n \sim N(0, 1/n)$ (or equivalently, that $10 \cdot \bar{X}_n \sim N(0, 1)$), we'd only expect to see a value of $\bar{X}_n$ as large as 0.3 once in about 1000 samples. We might thus be willing to rule $\mu = 0$ out as a possibility. This is an example of a *hypothesis test*, which will be covered in Section 5.4.1.

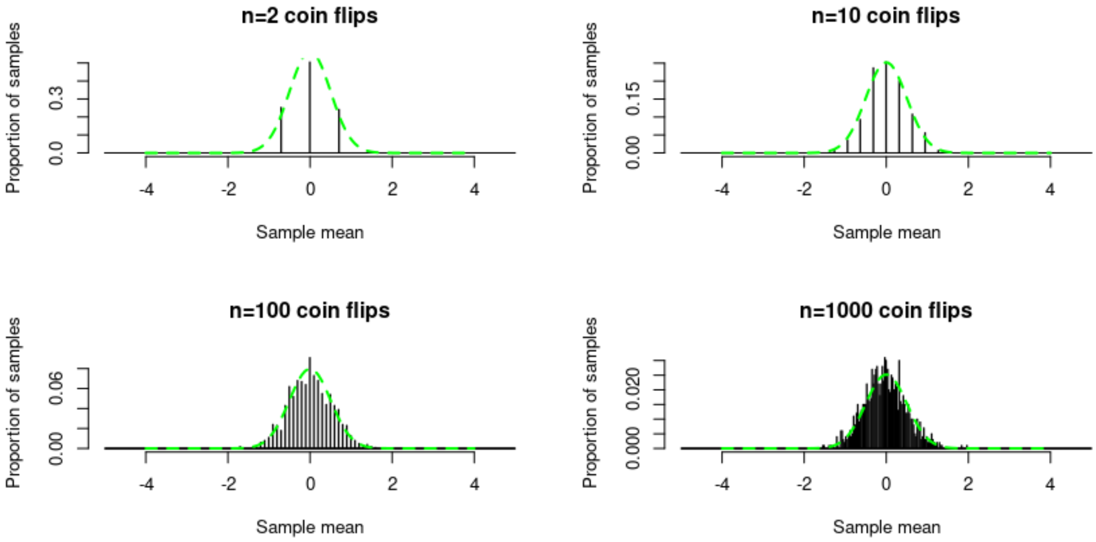


Figure 4.3: The same simulation as in Figure 4.2, except now we plot the distribution of $\sqrt{n}(\bar{X}_n - 1/2)$ rather than of $\bar{X}_n$. The CLT tells us that $\sqrt{n}(\bar{X}_n - 1/2) \xrightarrow{d} N(0, 1/4)$, since $1/4$ is the variance of X_i. Green dashed lines depict what is predicted by the distribution $N(0, 1/4)$, which we can see becomes close to what we see for larger values of n.

Illustrating the CLT:

Figures 4.2 and 4.3 illustrate the CLT in action. Recall that in this example X_i has a two-point distribution $P(X_i = 0) = 1/2$ and $P(X_i = 1) = 1/2$). The distribution of $\bar{X}_n$ becomes closer and closer to a normal distribution centered around $\mu = 1/2$ as n gets large. To the eye, the distribution of $\bar{X}_n$ definately does not look normal for $n = 2$ or for $n = 10$ in Figure 4.2. But by the time we have $n = 100$, it starts to take on the bell-curve shape. We see the variance Σ/n falling as we compare $n = 100$ and $n = 1000$: the latter has a variance about 1/10 as large. In Figure 4.3, we plot the distribution of $\sqrt{n}(\bar{X}_n - 1/2)$ overlaid with its limiting distribution.

Thought experiments like this simulation experiment are useful for getting intuition about the CLT. Accorindly, you often hear descriptions of the CLT along the lines of: "the sample mean becomes normal as the sample gets bigger and bigger". This isn't wrong, but can be a little misleading. A given real-world sample never gets bigger: it always has a single finite size n! Similarly, the sample size n never "goes to infinity"–though we can get pretty close by simulating a sequence of samples on a computer! Imagining an infinite sequence of samples having means $\bar{X}_1$, $\bar{X}_2$, and so on, is just a useful abstraction.

The following proof of the CLT is not necessary for you to know, but you may find it interesting, and being able to follow it is a good study device.

Proof of the CLT:

We'll consider a proof for the univariate case, which can be extended to random vectors using the Cramér-Wold theorem introduced in Section 4.5. The proof here will use the concept of a *moment generating function*:

$$M_X(t) := \mathbb{E}[e^{t\cdot X_i}] = 1 + t \cdot \mathbb{E}[X_i] + \frac{t^2}{2} \cdot \mathbb{E}[X_i^2] + \frac{t^3}{3!} \cdot \mathbb{E}[X_i^3] + \dots \tag{4.1}$$

where the second equality uses the Taylor expansion of e^{tx}. This will be a useful expression for the moment generating function $M_X(t)$. Note that $M_X(t)$ is a (non-random) function of t: the randomness in X_i has been averaged out.

A useful result (that we will not prove here) is that if two random variables X and Y have the same moment generating function $M_X(t) = M_Y(t)$ for all t, then they have the same distribution. Our goal will be to show that whatever the distribution of X_i, the moment generating function of $\sqrt{n}(\bar{X}_n - \mu)$ converges to that of a normal random variable with variance $\sigma^2 = Var(X_i)$.

Let us divide out the variance to rewrite the CLT (in the univariate case) as $\sqrt{n}\cdot\frac{\bar{X}_n-\mu}{\sigma} \xrightarrow{d} N(0,1)$. The moment generating function of the standard normal distribution is:

$$M_Z(t) := \frac{1}{\sqrt{2\pi}} \int dx \cdot e^{tx} \cdot e^{-\frac{x^2}{2}} = e^{-\frac{t^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} \int dx \cdot \cdot e^{-\frac{(x-t)^2}{2}} = e^{-\frac{t^2}{2}}$$

where we've used that $(x-t)^2 = x^2 - 2tx + t^2$ and that the final integral is over the density of a normal random variable with mean t and variance 1.

Now for the magic part. We'll show that whatever the distribution of X_i is, and hence whatever the moment generating function of X_i, the moment generating function of

$$\sqrt{n} \cdot \frac{\bar{X}_n - \mu}{\sigma} = \frac{1}{\sqrt{n}}\frac{X_1 - \mu}{\sigma} + \frac{1}{\sqrt{n}}\frac{X_2 - \mu}{\sigma} + \dots + \frac{1}{\sqrt{n}}\frac{X_n - \mu}{\sigma}$$

will end up being $e^{-\frac{t^2}{2}}$!

First, note that when Y and Z are independent of one another, the moment generating function of $Y + Z$ is equal to the product of each of their moment generating functions, i.e. $\mathbb{E}[e^{t(X_i+Z_i)}] = \mathbb{E}[e^{tY_i} \cdot e^{tZ_i}] = \mathbb{E}[e^{tY_i}\mathbb{E}[e^{tZ_i}]$. Applying this to the above expression, we have that:

$$M_{\sqrt{n}\cdot\frac{\bar{X}_n-\mu}{\sigma}}(t) = M_{\frac{1}{\sqrt{n}}\frac{X-\mu}{\sigma}}(t) \cdot M_{\frac{1}{\sqrt{n}}\frac{X-\mu}{\sigma}}(t) \cdot \dots \cdot M_{\frac{1}{\sqrt{n}}\frac{X-\mu}{\sigma}}(t) = \left(\frac{1}{\sqrt{n}} M_{\frac{X-\mu}{\sigma}}(t)\right)^n$$

Note that for any random variable Y, $M_{\frac{1}{\sqrt{n}}\cdot Y}(t) = M_Y(t/\sqrt{n})$. Therefore, we wish to show that

$$\lim_{n\to\infty} \left(M_{\frac{X-\mu}{\sigma}}(t/\sqrt{n})\right)^n = e^{-\frac{t^2}{2}}$$

for any t.

Applying the Taylor series expansion of the moment generating function in Equation 4.1, we have that:

$$M_{\frac{X-\mu}{\sigma}}(t/\sqrt{n}) = 1 + \frac{t}{\sqrt{n}} \cdot \mathbb{E}\left[\frac{X_i - \mu}{\sigma}\right] + \frac{t^2}{2n} \cdot \mathbb{E}\left[\left(\frac{X_i - \mu}{\sigma}\right)^2\right] + \frac{t^2}{n} \cdot g\left(\frac{t}{\sqrt{n}}\right)$$

where by the Taylor theorem $\lim_{n\to\infty} g\left(\frac{t}{\sqrt{n}}\right) = 0$. Note that $\mathbb{E}\left[\frac{X_i-\mu}{\sigma}\right] = 0$ and $\mathbb{E}\left[\left(\frac{X_i-\mu}{\sigma}\right)^2\right] = 1$, and thus we wish to show that

$$\lim_{n\to\infty}\left(1 + \frac{t^2}{2n} + \frac{t^2}{n} \cdot g\left(\frac{t}{\sqrt{n}}\right)\right)^n = e^{-\frac{t^2}{2}}$$

Recall the identity that $\lim_{n\to\infty}(1 + x/n)^n = e^x$. If we can ignore the g term then we are done. To show that the g term indeed does not contribute in the limit, consider taking the natural logarithm of both sides of the above equation (since the log is continuous function, it preserves limits):

$$\begin{aligned}\lim_{n\to\infty} \ln\left\{\left(1 + \frac{t^2}{2n} + \frac{t^2}{n} \cdot g\left(\frac{t}{\sqrt{n}}\right)\right)^n\right\} &= \lim_{n\to\infty} n \cdot \ln\left(1 + \frac{t^2}{2n} + \frac{t^2}{n} \cdot g\left(\frac{t}{\sqrt{n}}\right)\right) \\ &= \lim_{n\to\infty} n \cdot \left(\frac{t^2}{2n} + \frac{t^2}{n} \cdot g\left(\frac{t}{\sqrt{n}}\right)\right) = -\frac{t^2}{2} + t^2 \cdot \lim_{n\to\infty} \cdot g\left(\frac{t}{\sqrt{n}}\right) \\ &= -\frac{t^2}{2}\end{aligned}$$

where we've used the Taylor theorem for the natural logarithm: $\ln(1 + z) = z + z \cdot h(z)$ where $\lim_{z\to 0} h(z) = 0$, and we have that $\lim_{n\to 0}\left(\frac{t^2}{2n} + \frac{t^2}{n} \cdot g\left(\frac{t}{\sqrt{n}}\right)\right) = 0$.

4.5 Properties of convergence of random variables

This section presents several results that are useful in the analysis of large samples. We will make heavy use of them, for example, when we study the asymptotic properties of the linear regression estimator.

4.5.1 The continuous mapping theorem

The *continuous mapping theorem* (CMT) states that the notions of convergence in probability and convergence in distribution are preserved when we apply a continuous function to each random vector in a sequence Z_n, that is:

Theorem 3 (continuous mapping theorem). *Consider a sequence Z_n of random vectors and a continuous function h. Then:*

- *if $Z_n \xrightarrow{p} Z$, then $h(Z_n) \xrightarrow{p} h(Z)$*
- *if $Z_n \xrightarrow{d} Z$, then $h(Z_n) \xrightarrow{d} h(Z)$*

Example: By the large of large numbers and the CMT: $\left(\bar{X}_n + 5\right) \xrightarrow{p} (\mu + 5)$, where $\mu = \mathbb{E}[X_i]$.

Example: Let $Z_n = \sqrt{n}(\bar{X}_n - \mu)$. Then by the CLT and CMT: $Z_n^2 = n\left(\bar{X}_n - \mu\right)^2 \xrightarrow{d} \chi_1^2$, where χ_1^2 is the chi-squared distribution with one degree of freedom (this is the distribution of a standard normal $N(0, 1)$ random variable squared).

Note: The assumption that h is (globally) continuous can be weakened, which is often important in applications.

- When Z is a contant (call it c), then the convergence in probability part of the CMT only requires that $h(z)$ be continuous at c, rather than everywhere.
- The convergence in distribution part of the CMT can be extended to cases in which h has a set of points $z \in \mathcal{D}$ at which it is discontinuous, provided that $P(Z \in \mathcal{D}) = 0$. This is useful when combined with the CLT, for which Z is continuously distributed. Hence applying an arbitrary function h to $Z_n = \sqrt{n}(\bar{X}_n - \mu)$ allows us to use the CMT provided that h has only a discrete set of points of discontinuity.

A set of useful/common applications of the CMT are summarized by the so-called *Slutsky's Theorem*:

Theorem 4 (Slutsky's Theorem). *Suppose $Z_n \xrightarrow{d} Z$ and $Y_n \xrightarrow{p} c$ with c a constant. Then:*

- $Z_n + Y_n \xrightarrow{d} Z + c$
- $Z_n \cdot Y_n \xrightarrow{d} cZ$
- $Z_n/Y_n \xrightarrow{d} Z/c$ *if* $c \neq 0$.

To see how these results follow from Theorem 3, note that since c is a constant, $Z_n \xrightarrow{d} Z$ and $Y_n \xrightarrow{p} c$ is equivalent to

$$\begin{pmatrix} Z_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} Z \\ c \end{pmatrix}$$

(see discussion following Proposition 4.2). Then we can apply the CMT to the sequence $\begin{pmatrix} Z_n \\ Y_n \end{pmatrix}$, with the following continuous functions h, respectively:

- $h(Z, Y) = Z + Y$
- $h(Z, Y) = Z \cdot Y$
- $Z_n/Y_n = h(Z, Y) = Z/Y$

4.5.2 The delta method

Note that when combined with the CLT, the continuous mapping theorem allows us to talk about the asymptotic distribution of $h\left(\sqrt{n}(\bar{X}_n - \mu)\right)$ for a continuous function h. What is often more useful is to talk about the asymptotic distribution of $\sqrt{n}(h(\bar{X}_n) - h(\mu))$. That is, when we apply a function h to our sample mean, how does the limiting distribution of $h(\bar{X}_n)$ look as it converges around $h(\mu)$? (Exercise: which result allows us to know that $h(\bar{X}_n)$ *does* converge around $h(\mu)$?)

The delta method gives us a tool to address exactly this question:

Theorem 5 (the delta method). *If $\sqrt{n}(Z_n - \mu) \xrightarrow{d} \xi$ for some random vector ξ, then if $h(z)$ is continuously differentiable in a neighborhood of $z = \mu$:*

$$\sqrt{n}(h(Z_n) - h(\mu)) \xrightarrow{d} \nabla h(\mu)'\xi$$

where $\nabla h(z) = (\frac{d}{d_1}h(z), \frac{d}{d_2}h(z), \dots)'$ is a vector of the derivatives of h with respect to each component of Z.

Consider now what this implies in the case of the CLT:

Corollary 1. *If X_i are i.i.d random vectors, $h(x)$ is a function that is continuously differentiable at $x = \mu$, and $\mathbb{E}[X_i'X_i] < \infty$, then*

$$\sqrt{n}(h(\bar{X}_n) - h(\mu)) \xrightarrow{d} N(0, \nabla h(\mu)'\Sigma\nabla h(\mu))$$

where $\Sigma = Var(X_i)$ and $\mu = \mathbb{E}[X_i]$.

Proof. Beginning from Theorem 5, we only need to show that for a random variable $Z \sim N(0,\Sigma)$, $h(\mu)'Z \sim N(0, \nabla h(\mu)'\Sigma\nabla h(\mu))$. We can see this in two steps. First of all, since a linear combination of normal random variables is also normal, we know that $\mathbf{a}'Z$ is normal for any normally-distributed $k-$component random vector and $k-$component vector $\mathbf{a}$. We thus need only to work out the mean and variance of $h(\mu)'Z$ to characterize its full distribution. By linearity of the expectation, $\mathbb{E}[h(\mu)'Z] = 0$, since each component of Z has mean zero. You also showed in HW 3 that the variance of $\mathbf{a}'Z$ is $\mathbf{a}'Z\mathbf{a}$. Substituting $\mathbf{a} = h(\mu)$ completes the proof. □

The most important special case of the corollary above is when X_i is a random variable. In this case, we don't need any matrix multiplication and we have that:

$$\sqrt{n}\left(h(\bar{X}_n) - h(\mu)\right) \xrightarrow{d} N\left(0, \left(\frac{d}{dx}h(\mu)\right)^2 \cdot \sigma^2\right)$$

Note that if the function h is very sensitive to the value of x near μ, i.e. $\frac{d}{dx}h(\mu)$ has a large magnitude, then the asymptotic variance of $h(\bar{X}_n)$ will be large, since the funciton h blows up the variance of X_i by a factor $\left(\frac{d}{dx}h(\mu)\right)^2$.

4.5.3 The Cramér–Wold theorem*

The following theorem, referred to as the Cramér–Wold theorem or the Cramér–Wold "device", is another tool in asymptotic analysis. We won't find it as useful as CMT or delta method, but it's worth seeing so I mention it here:

Theorem 6 (the Cramér–Wold device). *If Z_n is a sequence of random vectors having k components, then $Z_n \xrightarrow{d} Z$ if and only if $\mathbf{a}'Z_n \xrightarrow{d} \mathbf{a}'Z$ for all (non-random) $k-$component vectors $\mathbf{a}$.*

One very important application of the Cramér–Wold device is in extending the central limit theorem to random vectors. In Section 4.4, we only proved the CLT for a random variable. The following exercise asks you to derive the multivariate CLT from the univariate CLT.

Exercise: Use the Cramér–Wold device to show that if Theorem 2 applies to random variables X_i, then it applies to a random vector $X_i = (X_{1i}, X_{2i}, \dots X_{ki})'$ as well (assume that any necessary moments exist).

4.6 Limit theorems for distribution functions*

While the law of large numbers might appear to be somewhat limited, in that it only talks about the mean, it is surprisingly versatile. For example, it implies that sample probabilities converge to their population counterparts. Suppose we have an *i.i.d.* collection of X_i and are interested in $F(x)$, the population CDF of X_i evaluated at some specific x. Then we can define $Z_i = \mathbb{1}(X_i \leq x)$, a random variable that takes a value of 1 if $X_i \leq x$, and zero otherwise. Since the collection $\{Z_1, Z_2, \dots Z_n\}$ is *i.i.d*, and has the finite mean:

$$\mathbb{E}[Z_i] = \mathbb{E}[\mathbb{1}(X_i \leq x)] = P(X_i \leq x) = F(x)$$

the law of large numbers implies that the sample mean of Z_i converges in probability to $F(x)$. The sample mean of Z_i is simply

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(X_i \leq x) = \frac{\textit{number of } i \text{ for which } X_i \leq x}{\textit{number of } i \text{ in sample}},$$

the proportion of the sample for which $X_i \leq x$. When considering this quantity across all x, we call the resulting function the *empirical CDF* of X_i, denoted as $F_n(x)$.

Thus, for each x the empirical CDF evaluated at x converges in probability to the population CDF evaluated at x, i.e. $F_n(x) \xrightarrow{p} F(x)$. This result can be strengthened in two ways (which are not implied by the weak law of large numbers). Consider the error in $F_n(x)$ as an approximation of $F(x)$, $|F_n(x) - F(x)|$ as a function of x. This may be larger or smaller depending on x. The *Glivenko-Cantelli theorem* states that even the largest error, over all x, converges to zero, and furthermore that this convergence is almost sure convergence (see box at the end of Section 4.3), rather than convergence in probability:

Theorem 7 (Glivenko-Cantelli theorem). *If X_i are i.i.d, then:*

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \overset{a.s.}{\to} 0$$

We won't use Theorem 7 in this class, but it can be useful for proving properties of asymptotic sequences that involve quantities that cannot be written as a function of $\bar{X}_n$.

Chapter 5

Statistical decision problems

This chapter presents a formal view of the goals of using statistics for econometrics. It starts with the question: what is it that we would like to learn? Once we've defined our "parameter of interest", we can separate much of econometrics into three parts: identification, estimation and inference.

I will not attempt at a thorough or rigorous treatment of many of the concepts this chapter touches upon. Rather, I hope it can present a unified way to think about several concepts you have probably seen in one form or another in previous courses, and serve either as a reference or a starting point to exploring terms in econometrics as you come across them in your own research.

5.1 Step one: defining a parameter of interest

Why do we use statistics? A short answer is that we want to learn things about the world, and data is the lens with which we investigate some population within it. A more careful answer, which is well-aligned with the specific approach that econometrics takes to using statistics, is that there are specific features θ of the world that we care about.

We can distinguish between three types of *parameter of interest*, θ.

First type (model parameters): Think back to the idea of a parametric statistical model, introduced in Chapter 3.1. Suppose we observe *i.i.d* data X_i, where the distribution of X_i is thought to belong to a parametric family $F(\cdot;\theta)$ for some $\theta \in \Theta$. For example, we might be willing to assume that X_i is a normally distributed random variable, with unknown mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. In this case, $\theta = (\mu, \sigma)$, and in the absence of any further assumptions about θ: $\Theta = \mathbb{R} \times \mathbb{R}^+$, where $\mathbb{R}^+$ is the set of all positive real numbers (the variance cannot be negative). In this context, it is natural to take the full vector of model parameters θ to be our parameter of interest (of course, we might only be interested in e.g. μ, in which case μ alone is our parameter of interest, and similarly with σ).

Second type (features of observed variables in the population): We don't need parametric statistical models to talk about parameters of interest, however. If we have *i.i.d.* data drawn from any population distribution F, we might think of some aspect of F that we'd like to know. For example, we might be interested in $\mathbb{E}[X_i]$, but don't want to assume that X_i is normally distributed, as in the last example. Then, our parameter of interest is $\theta = \mathbb{E}[X_i]$. Another parameter of interest might be the median of F, the point x at which $F(x) = 1/2$. In this case, $\theta = \inf\{x : F(x) \geq 1/2\}$ (this general definition allows for a non-continuous F, in which case there may be no x such that $F(x) = 1/2$ exactly).

Third type (quantities that depend on unobservables): One of the exciting and difficult things about applied econometrics is that often our parameters of interest do not depend solely on the distribution F of the vector of variables X that we observe in our data. Rather, θ often depends also on the distribution of some other variables U that are *not* observed. This situation most often arises when discussing causality, for example when our parameter of interest summarizes the causal effect of a policy. Talking about causality involves some new notation and concepts, so we'll defer further discussion to Chapter 6. As a simpler example of a situation that involves unobservables, let us consider a different important practical problem: measurement error.

Suppose our parameter of interest is $\theta = \mathbb{E}[Z_i]$, the average value of some random variable Z_i. However, our data was not recorded perfectly, and instead of an *i.i.d* sample of Z_i, we observe an *i.i.d* sample of $X_i = Z_i + U_i$, where U_i represents unobserved "measurement error". In this case, our parameter of interest can be written as $\mathbb{E}[X_i - U_i]$, which depends both upon the distribution of X and the distribution of U.

5.2 Identification

Once we have a parameter of interest in mind, a good starting point is often to ask the question: "could I determine the value of θ if I had access to the population distribution F underlying my data?".

If the answer is no, then no amount of statistical wizardry will allow you to learn the value of θ. If the answer is yes, then we say that θ is *identified*.

Definition 5.1. *Given a statistical model $\mathcal{F}$ for (X, U), we say that θ is* ***identified*** *when there is a unique value θ_0 of θ compatible with F_X, the population CDF of observable variables X.*

Often identification is described as saying that if we observed an "infinite" sample, we could determine the value of θ. The reason for this is that by the law of large numbers, we can learn the entire population distribution of X from an *i.i.d* sample X_i, as the sample size goes to infinity (see discussion in Section 4.6). Of course, we never observe an infinitely large dataset, but defining identification in terms of what we *could* know if we did cleanly separates problems of research design from the statistical problem of having too small a sample.

Whenever our parameter of interest is defined directly from the population distribution F_X of observables (e.g. $\theta = \mathbb{E}[X_i]$), it will be identified. Thus, parameters of the second type are always identified. This logic often applies to parameters of the first type as well, except in cases when $F(\cdot;\theta)$ doesn't always change with θ (see example below). Questions of identification usually arise in the third case, when our parameter of interest θ depends on the distribution of unobservables: for example when we're interested in causality, have measurement error, or have "simultaneous equations".

Example: Suppose X_i are *i.i.d* draws from $N(\mu, \sigma^2)$. Then the parameters μ and σ are *identified*, because each pair (μ, σ) gives rise to a different CDF F_X of X_i.

Example: Suppose X_i are *i.i.d* draws from $N(\min\{\theta, 5\}, \sigma^2)$. Then θ is not identified, because different values of θ (e.g. $\theta = 6$ vs. $\theta = 7$), do not give rise to a different CDF F of X_i.

Example: In the measurement error example, suppose that we're willing to assume that $\mathbb{E}[U_i] = 0$, that the measurement error averages out to zero (e.g. there are equal chances of getting positive and negative errors of the same magnitude). Then $\theta = \mathbb{E}[Z_i]$ is identified, since now $\mathbb{E}[Z_i] = \mathbb{E}[X_i]$. This example underscores the role of $\mathcal{F}$ in Definition 5.1. Whether or not θ is identified often depends on what assumptions we are willing to make, which restrict the set $\mathcal{F}$ of possible joint-distributions for (X, U).

Below I discuss some additional issues related to identification, which may relate to terms you've heard floating around about identification:

Parametric vs. non-parametric identification: When $\mathcal{F}$ is a non-parametric statistical model, in the sense described in Section 3.1, we say that θ is *non-parametrically identified.* We have non-parametric identification when we do not need to specify a parametric functional form for the distribution of observables or unobservables. Sometimes we only have parametric identification but not non-parametric identification. Suppose, in the measurement error example, our parameter of interest is full distribution function F_Z of Z_i, and are willing to assume that $U_i \perp Z_i$. Then F_Z is identified if we are willing to specify the exact form of F_U, e.g. $U_i \sim N(0, 1)$, through a technique known as *deconvolution.* However, F_z is not non-parametrically identified.

Partial vs. point identification: Sometimes knowing F_X is not enough to pin down the value of θ, but it is enough to determine a *set* of values that θ might take. For example, we may be able to determine upper and lower bounds for θ. In such cases we often say that θ is *partially identified.* This can be contrasted with Definition 5.1, which describes *point identification.*

Identification of a parametric model: Suppose we have an *i.i.d* sample of observables X_i and a parametric statistical model for (X_i, U_i), in the language of Section 3.1. Then we might say the model is identified, when the full vector θ of model parameters are identified in the sense of Definition 5.1:

Definition 5.2 (full identification of a model). *Given a statistical model $\mathcal{F}$ for (X, U), we say that the model is **identified** when when the set $\{\theta \in \Theta : F_X(\cdot) = F_X(\cdot, \theta)\}$ is a singleton, where F_X is the CDF of X.*

Definition 5.1 says that there is a unique value $\theta_0 \in \Theta$ such that $F_X(\cdot, \theta_0\}$ is equivalent to the population distribution of observables X_i. This situation arises often in econometrics in the context of so-called *structural* models in which the entire model can be characterized by a finite set of model parameters.

5.3 Estimation

If our parameter of interest θ is identified, then we can move on to our next question: how can we estimate it?

In this section, we treat the task of estimating θ as a decision problem. In the next section, we'll take the same approach to testing hypotheses about θ. This way of thinking about estimation and inference is called *statistical decision theory.*

Let's think about the task of estimating θ as a problem of choosing an optimal strategy in a particular game, which we play along with "nature". Nature goes first, giving us a sample $\mathbf{X}$, the distribution of which we denote abstractly as P (this is equivalent to the joint-CDF of all of the components of $\mathbf{X}$). Our goal is to think about how to form $\hat{\theta} = g(\mathbf{X})$ as a function of the data $\mathbf{X}$. How should we proceed?

Recall that in game theory, a *strategy* is a complete profile of what we would do, given whatever the other players do. In this context, we a strategy is not a particular numerical estimate of θ, but the function g. For example, if our estimator is the sample mean, then $g(\mathbf{X}) = g(X_1, X_2, \dots X_n) = \frac{1}{n}\sum_{i=1}^n X_i$, which will depend upon the particular values of X_i occur in our sample.

As in game theory, our best-response to the actions of nature will depend upon our preferences (a.k.a. our utility function). In statistical decision theory this takes the form of a "loss function": $L(\hat{\theta}, \theta_0)$, where θ_0 is the true value of θ. For the most part, we consider the so-called quadratic loss function:

$$L(\hat{\theta}, \theta_0) = ||\hat{\theta} - \theta_0||_2^2 := (\hat{\theta} - \theta_0)'(\hat{\theta} - \theta_0)$$

When θ is a scalar, then this is just the square of the difference between our estimator $\hat{\theta}$ and the true value θ_0.

However, remember that $\hat{\theta} = g(\mathbf{X})$ is a random variable/vector, which depends on our randomly drawn dataset $\mathbf{X}$. Thus to pick a strategy g, we need to define our preferences over "lotteries", again–as in standard game theory. In line with expected utility theory, the convention here is to take our optimal action g to be the minimizer of *expected* loss: $\mathbb{E}[L(\hat{\theta}, \theta_0)]$ where the expectation is over the distribution of $\mathbf{X}$. The *risk* function $R_g(\theta)$ of estimator g views the expected loss as a function of the true value of θ. It is common to write this as $\mathbb{E}_\theta[L(\hat{\theta}, \theta)]$, where the notation $\mathbb{E}_\theta$ makes it clear that the distribution of $\mathcal{X}$ must depend in some way on the value of θ. This is motivated by cases in which we have *i.i.d.* data from a parametric statistical model where θ indexes the population distribution of X_i. Then the distribution of $\mathbf{X}$ depends on just two things: n and the true value of θ.

When we use the quadratic loss function, the optimal estimator g would be

$$g^* := \underset{g}{\operatorname{argmin}}\ \mathbb{E}[||g(\mathbf{X}) - \theta_0)||_2^2] \tag{5.1}$$

However, solving this problem is not easy, because we generally don't know the distribution of $\mathbf{X}$ ex-ante. However, statisticians have developed various strategies to try to keep $\mathbb{E}[||g(\mathbf{X}) - \theta_0)||_2^2]$ small. These strategies are best understood as ways to navigate the so-called *bias-variance tradeoff.* The following proposition shows that expected quadratic loss can be decomposed into two terms: one capturing the square of the "bias" of the estimator, and the other capturing its variance.

For simplicity, we state this result in the special case that θ is a scalar. We'll also just write $\hat{\theta}$ rather than $g(\mathbf{X})$, to keep the notation simple.

Proposition 5.1 (the bias-variance decomposition).

$$\underbrace{\mathbb{E}[(\hat{\theta}-\theta_0)^2]}_{expected\ loss} = \underbrace{\mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2]}_{variance\ of\ \hat{\theta}} + \left(\underbrace{\mathbb{E}[\hat{\theta}]-\theta_0}_{bias\ of\ \hat{\theta}}\right)^2 \tag{5.2}$$

Proof. Add and subract $\mathbb{E}[\hat{\theta}]$ to obtain:

$$\mathbb{E}[\left\{(\hat{\theta}-\mathbb{E}[\hat{\theta}])+(\mathbb{E}[\hat{\theta}]-\theta_0)\right\}^2] = \mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2] + 2\cdot\mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}]-\theta_0)] + \left(\mathbb{E}[\hat{\theta}]-\theta_0\right)^2$$

Now observe that the middle term is zero, because

$$\mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}]-\theta_0)] = (\mathbb{E}[\hat{\theta}]-\theta_0)\cdot\mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}])] = (\mathbb{E}[\hat{\theta}]-\theta_0)\cdot\left(\cancel{\mathbb{E}[\hat{\theta}]}-\cancel{\mathbb{E}[\hat{\theta}]}\right) = 0$$

since $(\mathbb{E}[\hat{\theta}]-\theta_0)$ is just a non-random number. □

Equation (5.2) is described as a bias-variance *tradeoff* because often strategies to decrease bias come at the expense of increasing variance, and vice-versa. Suppose for example that we just pick $g(\cdot) = 5$, estimating θ to be 5, regardless of what sample we see. This estimator will have zero variance! But we can expect the bias $5-\theta_0$ to be quite large. On the other hand, extremely flexible estimation methods are often good at minimizing bias, but doing so may increase variance. The field of *non-parametric* estimation chooses estimators to explicitly navigate this tradeoff.

5.3.1 Desirable properties of an estimator

This section investigates some desirable properties of an estimator, in light of the bias-variance tradeoff. It is not meant to deliver a detailed account of these properties, but simply to serve as a reference for what the associated terms mean. Please see the course textbooks for more details.

5.3.1.1 Consistency

The first thing that we might ask of our estimator is that it be *consistent.* What we mean by that is that

$$\hat{\theta} \xrightarrow{p} \theta_0$$

regardless of the value of θ_0. Consistency means that as n goes to infinity, the entire expected loss in Equation (5.2) converges to zero.

5.3.1.2 Rate of convergence

Consider the sample mean $\bar{X}_n$ viewed as an estimator of the population mean $\mu = \mathbb{E}[X_i]$. We know by the LLN that $\bar{X}_n$ is a consistent estimator of μ, and we furthermore know by the CLT that

$$n^p(\bar{X}_n-\mu) \xrightarrow{d} N(0, Var(X_i))$$

if we set $p = 1/2$. Note that if we set the power p on n to be any larger than $1/2$, then the LHS would blow up, rather than converging in distribution to anything (like a normal distribution). On the other hand, if we had set $p < 1/2$, then $n^p(\bar{X}_n-\mu)$ will simply converge in probability to zero. $1/2$ is "Goldilocks" level of p in which we get a non-degenerate asymptotic distribution for $n^p(\bar{X}_n-\mu)$.

In general, when we have a consistent estimator $\hat{\theta}$, we call the maximum value of p such that $n^p(\hat{\theta}-\theta_0)$ converges in distribution to something (technically, to some distribution that is "bounded in probability") the *rate of convergence* of $\hat{\theta}$. The rate of convergence of the sample mean is $1/2$, and we often say that it is $\sqrt{n}-$*consistent*. $\sqrt{n}$-consistency is a desirable property, which is shared by many common estimators. However, some estimators have a slower rate of convergence. For example, suppose we'd like to estimate the density $\theta = f(\mathbf{x})$ of a $d-$dimensional random vector X_i at some point $\mathbf{x}$, and we'd like to make this estimation *non-parametric*—that is, not based on assuming a parametric model for $f(\mathbf{x})$.

We can do so using the so-called kernel density estimator $\hat{f}_K(\mathbf{x})$, which has a rate of convergence no better than $p = \frac{2}{d+4}$. When $d = 1$, for example, we can only blow up $(\hat{f}_K(\mathbf{x}) - f(\mathbf{x}))$ by a factor of $n^{2/5}$ and get an asymptotic distribution. In practice, this means that we need a *larger* sample n for asymptotic arguments to provide good approximations to the sampling distribution of $\hat{f}_K(\mathbf{x})$. This becomes a real problem as d starts to increase: for example, the rate of convergence $\hat{f}_K(\mathbf{x})$ when $d = 5$ is just $2/9$. This problem is often referred to as the *curse of dimensionality*, and is why we need very large samples—and/or even more clever techniques—to do non-parametric estimation with many covariates.

5.3.1.3 Unbiasedness

If an estimator is *unbiased* if it manages to make the second term in Equation (5.2) zero, that is:

$$\mathbb{E}[\hat{\theta}] = \theta_0$$

Unbiasedness has a nice interpretation: we know that $\hat{\theta} \neq \theta_0$ in general, but we know that $\hat{\theta}$ will be right *on average*, over different realizations of our dataset.

An example of an unbiased estimator is the sample mean, when our parameter of interest θ is the population mean $\mathbb{E}[X_i]$. In Section 4.1, we showed indeed that $\mathbb{E}[\bar{X}_n] = \mathbb{E}[X_i]$, regardless of n or the true value of $\mathbb{E}[X_i]$ (so long as it exists).

Note that an estimator can be consistent without being unbiased. For example, the estimator $\hat{\theta} = \frac{n+1}{n}\bar{X}_n$ is biased as an estimator for $\theta_0 = \mathbb{E}[X_i]$, because

$$\mathbb{E}[\hat{\theta}] - \theta_0 = \frac{n+1}{n} \cdot \theta_0 - \theta_0 = \frac{\theta_0}{n} \neq 0$$

unless $\theta_0 = 0$. However, this $\hat{\theta}$ is consistent. This implies that as n approaches infinity, both its bias and its variance converge to zero. If an estimator has an asymptotic bias (that is, a bias that doesn't go away with n), then it cannot be consistent.

5.3.1.4 Efficiency

Econometricians often speak of an estimator as being efficient. Loosely speaking, this typically means that $\hat{\theta}$ minimizes mean squared error (5.2) among some class of estimators.

For example, we might consider the class of unbiased estimators, and ask whether a given $\hat{\theta}$ minimizes Eq. (5.2). Since the bias term is zero for all estimators in this class, the efficient estimator will be the one that minimizes variance.

In the context of parametric models, the *Cramer-Rao lower-bound* establishes the smallest variance that an unbiased estimator can possibly have (even when θ is a vector, though the definition of "smallest" here requires qualification). The maximum likelihood estimator, discussed in the next section, achieves this bound: it is thus efficient, whenever it happens to be unbiased (which is not guaranteed in general).

A related notion is *asymptotic efficiency*, which says that an estimator is efficient as $n \to \infty$.

5.3.2 Important types of estimator*

In this section we'll introduce a few types of estimator that are common or important in econometrics.

In light of the asymptotic theorems of Chapter 4, it should be clear then when θ can be expressed as a continuous function of the population expectations of various random variables, then an estimator composed of that same function applied to sample means might be expected to have desirable properties. The ordinary least-squares (OLS) linear regression estimator can be seen as an estimator of this kind, and Chapter 7 will offer a detailed look at it.

In this section, I present a few prominent examples of a category of estimator called *extremum-estimators*. This category overlaps the one I just described: OLS can also be seen as an extremum estimator. An extremum-estimator takes the form:

$$\hat{\theta} = \underset{\theta \in \Theta}{\text{argmin}}\ Q(\mathbf{X}, \theta) \tag{5.3}$$

where $Q(\mathbf{X}, \theta)$ is some function of θ and our sample. Why might (5.3) be a good way to construct an estimator? The basic idea is to find cases in which $Q(\mathbf{X}, \theta) \xrightarrow{p} Q_0(\theta)$ for some function $Q_0(\theta)$ which has a unique minimizer within a space Θ of possible values for θ. As the sample gets larger, our $\hat{\theta}$—which solves the sample minimization problem—converges to θ_0, which solves the population problem $Q_0(\theta)$. The function Q might be provided by economic theory, a statistical model, or some combination of the two.

An important subclass of extremum estimators are those in which $Q(\mathbf{X}, \theta)$ takes the form of a sample mean: $Q(\mathbf{X}, \theta) = \frac{1}{n} \cdot \sum_{i=1}^{n} m(X_i, \theta)$ for some $m(X_i, \theta)$. Such estimators are called *M-estimators*. Given the law of large numbers, M-estimation will work well, in the sense of at least being *consistent*, under fairly general technical conditions on m and the distribution of X_i (which we will not discuss here).

5.3.2.1 Method of moments*

The basic *method of moments* considers a $k-$dimensional parameter-of-interest θ, and a set of l *moment-conditions*:

$$\begin{aligned} \mathbb{E}[g_1(X_i, \theta)] &= 0 \\ \mathbb{E}[g_2(X_i, \theta)] &= 0 \\ &\cdots \\ \mathbb{E}[g_l(X_i, \theta)] &= 0 \end{aligned} \tag{5.4}$$

which we can gather into a vector equation $\mathbb{E}[\mathbf{g}(X_i, \theta)] = \mathbf{0}$. The most prominent example of a model like (5.4) is the linear regression model, which we will study in detail in Chapter 7. Another example is the linear instrumental variables model with l instruments and k endogenous variables. The moment conditions (5.4) are a statistical model, in the sense of Section 3.1, because they restrict the distribution of X_i to be among those for which $\mathbb{E}[\mathbf{g}(X_i, \theta_0)] = \mathbf{0}$, where θ_0 is the true value of θ. If Equation (5.4) has only one solution (which must then be equal to θ_0), then the parameter θ is *identified*. In general, a necessary condition for identification is $l \geq k$.

The basic method-of-moments estimator $\hat{\theta}_{MM}$ considers the case in which $k = l$; i.e. we have the same number of moments and parameters we wish to estimate. In this setting, we simply replaces the population expectations in (5.4) by their sample counterparts, that is we solve choose $\hat{\theta}_{MM}$ such that $\frac{1}{n} \sum_{i=1}^{n} \mathbf{g}(X_i, \hat{\theta}_{MM}) = \mathbf{0}$. Note that this equation is the first-order-condition of the extremum-estimator in which

$$Q(\mathbf{X}, \theta) = \left|\left| \frac{1}{n} \sum_{i=1^n} \mathbf{g}(X_i, \theta)) \right|\right|_2^2 = \left(\frac{1}{n} \sum_{i=1^n} \mathbf{g}(X_i, \theta) \right)' \left(\frac{1}{n} \sum_{i=1^n} \mathbf{g}(X_i, \theta) \right)$$

The most famous example of a method-of-moments estimator is the ordinary least squares estimator for the linear regression model, which we will study in depth in Section 7.5.

The so-called *generalized method of moments* (GMM) considers the more general case in which $l \geq k$. When we have more moment equations than parameters we aggregate over them with some $k \times l$ matrix B to yield k equations: $\mathbf{B}\mathbb{E}[\mathbf{g}(X_i, \theta)] = \mathbf{0}$ which are linear combinations of the original ones (where now $\mathbf{0}$ has k components). The GMM estimator $\hat{\theta}_{MM}$ minimizes the sample analog of this set of equations, $\hat{\theta}_{MM} = \underset{\theta \in \Theta}{\text{argmin}}\ Q(\mathbf{X}, \theta)$, where

$$Q(\mathbf{X}, \theta) = \left|\left| \mathbf{B} \left(\frac{1}{n} \sum_{i=1^n} \mathbf{g}(X_i, \theta)) \right) \right|\right|_2^2 = \left(\frac{1}{n} \sum_{i=1^n} \mathbf{g}(X_i, \theta) \right)' \mathbf{W} \left(\frac{1}{n} \sum_{i=1^n} \mathbf{g}(X_i, \theta) \right)$$

where $\mathbf{W} := \mathbf{B}'\mathbf{B}$ is an $l \times l$ matrix that is positive-semidefinite. Note that $\hat{\theta}_{GMM}$ depends on our choice of $\mathbf{B}$ through $\mathbf{W}$. *Efficient GMM* uses a data-driven weighting matrix $\hat{\mathbf{W}}$ in order to minimize the asymptotic variance of $\hat{\theta}_{GMM}$.

Note: A related method uses *moment inequalities*, in which the equals signs in Equation (5.4) are replaced with inequalities $\leq$. Models with moment inequalities typically result in partial identification of θ.

Note that the method of moments does not require us to have a fully-specified statistical model $F(\cdot;\theta)$, in the sense of Section 3.1. Rather, it is typically sufficient to have $l \leq k$ moment conditions of the form Equation (5.4) that involve our parameter of interest θ. The method of moments can thus be thought of as employing a *semi-parametric* model.

5.3.2.2 Maximum likelihood estimation*

When we do have a fully parametric model of the form $F(\cdot;\theta)$, then maximum likelihood estimation is often a good bet. The *maximum likelihood* approach to estimation begins with an *i.i.d.* sample X_i, in which we assume a parametric model $F(x;\theta)$ for the population distribution of X_i. For simplicity of notation, suppose that X_i is continuously distributed so that $F(\mathbf{x};\theta)$ admits a density function $f(\mathbf{x};\theta)$.

Consider the function $L(\theta) = \mathbb{E}[\log(f(X_i;\theta))]$. It's not obvious, but we have the following very interesting result:

Proposition 5.2. *The true value of θ_0 maximizes $L(\theta)$.*

Proof. First write, for any $\theta \neq \theta_0$:

$$L(\theta) - L(\theta_0) = \mathbb{E}\left[\log(f(X_i,\theta))\right] - \mathbb{E}\left[\log(f(X_i,\theta_0))\right] = \mathbb{E}\left[\log(f(X_i,\theta)) - \log(f(X_i,\theta_0))\right] = \mathbb{E}\left[\log\left(\frac{f(X_i,\theta)}{f(X_i,\theta_0)}\right)\right]$$

Because log is a concave function, we have by a property of expectation known as *Jensen's inequality* that $\mathbb{E}\left[\log\left(\frac{f(X_i,\theta)}{f(X_i,\theta_0)}\right)\right] \leq \log\left(\mathbb{E}\left[\frac{f(X_i,\theta)}{f(X_i,\theta_0)}\right]\right)$. Now, since $f(\cdot,\theta_0)$ is the true density of X_i, this is in turn equal to

$$\log\left(\mathbb{E}\left[\frac{f(X_i,\theta)}{f(X_i,\theta_0)}\right]\right) = \log\left(\int \cancel{f(\mathbf{x},\theta_0)} \cdot \frac{f(\mathbf{x},\theta)}{\cancel{f(\mathbf{x},\theta_0)}} \cdot \mathbf{dx}\right) = \log(1) = 0$$

Thus, we've shown that $L(\theta) - L(\theta_0) \leq 0$ for any θ! □

Note: Provided that $P\left(f(X_i,\theta) \neq f(X_i,\theta_0)\right) > 0$ for any $\theta \neq \theta_0$, the above can be strengthened to say that θ_0 is the *unique* maximizer of $L(\theta)$.

The maximum likelihood estimator $\hat{\theta}_{MLE}$ minimizes the sample analog of $L(\theta)$ with respect to θ. In particular, let the *likelihood function* of the data $\mathcal{L}(\mathbf{X};\theta)$ be:

$$\mathcal{L}(\mathbf{X};\theta) := \prod_{i=1}^{n} f(X_i,\theta)$$

Given that the X_i are *i.i.d.*, $\mathcal{L}(\mathbf{X};\theta)$ is the joint density of the observed dataset $mathbf X = \{X_1, X_2, \dots X_n\}$, given the model $f(\cdot;\theta)$. Maximizing $\mathcal{L}(\mathbf{X};\theta)$ with respect to θ is equivalent to maximizing the logarithm of $\mathcal{L}(\mathbf{X};\theta)$ with respect to θ, leading to the more-familiar "log-likelihood" expression for $\hat{\theta}_{MLE}$:

$$\hat{\theta}_{MLE} = \underset{\theta\in\Theta}{\text{argmax}}\ \log\left(\prod_{i=1}^{n} f(X_i,\theta)\right) = \underset{\theta\in\Theta}{\text{argmax}}\ \frac{1}{n}\ \underbrace{\sum_{i=1}^{n} \log\left(f(X_i,\theta)\right)}_{\text{“log-likelihood function”:}}$$

where we insert a factor of 1/4 (which doesn't change the argmax) in order to view this as a sample analog of Proposition (5.2). This problem has a unique solution if for example the the log-likelihood function $\log(\mathcal{L}(\mathbf{X};\theta))$ is globally-concave, as it does in many familiar models (e.g. the "probit" model). The above expression also reveals that $\hat{\theta}_{MLE}$ is an example of an M-estimator in which the function $m(x,\theta) = \log\left(f(x,\theta)\right)$.

When $\hat{\theta}_{MLE}$ is an unbiased estimator (which is only true in some applications), it is efficient among all unbiased estimator (look up the Cramer-Rao bound for details). It is also asymptotically efficient among so-called $\sqrt{n}$-consistent "regular" estimators. These properties however require the model $f(\mathbf{x};\theta)$

to be correctly specified, which may be a hard assumption to defend. When the model is misspecified, then generally $\hat{\theta}_{MLE} \xrightarrow{p} \theta^*$, where

$$\theta^* = \underset{\theta \in \Theta}{\text{argmin}} \ -\mathbb{E}\left[\log\left(\frac{f(\mathbf{x};\theta)}{f(\mathbf{x})}\right)\right]$$

where $f(\mathbf{x})$ is whatever the true density function of X_i is. The quantity inside the minimization is called the Kullback-Leibler divergence between $f(\mathbf{x};\theta)$ and $f(\mathbf{x})$, and can be thought of as a kind of "distance" between the two density funcitons. Thus, when the model is misspecified, $\hat{\theta}_{MLE}$ is consistent for a "pseudo-parameter" that finds the closest point in Θ to θ_0, as measured by the Kullback-Lieber divergence. Misspecification occurs when Θ is not a sufficiently big or rich set to contain $f(\mathbf{x}) = f(\mathbf{x};\theta_0)$ for some $\theta_0 \in \Theta$.

5.3.2.3 Bayes estimators*

An alternative approach to estimation is based on the principles of *Bayesian statistics*. The basic idea of the Bayesian approach to estimation is to imagine that both our sample $\mathbf{X}$, *and* our parameter of interest θ are random.This accords with the Bayesian notion that probabilities reflect degrees of belief. You don't need to take this interpretation seriously however to define a Bayes estimator. You can think of it instead as a way to think of a class of interesting estimators for θ.

In our interpretation of $(\mathbf{X}, \theta)$ as jointly random, the marginal distribution of θ is called the *prior* and is denoted here as $\pi(\theta)$. It can be interpreted as our beliefs about θ before we see the data $\mathbf{X}$. We call the conditional distribution of θ given the data $\mathbf{X}$, denoted $\pi(\theta|\mathbf{X})$, the *posterior* distribution of θ. The two are related by Bayes rule (albeit with some different notation than we saw before):

$$\pi(\theta|\mathbf{X}) = \frac{\pi(\theta)}{P(\mathbf{X})} \cdot \mathcal{L}(\theta;\mathbf{X}) \tag{5.5}$$

Note: We're using somewhat abstract notation here, but π can be interpreted as a probability density function over θ. $P(\mathbf{X})$ can be interpreted as a density function or as a probability mass function of the data. $\mathcal{L}(\theta;\mathbf{X})$ is the likelihood function we saw in the last section on maximum likelihood estimation: this corresponds to a parametric statistical model, and represents the probability of drawing sample $\mathbf{X}$, given θ.

Given our loss function $L(\hat{\theta}, \theta_0)$ as in Section 5.3, we can define expected loss over the full joint-distribution of $(\mathbf{X}, \theta)$. This is now

$$\mathbb{E}[L(\hat{\theta},\theta)] = \mathbb{E}\left[\mathbb{E}\left[L(\hat{\theta},\theta)|\mathbf{X}\right]\right] \tag{5.6}$$

by the law of iterated expectations. The inner expectation represents an average over values of θ with respect to the posterior distribution $\pi(\theta|\mathbf{X})$, and delivers a function of $\mathbf{X}$ (and θ_0). The outer expectation integrates over the marginal distribution of the data, just as in Equation 5.1.

Our goal is now to minimize (5.6) with respect to $\hat{\theta}$, to provide an estimate of θ_0. Note that since $\hat{\theta} = g(\mathbf{X})$ is a function of $\mathbf{X}$, the optimal g must minimize $\mathbb{E}\left[L(\hat{\theta},\theta)|\mathbf{X}\right]$ for *every* possible value $\mathbf{x}$ of $\mathbf{X}$ (otherwise we could just change $g(\mathbf{x})$ for that $\mathbf{x}$ along while decreasing (5.6)). Thus, we can focus on the problem of minimizing the risk $\mathbb{E}\left[L(\hat{\theta},\theta)|\mathbf{X}\right]$ with respect to θ, for a given value of $\mathbf{X}$. The function of $\mathbf{X}$ that does this is called the *Bayes estimator*, $\hat{\theta}_B$ (let's assume the problem is such that it is unique).

When we use the quadratic loss function $L(\hat{\theta},\theta) = (\hat{\theta}-\theta)^2$, as in Equation 5.1 (focusing on the case of a scalar θ for simplicity), our problem can now be written as

$$\hat{\theta}_B = \underset{\theta \in \Theta}{\text{argmin}} \ \mathbb{E}\left[L(\hat{\theta},\theta)|\mathbf{X}\right] = \underset{\theta \in \Theta}{\text{argmin}} \ \left\{\int \pi(\theta|\mathbf{X}) \cdot (\hat{\theta}-\theta)^2 \cdot d\theta\right\}$$

The first order condition of this minimization problem yields that

$$\hat{\theta}_B = \int \pi(\theta|\mathbf{X}) \cdot \theta \cdot d\theta,$$

i.e. Bayes estimator $\hat{\theta}_B$ is equal to the so-called *posterior-mean* of θ. Note that we can use Eq. (5.5) to write the posterior mean in terms of our prior π on θ and the likelihood function:

$$\hat{\theta}_B = P(\mathbf{X})^{-1} \cdot \int \pi(\theta) \cdot \mathcal{L}(\theta;\mathbf{X}) \cdot \theta \cdot d\theta = \frac{\int \pi(\theta) \cdot \mathcal{L}(\theta;\mathbf{X}) \cdot \theta \cdot d\theta}{\int \pi(\theta) \cdot \mathcal{L}(\theta;\mathbf{X}) \cdot d\theta}$$

where pull the $P(\mathbf{X})$ factor out of the integral since it does not depend on θ, and then express it in terms of the likelihood function given that $\pi(\theta|\mathbf{X})$ has to integrate to one. Although it takes the form of an extremum estimator in which $Q(\mathbf{X},\theta)=\mathbb{E}\left[L(\hat{\theta},\theta)|\mathbf{X}\right]$, we do not need to solve an explicit minimization problem to compute the Bayes estimator. Rather, we only need to be able to evaluate the above integral (typically numerically), given the likelihood function implied by our model and the prior $\pi(\theta)$.

$\hat{\theta}_B$ thus has two ingredients provided by the researcher: a statistical model which delivers $\mathcal{L}\theta;\mathbf{X}$, and a prior π on θ. Since the value of $\hat{\theta}_B$ depends on how the researcher chooses this prior, you might wonder whether how we could ever be confident that $\hat{\theta}_B$ could even be consistent for θ_0, even if the model is correctly specified. Couldn't you pick a really bad prior? The *Bernstein-von Mises theorem* says that for finite-dimensional θ, the posterior mean and the maximum likelihood estimator $\hat{\theta}_{MLE}$ will converge asymptotically under general conditions.

Essentially, the influence of the prior $\pi(\theta)$ wanes as the sample becomes large, and the posterior mean ends up finding the peak of the likelihood function, just as $\hat{\theta}_{MLE}$ does. One way to see this is to take the log of the posterior:

$$\log(\pi(\theta|\mathbf{X}))=\log\left(\frac{\pi(\theta)}{P(\mathbf{X})}\cdot\prod_{i=1}^{n}f(X_i,\theta)\right)=-C_n+\log(\pi(\theta))+\sum_{i=1}^{n}\log\left(f(X_i,\theta)\right)$$

As n gets large, the data contribute n terms to this sum, while the prior always contributes the same $\log(\pi(\theta))$. The other term $C_n:=\log(P(\mathbf{X}))$ is also growing with n, but doesn't depend on θ. Thus, the dependence of the posterior distribution on θ is entirely driven by the likelihood, asymptotically.

5.4 Inference*

In Section 5.3, our goal was to deliver a *point-estimate* $\hat{\theta}$ of our parameter of interest. That is, we want a number that yields something close to the true value θ_0 of θ.

Sometimes we can settle for a less ambitious goal, which is to ask not what the exact value of θ_0 is, but rather we want to know whether or not θ_0 belongs to some *set* of values. I will discuss two approaches of this type: i) *hypothesis testing*, in which we want to test whether $\theta_0\in\Theta_0$ for some fixed set Θ_0; and ii) *interval estimation*, in which we want to construct a set $\hat{\Theta}$ that has some desirable relationship to θ_0 (for example contains θ_0 with high probability)

5.4.1 Hypothesis testing

Beginning with some overall space of admissible values Θ (e.g. the real numbers), let us carve the space into two sets: Θ_0 and Θ_1, where $\Theta_0\cup\Theta_1=\Theta$ and $\Theta_0\cap\Theta_1=\emptyset$. We call our hypothesis that $\theta_0\in\Theta_0$ the *null-hypothesis*:

$$\text{(Null hypothesis)}\quad H_0:\theta_0\in\Theta_0\qquad\qquad\text{(Alternative hypothesis)}\quad H_1:\theta_0\in\Theta_1$$

Note that provided that our model $\theta_0\in\Theta$ is correctly specified, either the null hypothesis H_0 or the alternative hypothesis H_1 holds.

Continuing of the approach of statistical decision theory, we may think of our action space as now as consisting of two actions $d\in\{a,r\}$, either accept (a) or reject (r) the null-hypothesis H_0. This can be contrasted with estimation, in which our action space was to pick a specific value in Θ to serve as an estimate for θ.

In this context, a *strategy* is a mapping from the possible datasets $\mathbf{X}$ that we might see to an action $\{a,r\}$. This function $d(\mathbf{X})$ is referred to as a decision rule, or a *test*. To think about what kind of a test might be optimal, we again need to specify our preferences, or a loss function, over actions. Compared with estimation, in which our loss function took the form $L(\hat{\theta},\theta)$, it now takes the form $L(d,\theta_0)$: how happy would we be with our decision $d\in\{a,r\}$, if we learned the true value of θ was θ_0?

Compared with estimation—where the quadratic loss function is very standard—in testing it is less obvious what our cost function could be. One thing is clear however, we'd prefer not to be *wrong*: we don't want to reject the null hypothesis (often referred to as *failing to accept* the null) when in fact $\theta\Theta_0$, and we also don't want to accept the null hypothesis when in fact $\theta_0\in\Theta_1$. The first of these errors is called a Type-I error (falsely rejecting H_0) while the second is called a Type-II error (incorrectly accepting H_0).

The most basic loss function we might think of is called *0-1 loss*, and only cares about *whether* we are right or not, i.e. $L(d, \theta_0)$ when either $d = a$ and $\theta_0 \in \Theta_0$ or $d = r$ and $\theta_0 \in \Theta_1$ (i.e. we are right), and $L(d, \theta_0)$ otherwise (we are wrong). Recall that since $\mathbf{X}$ is random, our decision $d(\mathbf{X})$ will be random, and thus we can again think about the *risk*, or expected loss, due to a particular strategy d. With the $0-1$ loss function:

$$\mathbb{E}[L(d(\mathbf{X}), \theta_0)] = \begin{cases} P(d(\mathbf{X}) = r) & \text{if } \theta_0 \in \Theta_0 \qquad \text{(Type-I error)} \\ P(d(\mathbf{X}) = a) & \text{if } \theta_0 \in \Theta_1 \qquad \text{(Type-II error)} \end{cases}$$

It is clear from the above that whether or not the null is actually true determines *which* probability matters in determining the risk of the test.

Since the value of θ pins down some aspect of the distribution of $\mathbf{X}$, the probability of rejecting the null will depend upon what the true value of θ_0 in fact is. Like the risk function that we saw in estimation, let us use the notation $P_\theta(d(\mathbf{X}) = r)$ to denote the probability of rejecting when the true value is θ. Viewing this as a function of θ, we define the *power function* $\beta(\theta)$ of test d.

Beyond the $0-1$ loss function, we might put a different penalty on Type-I vs. Type-II errors:

$$L(a, \theta_0) = \begin{cases} 0 & \text{if } \theta_0 \in \Theta_0 \\ \ell_{II} & \text{if } \theta_0 \notin \Theta_0 \end{cases} \qquad \text{while} \qquad L(r, \theta_0) = \begin{cases} 0 & \text{if } \theta_0 \in \Theta_1 \\ \ell_I & \text{if } \theta_0 \notin \Theta_1 \end{cases}$$

The ratio ℓ_{II}/ℓ_I will govern whether our test d should be more conservative about avoided Type-I errors, or about avoiding Type-II errors.

5.4.2 Desirable properties of a test

As with estimation problems, choosing the optimal test d is a hard problem because we don't know the distribution of $\mathbf{X}$, we can only approximate it using the dataset $\mathbf{X}$ that we actually observe, along with whatever assumptions we are willing to make. As again with estimation, there are a few principles that are used to help guide the design of statistical tests.

5.4.2.1 Size

The *size* α of a test d is the maximum probability of making a Type-I error (falsely rejecting), over all $\theta \in \Theta_0$. We can write this in terms of the power-function $\beta(\theta)$ as:

$$\alpha = \sup_{\theta_0 \in \Theta_0} \beta(\theta)$$

We'd like the size of a test d to be small; we therefore often design tests to control their size (keep it below a certain value). Often we can do this in the asymptotic limit (as $n \to \infty$) even if we do now know the size of a test in finite sample.

5.4.2.2 Power

The power of a test is given by its power function $\beta(\theta)$. We generally want to increase $\beta(\theta)$ among the $\theta \in \Theta_1$, to reduce the probability of a Type-II error.

5.4.2.3 Navigating the tradeoff

In general, the two desiderata of a) a small size; and b) large power, are at tension with one another. A test that always rejects, regardless of the data $\mathbf{X}$, will never make a Type-II error (have lots of power), but may be extremely likely to make a Type-I error (have large size). On the other hand, a test that always accepts will never make a Type-I error (have low size) but may be making lots of Type-II errors (have low power). Often we approach testing by choosing a *significance-level* p ex-ante, (e.g. $p = .05$), and then design the test so that it's size is no greater than p. Given that constraint, we then try to make the power of the test as large as possible (which usually means making it's size exactly p).

5.4.3 Constructing a hypothesis test

The most common variety of hypothesis test takes the following form: from the data $\mathbf{X}$ we compute some *test statistic*, call it T_n. Then we compare T_n to some *critical value* c, and choose to reject the null-hypothesis if and only if $|T_n|$ exceeds the critical value (a so-called *two-sided test*), or alternatively if T_n exceeds the critical value (a so-called *one-sided test*).

Tests of this form are usually motivated by knowing the asymptotic distribution of T_n, i.e. $T_n \xrightarrow{d} T$ where T has some known distribution. Then we can control the size of our test by choosing c to be such that $P(T \leq c) \geq 1 - \alpha$. We then maximize power subject to this contraint on size by choosing c to be exactly the $1 - \alpha$ quantile of T (and no lower), so that $P(T \leq c) = 1 - \alpha$.

Example: Let us close by illustrating some of the concepts of this section with an example. Suppose our statistical model is that $X_i \sim N(\theta_0, 1)$, i.e. a normal random variable with unit variance but unknown mean θ_0. We wish to test whether $H_0 : \theta_0 = 0$, that is: $\Theta_0 = \{0\}$ and $\Theta_1 = \mathbb{R}/\{0\}$. Let our test statistic be $\sqrt{n}$ times the sample mean $T_n = \sqrt{n}{\cdot}\bar{X}_n$. Given our model, the sample mean has the exact distribution $\bar{X}_n \sim N(\theta_0, 1/n)$ for any n, and hence $T_n \sim N(\theta_0, 1)$. Under the null, T_n is a standard normal (since then $\theta_0 = 0$) and hence for a two-sided test we can choose our critical value c to be the $1 - \alpha/2$ quantile of the standard normal distribution (then $P(|T_n| > c) = P(T_n < -c) + P(T_n > c) = \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha$). Note that the power function $\beta(\theta)$ of this test is the probability that a $N(\theta, 1)$ random variable has absolute value greater than c, which is equal to $\Phi(c - \theta) + \Phi(-c - \theta)$, where Φ denotes the standard normal CDF.

5.4.4 Interval estimation and confidence intervals

The goal of interval estimation is to choose an a set $\hat{\Theta}$ of values that with high probability contains the true value θ_0. We call this interval estimation because $\hat{\Theta}$ typically corresponds to an interval $[a, b]$ (if θ is one-dimensional), or some higher-dimensional analog of an interval (e.g. a region). By contrast, estimation in the sense of Section 5.3 is by contrast referred to as *point*-estimation.

As with point estimation and testing, our action $\hat{\Theta}$ is a function of the data (however now this is a set-valued function)—call it $s(\mathbf{X})$. The *coverage probability* of an interval estimator s is the probability that it contains the true value of θ_0 Here the tradeoff is between increasing the coverage probability, but without making the the interval too big (in which case we haven't learned much about the value of θ_0). Thus with interval estimation, we might define our loss function to depend both on the coverage probability and the length of the interval estimate.

As with estimation, we *do* care about the specific value of θ, not just whether or not some hypothesis H_0 about it is true. However, we'll now see that there is a very close connection between interval estimation and hypothesis testing.

One scenario in which we might implement interval estimation is when our parameter of interest is only partially identified (see Section 5.2). In such a setting, for example, our model might only imply that $\theta_0 \in [\theta_L, \theta_H]$, where the bounds θ_L and θ_H are themselves point identified. Then we can construct an interval estimate of θ_0 with the set $\hat{\Theta} = [\hat{\theta}_L, \hat{\theta}_H]$, given estimators of each of the two bounds.

The much more common scenario in which we engage in interval estimation is when constructing a *confidence interval* for θ_0. We do this even when θ_0 is identified and we have a consistent estimator for it. A confidence interval makes a much more credible than a point estimate. In fact, point-estimation is just a special case of interval estimation in which we constrain our $\hat{\Theta}$ to be a singleton. While singleton will sets typically have zero probability containing θ_0 (though they may be very close to it with high probability), confidence intervals allow us to deliver an interval estimate of θ_0 that takes sampling uncertainty into account.

5.4.4.1 Confidence intervals by test inversion

The most popular method for constructing confidence intervals is to perform a hypothesis test having size α for the null $H_0 : \theta_0 = \theta$, for each conceivable value of θ. Then, collect the set of all values θ that are not rejected by that test to form our interval estimate of θ_0. That is:

$$\hat{\Theta} = \{\theta \in \Theta : d(\mathbf{X}) = a\}$$

This process is often referred to as *test inversion*, and the resulting $\hat{\Theta}$ is called a $(1 - \alpha)$-confidence interval $\mathcal{CI}^{1-\alpha}$. For example, if we used a test with size 5%, then the resulting confidence interval is called a 95% confidence interval.

Example: Suppose we apply this principle to the example in Section 5.4.3 in which $X_i \sim N(\theta_0, 1)$. There we constructed a test for the null hypothesis that $\theta_0 = 0$, but now we need to consider more general hypotheses of the form $H_0 : \theta_0 = \theta$. If we revise our test statistic to be $T_n(\theta) = \sqrt{n} \cdot (\bar{X}_n - \theta)$, we again have that T_n has a standard normal distribution asymptotically, and thus our critical value c is unchanged from the $\theta = 0$ case. A $1 - \alpha$ confidence interval would thus be:

$$\mathcal{CI}^{1-\alpha} = \{\theta \in \mathbb{R} : |T_n(\theta)| \leq c\} = \{\bar{X}_n - c/\sqrt{n}, \bar{X}_n + c/\sqrt{n}\}$$

where c is the $1 - \alpha$ quantile of the standard normal distribution.

Chapter 6

Brief intro to causality

Most interesting questions in social science concern *causality.* We aren't just interested in observing what happens in the social world, but why and how. And we're usually interested in what changes to policy or behavior could lead to changes that we might deem desirable.

These types of questions concern causality. The meaning of the term "causal" is a long-standing philosophical question; see Lewis (1973) for a fairly modern treatment that will accord with our approach in this class. We will take a very simple perspective: A causes B if B *would* be different if A were different. For example, on a day in which rain was forecast and I took my umbrella to school, we might say that the rain forecast caused me to bring my umbrella, if I *wouldn't have* taken the umbrella, absent the forecast for rain. We of course can't directly observe what would have happened if the forecast had been different; we call this a *counterfactual.*

6.1 Causality as counterfactual contrasts

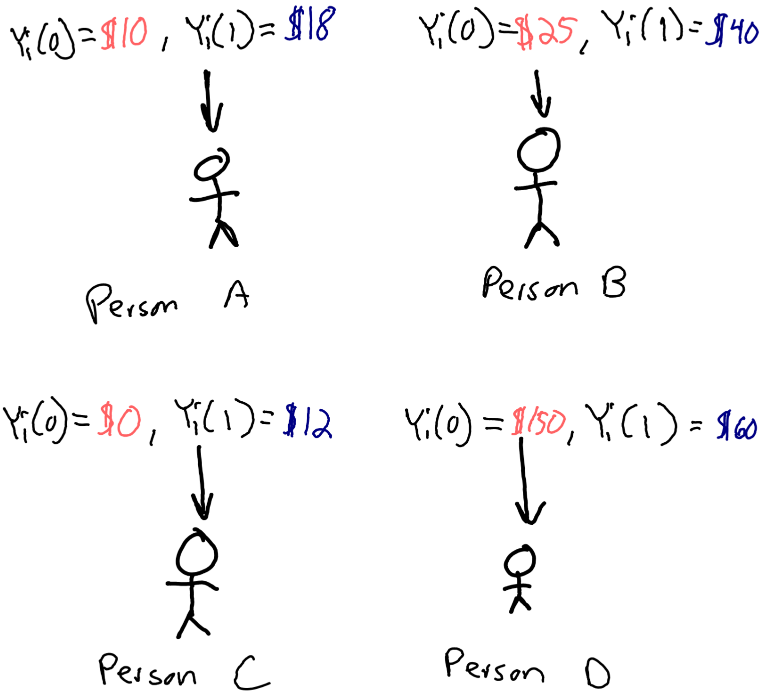
The potential outcomes framework offers an elegant and tractable way to talk about counterfactuals, in the language of random variables (Rubin, 1974). This connects questions of causality to questions of statistics, which we have been developing tools to study.

As a running example, consider the question of the effect of obtaining a college degree on a worker's earnings. Suppose we have data in the form of an *i.i.d* sample of (D_i, Y_i), where $D_i \in \{0, 1\}$ indicates whether individual i completed a college degree, and Y_i indicates the workers average hourly earnings at age 30. We call D_i our *treatment* variable, and Y_i our *outcome* variable. We're interested in the causal effect of the treatment variable on the outcome. This is a setting in which we have a *binary* treatment. We'll start here because it's the simplest setting to develop the concept of causality. In Section 6.5 I'll discuss how these ideas generalize beyond a binary treatment.

Definition 6.1. *An individual's* ***potential outcomes*** *are:* $(Y_i(1), Y_i(0))$*, where* $Y_i(1)$ *is the outcome they would receive if they received the treatment, and the outcome* $Y_i(0)$ *they would receive if they did not.*

In the returns-to-college example, $Y_i(0)$ is the earnings i would have if they didn't go to college, and $Y_i(1)$ is the earnings that i would have if they did go to college. The key thing to keep in mind in the definition of counterfactuals is that we assume each individual i has a well-defined value both of $Y_i(0)$ *and* of $Y_i(1)$. Regardless of whether i went to college or not, there is an answer to the question of how much they would earn if they did go to college, and how much they would earn if they did not.

Consider for example a population composed of four individuals, pictured below. Person A would earn \$10 an hour if they didn't graduate college, but if they did go to college they would get a higher-paying job that paid them \$18 an hour at age 30. Person B is a higher earner, and would earn \$25 an hour without a college degree, and would earn \$40 and hour with one. Person C would choose to leave the labor force and earn \$0 without a degree, but with a college degree would find a job that pays \$12 an hour. Notice that for all three of these individuals, $Y_i(1) > Y_i(0)$: the causal effect of college on their earnings is positive. But this not need be the case: suppose person D would found a successful company if they didn't go to college, earning them \$150 an hour by age 30, but if they did go to college they would have missed a chance opportunity to start the company and earned \$60 as an employee somewhere else.



Definition 6.2. *An individual's **treatment effect** is defined as: $\Delta_i = Y_i(1) - Y_i(0)$, the difference between their treated and untreated potential outcomes.*

In the above example, the treatment effects Δ_i are \$8 an hour for Person A, \$15 an hour for Person B, \$12 an hour for Person C, and $\$-90$ an hour for Person D. On average, treatment effects or positive—although Person D's individual treatment effect is negative.

The leap of faith that you need to take withe the potential outcomes is to believe that there exists a value $Y_i(0)$ *and* $Y_i(1)$ for each individual, regardless of whether they actually went to college. If i does graduate college (i.e $D_i = 1$), then their actual earnings Y_i, will be $Y_i = Y_i(1)$. Similarly, if they don't go to college, then their earnings will be $D_i = 0$. Another way of writing this is that, for each i:

$$Y_i = D_i \cdot Y_i(1) + (1 - D_i) \cdot Y_i(0)$$

Notice that since $D_i \in \{0, 1\}$, there is always one of the above terms that is equal to zero, and the other term gives us the appropriate potential outcome.

In the above example, suppose that Persons A and D do go to college and graduate, while B and C do not. Then if we measure the earnings and college-graduation status of each of the four individuals, our data will be $\{(Y_i, D_i)\}_{i=1,2,3,4} = \{(\$18, 1), (\$25, 0), (\$0, 0), (\$60, 0)\}$.

What can we say about *treatment effects*, given this data? Consider for example individual D, who in reality missed their opportunity to start the business and earn \$150 an hour. This is a *counterfactual*, something that would have happened if the world were different. Since we can't observe what would have happened, we'll never be able to answer the question of what person D's value of Δ_i is, empirically.

Definition 6.3. *The **fundamental problem of causal inference** is that for a given i, we only observe one of the two potential outcomes: either $Y_i(1)$ if $D_i = 1$, or $Y_i(0)$ if $D_i = 0$. In other words, we only observe i's **realized value** $Y_i = Y_i(D_i)$, and not their other potential outcome.*

The fundamental problem of causal inference means that we have a problem of identification, in the language of Section 5.2. Suppose our parameter of interst is $\Delta_i = Y_i(1) - Y_i(0)$ for some particular individual i. Suppose that they did graduate from college, so $D_i = 1$. If we can't observe $Y_i(0)$, we can't identify their treatment effect. However, we'll see that we can still sometimes make statements about average treatment effects, by using *other* students who didn't go to college as a comparison group.

6.2 The difference in means estimator and selection bias

The *difference-in-means estimator* takes the difference in the sample average of the outcome variable among the "treatment group" $D_i = 1$ and "control group" $D_i = 0$:

$$\hat{\theta}_{DM} = \frac{1}{N_1} \sum_{i:D_i=1} Y_i - \frac{1}{N_0} \sum_{i:D_i=0} Y_i$$

where N_1 is the number of individuals i in the sample such that $D_i = 1$, and N_1 is the number of individuals in the sample such that $D_i = 0$. Of course, the total sample size $n = N_0 + N_1$.

We know from the results of 4.6, and the midterm that for large samples $\hat{\theta}_{DM}$ converges in probability to its population counterpart:

$$\hat{\theta}_{DM} \xrightarrow{p} \{\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]\} \tag{6.1}$$

Suppose that in our data, we observe that Y_i and D_i are positively correlated, that is $\hat{\theta}_{DM} \geq 0$ (as you showed in a homework problem). This suggests that $\mathbb{E}[Y_i|D_i = 1] \geq \mathbb{E}[Y_i|D_i = 0]$. Can we conclude from our data that going to college causes ones earnings at age 30 to be higher?

We know from Definition 6.3 that for any individual for whom $D_i = 1$, our observed Y_i is $Y_i = Y_i(1)$. Similarly for any individual who doesn't go to college, $Y_i = Y_i(0)$. Thus, we can rewrite the estimand of our difference-in-means estimator as:

$$\theta_{DM} = \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \tag{6.2}$$

Notice that the first term in Eq. (6.2) conditions on the event $D_i = 1$, and the second term conditions on $D_i = 0$. This means that the difference-in-means estimand compares two different groups, which might not be comparable to one another. For example, students who go to college might have higher SAT scores than students who do not.

Suppose for the moment that the second term in Eq. (6.2) also conditioned on the event $D_i = 1$ (rather than $D_i = 0$). If this were the case, then we could use linearity of the expectation to rewrite θ as being equal to $\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1]$, the average treatment effect Δ_i among students who do go to college. We call this the *average treatment effect on the treated*, or *ATT*. The ATT is a causal parameter of interest, because it compares the values of $Y_i(1)$ and $Y_i(0)$, on average, for the *same group*.

Note that by adding and subtracting $\mathbb{E}[Y_i(0)|D_i = 1]$ to equation (6.2), we can write:

$$\theta_{DM} = \underbrace{\{\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1]\}}_{ATT} + \underbrace{\{\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]\}}_{\text{selection bias}} \tag{6.3}$$

The parameter if interest *ATT* is not identified, unless the selection bias term $\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]$ is equal to zero. This term represents a measure of non-comparability between the students who go to college and the students who do not, in terms of their counterfactual earnings $Y_i(0)$.

For example, students who obtain a college degree may be more likely to come from family backgrounds in which their parent(s) had time and resources to help the student accumulate skills that are valued by the labor market. As a result, these students would have earned more on average, even if they didn't go to college and hence $\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]$. Many other stories also lead to a positive correlation between D_i and $Y_i(0)$: students whose parents are well-connected may be more likely to go to college, and earn more even if they didn't go to college, and any genetic traits that are associated with higher earnings are likely to also increase college attendance.

6.3 Randomization eliminates selection bias

A sufficient condition for the selection bias term to be zero is that $\mathbb{E}[Y_i(0)|D_i = 1] = \mathbb{E}[Y_i(0)|D_i = 0]$, which says that $Y_i(0)$ is *mean-independent* of D_i. One case in which this will hold is when D_i is assigned completely randomly, as in a randomized controlled trial. In this case, we have:

Definition 6.4. ***Random assignment*** *says that* $(Y_i(0), Y_i(1)) \perp D_i$

The random assignment assumption is stronger than we need to kill the selection bias term in Equation (6.3). All we need for that is $\mathbb{E}[Y_i(0)|D_i = 1] = \mathbb{E}[Y_i(0)|D_i = 0]$. This is implied by random assignment, because $(Y_i(0), Y_i(1)) \perp D_i$ implies that $Y_i(0) \perp D_i$, which in turn implies that $\mathbb{E}[Y_i(0)|D_i = 1] =$

$\mathbb{E}[Y_i(0)|D_i = 0]$ (*Note:* recall that independence implies uncorrelatedness, and furthermore that for a binary D_i and any random variable V_i, $V_i \perp D_i$ holds if and only if $\mathbb{E}[V_i|D_i = 1] = \mathbb{E}[V_i|D_i = 0]$. You may want to review the earlier homework problem on this to convince yourself).

When the selection-bias term in Equation (6.3) is equal to zero, we can say that the ATT is *identified*, in the language of Section 5.2. There is only one value of ATT compatible with the population distribution of observables, since (Y_i, D_i) is observed and $ATT = \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 1]$. However, under random-assignment we can actually say more. Not only is the ATT identified, but so is the *average treatment effect*:

$$ATE = \mathbb{E}[Y_i(1) - Y_i(0)] = P(D_i = 1) \cdot ATT + (1 - P(D_i = 1)) \cdot ATU$$

where $ATU := \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 0]$ is the average treatment effect on the untreated, and we've used the law of iterated expectations to decompose ATE into the ATT and ATU. Since the random-assignment assumption says that treated potential outcomes $Y_i(1)$ are also independent of treatment D_i, we have note only that $\mathbb{E}[Y_i(0)|D_i = 1] = \mathbb{E}[Y_i(0)|D_i = 0]$, but also that $\mathbb{E}[Y_i(1)|D_i = 1] = \mathbb{E}[Y_i(1)|D_i = 0]$, and thus the ATT, ATU, and ATE are all equal to one another.

In non-experimental settings, one may be able to identify a parameter like the ATT without being able to identify the ATE. An example of this is the difference-in-differences research design, which (in its basic, most common form) only yields identification of the ATT and not the ATU or ATE.

Note: Even the above argument that $ATT = ATU = ATE = \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$ only ever makes use of $Y_i(0)$ being independent of D_i, and $Y_i(1)$ being independent of D_i. This is still weaker than the assumption made above, that $Y_i(0)$ and $Y_i(1)$ are *jointly* independent of D_i. In practice, it's usually hard to come up for an argument for why only the marginal distributions of $Y_i(1)$ and $Y_i(0)$ would be independent of D_i, and not their joint distribution, which is why I've written it the way I have.

Note: Definition 6.4 corresponds to a randomized controlled trial with *perfect compliance*. In many real-world trials, the only thing that can be randomized is whether an individual is *assigned* to receive treatment. But subjects may still choose whether to actually receive treatment. This situation is very common in economics settings, see for example the homework problem about a lottery to migrate to New Zealand. In these cases, one can use the method of *instrumental variables* to estimate causal effects, which you'll see in later courses.

An assumption implicit in our use of random assignment to eliminate selection bias above is that each individual's potential outcomes does not depend on whether *other* individuals go to college. This is known as the *stable unit treatment value assumption*, or SUTVA. This is not always a harmless assumption, as it rules out spillover effects.

6.4 The selection-on-observables assumption

Outside of an actual experimental setting, the random-assignment assumption is very strong. Typically, economic agents "select into" treatment, meaning they choose for themselves whether or not $D_i = 1$ or $D_i = 0$. Their are usually a variety of reasons why the circumstances and preferences that lead to a choice of taking treatment can be expected to be correlated with potential outcomes.

Suppose we observe a vector of covariates X_i, along with Y_i and D_i. Then, the following assumption is often considered to be weaker than assuming fully random-assignment:

Definition 6.5. ***Selection-on-observables****, also referred to as* ***unconfoundedness****, says that*

$$\{(Y_i(0), Y_i(1)) \perp D_i\}|X_i$$

Selection-on-observables makes the same assumption as random assignment, but we assume it holds conditional on each value X_i. It is thus often also called a *conditional independence assumption*.

Why might selection-on-observables be more reasonable than random assignment? The basic idea is that if we observe a rich enough set of X_i, we might be able to control for confounding factors that lead to selection bias. For example, in the returns-to-college example, we might include in the vector X_i whether ot not i's parents graduated from college, their socio-economic status, and i's test scores in high school.

Imagine that we observed literally *everything* that matters for determining the outcome Y_i, in addition to treatment. In this case, we could write potential outcomes as

$$Y_i(0) = Y(0, X_i) \quad \text{and} \quad Y_i(0) = Y(0, X_i),$$

where the function $Y(d,x)$ is common to everybody: once we know d and x we can say exactly what is going to happen to you. Then selection-on-observables would be satisfied automatically, since if we condition on $X_i = x$, then $Y_i(d) = Y(d,x)$ for either $d \in \{0,1\}$. Notice that $Y(d,x)$ doesn't depend on i: it is no longer random once we've fixed X_i. It is hence is uncorrelated with D_i, since degenerate random variables are statistcally independent of everything! This can be seen as mimicking the logic of a carefully controlled experiment in the natural sciences, in which we make sure "everything else" that matters X_i is held fixed, while varying D_i between 0 and 1.

A similar logic would apply if X_i includes everything that determines D_i: e.g. $D_i = d(X_i)$ for some function d. Then we'd also get selection-on-observables for free. In practice, apart from vey specific settings, we'll never observe everything that determines outcomes Y_i, or selection into treatment D_i. However, if we can control for most of obvious threats to eliminating selection bias, we might be willing to think that our X_i get us most of the way there. For a clever and compelling example of using selection-on-observables, I recommend looking at Dale and Krueger (2002).

How does the selection-on-observables assumption help us? Note that if it holds then

$$\begin{aligned}\mathbb{E}[Y_i|X_i = x, D_i = 1] - \mathbb{E}[Y_i|X_i = x, D_i = 0] &= \mathbb{E}[Y_i(1)|X_i = x, D_i = 1] - \mathbb{E}[Y_i(0)|X_i = x, D_i = 0] \\ &= \mathbb{E}[Y_i(1)|X_i = x] - \mathbb{E}[Y_i(0)|X_i = x] \\ &= \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x] := ATE(x) \qquad (6.4)\end{aligned}$$

Thus, the average treatment effect, conditional on X is identified by a version of the difference in means estimand that conditions on any given value x of X_i. Let us denote this parameter as $ATE(x)$. Equation 6.4 shows that under selection-on-observables, it is identified. Since we also observe the marginal distribution of X_i, we can then recover for example the overall average treatment effect by integrating over values of the control variables:

$$ATE = \int \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x] \cdot dF(x) = \int ATE(x) \cdot dF(x)$$

which follows by the law of iterated expectations.

There are three main approaches to making use of the selection-on-observables assumption in this way: *inverse-propensity score weighting*, *matching*, and *regression*. In this class, we'll focus on the third of these, regression, but I briefly introduce the other two in the box at the end of this section. The three approaches can be thought of as essentially three different strategies to construct an estimator for $ATE(x)$, but are all fundamentally based off of the identification result (6.4).

Exercise: Show that the difference-in-means estimator won't work in general under selection-on-observables.

Solution: Suppose for the sake of argument that X_i is discrete. Then, by LIE:

$$\begin{aligned}
\theta_{DM} &= \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] \\
&= \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \\
&= \sum_x P(X_i = x|D_i = 1) \cdot \mathbb{E}[Y_i(1)|X_i = x, D_i = 1] - \sum_x P(X_i = x|D_i = 0) \cdot \mathbb{E}[Y_i(0)|X_i = x, D_i = 0] \\
&= \sum_x P(X_i = x|D_i = 1) \cdot \mathbb{E}[Y_i(1)|X_i = x] - \sum_x P(X_i = x|D_i = 0) \cdot \mathbb{E}[Y_i(0)|X_i = x] \\
&= \sum_x P(X_i = x|D_i = 1) \cdot \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x] + \sum_x \{P(X_i = x|D_i = 1) - P(X_i = x|D_i = 0)\} \cdot \mathbb{E}[Y_i(0)|X_i = x]
\end{aligned}$$

Note: When using the selection-on-observables assumption, it is important that the variables in the vector X_i are *unaffected* by treatment. That is, if we introduced potential outcomes $X_i(0)$ and $X_i(1)$, we would have $X_i(0) = X_i(1)$ for all i. To make sure of this, researchers typically consider variables X_i that are measured earlier in time than treatment D_i is assigned. When this condition fails, causal inference can fail even when the selection-on-observables assumption holds, via a problem often referred to as "bad-control".

Alternative approaches using selection-on-observables:

Inverse propensity score weighting: Under selection-on-observables:

$$ATE = \mathbb{E}\left[\frac{D_i \cdot Y_i}{P(D_i = 1|X_i)} - \frac{(1 - D_i) \cdot Y_i}{1 - P(D_i = 1|X_i)}\right]$$

This strategy is known as inverse propensity score weighting. If one estimates the function $\mathcal{P}(x) = P(D_i = 1|X_i = x)$ (known as the *propensity score function*) for all values of x, then one can form the object within the expectation above, using $\mathcal{P}(X_i)$ for each observation.
To see this, note that by the law of iterated expectations and selection-on-observables:

$$\begin{aligned}
&\mathbb{E}\left[\frac{D_i \cdot Y_i}{P(D_i = 1|X_i)} - \frac{(1 - D_i) \cdot Y_i}{1 - P(D_i = 1|X_i)}\right] \\
&\quad= \mathbb{E}\left\{\mathbb{E}\left[\frac{D_i \cdot Y_i}{P(D_i = 1|X_i)} - \frac{(1 - D_i) \cdot Y_i}{1 - P(D_i = 1|X_i)}\middle| X_i\right]\right\} \\
&\quad= \int \left\{\mathbb{E}\left[\frac{D_i \cdot Y_i}{P(D_i = 1|X_i)} - \frac{(1 - D_i) \cdot Y_i}{1 - P(D_i = 1|X_i)}\middle| X_i = x\right]\right\} \cdot dF_X(x) \\
&\quad= \int \left\{\frac{\mathbb{E}[D_i \cdot Y_i|X_i = x]}{P(D_i = 1|X_i = x)} - \frac{\mathbb{E}[(1 - D_i) \cdot Y_i|X_i = x]}{1 - P(D_i = 1|X_i = x)}\right\} \cdot dF_X(x) \\
&\quad= \int \left\{\frac{\mathbb{E}[D_i \cdot Y_i(1)|X_i = x]}{P(D_i = 1|X_i = x)} - \frac{\mathbb{E}[(1 - D_i) \cdot Y_i(0)|X_i = x]}{1 - P(D_i = 1|X_i = x)}\right\} \cdot dF_X(x) \\
&\quad= \int \left\{\frac{\cancel{P(D_i = 1|X_i = x)} \cdot \mathbb{E}[Y_i(1)|D_i = 1, X_i = x]}{\cancel{P(D_i = 1|X_i = x)}}\right. \\
&\qquad\qquad \left. - \frac{\cancel{(1 - P(D_i = 1|X_i = x))} \cdot \mathbb{E}[Y_i(0)|D_i = 0, X_i = x]}{\cancel{1 - P(D_i = 1|X_i = x)}}\right\} \cdot dF_X(x) \\
&\quad= \int \{\mathbb{E}[Y_i(1)|X_i = x] - \mathbb{E}[Y_i(0)|X_i = x]\} \cdot dF_X(x) \\
&\quad= \mathbb{E}[Y_i(1) - Y_i(0)] = ATE
\end{aligned}$$

Matching: The approach of *matching* finds within one's sample pairs of units having similar values of X_i. In the most basic version of this strategy (*one-to-one, exact* matching), we for each treated unit i find a control unit i' such that $X_i = X_{i'}$. We drop any control units that are not matched, and then apply the difference in means estimator to the modified sample, and consider this as an estimate of the ATT. However, finding pairs such that $X_i = X_{i'}$ can be difficult when X is a vector with many components, and will be impossible when it includes any components that are continuously distributed. In these cases we'd need to settle for finding an i' such that $X_i \approx X_{i'}$ (note that this issue is not specific to matching, we still have to worry about it when evaluating Eq. (6.4)).

However, a clever application of selection-on-observables (Rosenbaum and Rubin (1983)) allows us to simplify the problem considerably, leading to the idea of *propensity-score matching.* It also follows from the selection-on-observables assumption that for any $p \in (0,1)$ we have that:

$$\mathbb{E}[Y_i|D_i = 1, \mathcal{P}(X_i) = p] - \mathbb{E}[Y_i|D_i = 0, \mathcal{P}(X_i) = p] = \mathbb{E}[Y_i(1) - Y_i(0)|P(X_i) = p]$$

where $\mathcal{P}(x) = P(D_i = 1|X_i = x)$ is the propensity score function introduced above. This expression says that conditioning on values of the *propensity score* rather than on X_i itself is sufficient to estimate causal effects. This is useful because while X_i may have many components, the propensity score is always a scalar. Thus, we simply need to estimate the function $\mathcal{P}(x)$, and then match units i and i' such that $\mathcal{P}(X_i) \approx \mathcal{P}(X_{i'})$, rather than finding a good way to compare X on all dimensions.

6.5 Causality beyond a binary treatment*

In this chapter we've focused on a *binary* treatment, which takes just two values: $D_i = 1$ ("treatment"), and $D_i = 0$ ("control"). However, we're often interested in the causal effect of a treatment variable that takes on many values. For example, what is the effect of *years* of schooling on earnings, rather than just the effect of completing any college degree?

Setting up the notation for multivalued treatment variables is pretty straightforward. We can define our potential outcomes $Y_i(d)$ in the same way as before, where now d index all of the values that D_i might take. Here are some examples:

- Let d be the number of years of schooling student i completes, and $Y_i(d)$ be their earnings at age 30.
- Let d be the price of some good, and let the function $Y_i(d)$ be the demand function for that good in market i.
- Let d be the high school in Georgia that student i attends, and let $Y_i(d)$ be an indicator for whether they were accepted to UGA, e.g. $d \in \{\text{school A}, \text{school B}, \text{school C}, \textit{etc.}\}$.
- In a randomized experiment about the effect of social media on mental health, subjects i are assigned to three different treatments:

$$d \in \{\text{no social media}, \text{Facebook only}, \text{Twitter only}, \text{Facebook and Twitter}\}$$

Regardless of the setting, we can still define random assignment ($(Y_i(1), Y_i(0)) \perp D_i$) and selection-on-observables ($\{(Y_i(1), Y_i(0)) \perp D_i\}\, |X_i$) exactly as we did before.

However, with more than two values of treatment, there are now many different ways to think about treatment effects. For example, in the first example above, we can think about the effect of finishing grade 12 as:

$$Y_i(12) - Y_i(11),$$

while the effect of completing high-school versus dropping out after grade 10 is:

$$Y_i(12) - Y_i(10)$$

The overall average causal effect of the last year of schooling that each student actually completes would be

$$\mathbb{E}[Y_i(D_i) - Y_i(D_i - 1)]$$

In the first two examples above, the values of treatment D_i have a natural order to them. In the third and fourth examples, treatment is categorical, and there may not be a natural such order. With an unordered treatment, like in the last example, we might pick one comparison category and consider treatment effects with respect to it, e.g. separately estimating $\mathbb{E}[Y_i(\text{Facebook only}) - Y_i(\text{no social media})]$, $\mathbb{E}[Y_i(\text{Twitter only}) - Y_i(\text{no social media})]$ and $\mathbb{E}[Y_i(\text{Facebook and Twitter}) - Y_i(\text{no social media})]$.

6.6 Moving beyond average treatment effects*

Although our discussion here has been focused on parameters that *average* over treatment effects $\Delta_i = Y_i(1) - Y_i(0)$, this isn't the only type of causal question that we can answer with random-assignment or selection-on-observables.

Consider a binary treatment D_i and random assignment: $(Y_i(0), Y_i(1)) \perp D_i$. Note that we can apply any function $g(\cdot)$ to the potential outcomes, without destroying independence, i.e. $(g(Y_i(0)), g(Y_i(1))) \perp D_i$. Why is this useful? Consider the function $g(t) = \mathbb{1}(t \leq y)$ for some value y. Given that random-assignment implies that the random variable $\mathbb{1}(Y_i(1) \leq y)$ is independent of D_i, we have that

$$\underbrace{\mathbb{E}[\mathbb{1}(Y_i \leq y)|D_i = 1]}_{F_{Y|D=1}(y)} = \mathbb{E}[\mathbb{1}(Y_i(1) \leq y)|D_i = 1] = \underbrace{\mathbb{E}[\mathbb{1}(Y_i(1) \leq y)]}_{F_{Y(1)}(y)}$$

The term on the left is the conditional CDF of Y_i given $D_i = 1$, which can be computed from the data. The term on the right is the (unconditional) CDF of the treated potential outcome $Y_i(1)$. This expression shows that we can identify the CDF of $Y_i(1)$ at any point y. Collecting over all y, we can thus compute the entire distribution of $Y_i(1)$.

By the same logic, we can also identify the entire distribution of $Y_i(0)$, using $F_{Y|D=1}(y) = \mathbb{E}[\mathbb{1}(Y_i \leq y)|D_i = 1]$. That means that we can use random-assignment to uncover the effect of treatment on the entire *distribution* of outcomes. This lets us answer a new set of causal questions. For instance: what is the difference between the median value of $Y_i(1)$ and the median value of $Y_i(0)$? This is an example of a so-called *quantile-treatment effect*.

A natural question that you might hope to answer is: how many individuals in my population have a negative treatment effect $Y_i(1) < Y_i(0)$, versus a positive one? This is a harder type of question, because it depends on the joint distribution of potential outcomes. By contrast, random assignment (and similarly selection-on-observables, or quasi-experimental approaches), only let us identify each of the *marginal* distributions of $Y_i(0)$ and $Y_i(1)$, due to the fundamental problem of causal inference.

The situation is not completely hopeless: the marginal distributions of $Y_i(1)$ and $Y_i(0)$ do put some restrictions on the distribution of treatment effects. For instance, it can be shown that a lower bound on the proportion "harmed" by treatment $P(\Delta_i \leq 0)$ is the supremum of $F_{Y(1)}(y) - F_{Y(0)}(y)$ over all values of y (see e.g. Fan and Park, 2010 for details). We can also make additional assumptions that allow us to say more about the distribution of treatment effects. For example, the strong assumption of *rank-invariance* allows us to trace out the entire CDF of Δ_i, and in principle estimate the treatment effect for any given individual (see e.g. Heckman et al., 1997).

Chapter 7

Linear regression

Note on notation: In this section we'll simplify notation by dropping i subscripts when discussing population quantities. We'll add them back in Section 7.5 when we get to estimation. Remember that with *i.i.d.* data, it doesn't matter whether we include the i indices or not, because the distribution of variables in each observation i is the same as the population distribution.

7.1 Motivation from selection-on-observables

We saw in Chapter 6 that under the selection-on-observables assumption, and with a binary treatment variable, the average treatment effect conditional on $X = x$ can be calculated as:

$$ATE(x) = \mathbb{E}[Y(1) - Y(0)|X = x] = \mathbb{E}[Y|X = x, D = 1] - \mathbb{E}[Y|X = x, D = 0]$$

This requires having a way to estimate conditional expectations of the form $\mathbb{E}[Y|X = x, D = d]$ for $d = 0$ and $d = 1$. How should we do this?

If X is a discrete random variable, there is a pretty straightforward way we could do this. With i.i.d. data, a consistent estimator is simply the mean among the sub-sample of data for which $D_i = d$ and $X_i = x$:

$$\underbrace{\frac{1}{\#\text{ of observations } i \text{ for which } X_i = x \text{ and } D_i = d} \sum_{i:X_i=x\&D_i=d} Y_i}_{\hat{\mathbb{E}}[Y|X=x,D=d]} \quad \xrightarrow{p} \quad \mathbb{E}[Y|X = x, D = d]$$

But remember that for the selection-on-observables assumption, we want X to be an extensive-enough set of control variables to eliminate selection bias. So how should we proceed $X = (X_1, X_2, \dots X_k)$ is a vector of several random variables, some of which may be continuously distributed?

This is actually a hard problem, in practice. Recall that $\mathbb{E}[Y|X = x, D = d]$ is a function of x and d, which in the notation of 1.5.3 we might write as:

$$\mathbb{E}[Y|X = x, D = d] = m(d, x_1, x_2, \dots x_k)$$

where $x_1, x_2, \dots x_k$ are the components of the vector x. Provided that (Y, D, X) are all observed, the function m will be *identified* (see Section 5.2). That is, for fixed values (x, d) there is only one value of $m(d, x_1, x_2, \dots x_k)$ compatible with the joint distribution of our observables.

However, estimation is another thing. Given our finite sample, how do we uncover the function $m(d, x_1, x_2, \dots x_k)$? This turns out to be particularly straightforward when the function m is *linear*, that is:

$$m(d, x_1, x_2, \dots x_k) = \beta_0 + \beta_D d + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{7.1}$$

for some set of coefficients $(\beta_D, \beta_0, \beta_1, \dots \beta_k)$. In this case note that our parameter of interest, $ATE(x)$ is simply equal to $m(1, x) - m(0, x) = \beta_D$. Since this difference yields the same fixed number β_D regardless of x, the conditional-on-X ATE is the same as the overall average treatment effect, so $ATE = \beta_D$.

7.2 The linear regression model

Given a random variable Y and a random vector X, the *linear-regression model* says that

$$Y = X'\beta + \epsilon \tag{7.2}$$

where

$$\mathbb{E}[\epsilon|X] = 0 \tag{7.3}$$

We'll refer to the vector β appearing in Eq. (7.2) the *coefficient vector* from a regression of Y on X (as a reminder of notation: $\beta'X = \sum_j \beta_j \cdot X_j$). The term ϵ is often called an *error term* or *residual.*[1]

Remember from Section 3.1 that a statistical *model* places some kind of restriction on the joint distribution of random variables. The key restriction of the linear regression model is that the residual ϵ has a conditional mean of zero given any realization of X. The linear regression model holds for some β if and only if the conditional expectation function of Y on X is a linear function of X, that is:

$$\mathbb{E}[Y|X] = X'\beta \tag{7.4}$$

In almost all cases in which we use the linear regression model, one of the components of X is taken to be non-random and simply equal to one. It thus contributes a constant to the function $X'\beta$, for example:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \cdots + \beta_k \cdot X_k + \epsilon \tag{7.5}$$

where here we have started the numbering at 0, so that β has $k+1$ components. In this notation X also has $k+1$ components: $X = (1, X_1, \ldots X_k)'$. However, to keep notation compact, we'll often ignore the distinction between a constant and random elements in X.

Accordingly, if we let k be the total number of components in $X = (X_1, X_2, X_3 \ldots X_k)'$ (including any constant term), then notice that Eq. (7.3) implies the following k equations:

$$\mathbb{E}[X\epsilon] = \begin{pmatrix} \mathbb{E}[X_1 \cdot \epsilon] \\ \mathbb{E}[X_2 \cdot \epsilon] \\ \vdots \\ \mathbb{E}[X_k \cdot \epsilon] \end{pmatrix} = \begin{pmatrix} \mathbb{E}[X_1 \cdot (Y - X'\beta)] \\ \mathbb{E}[X_2 \cdot (Y - X'\beta)] \\ \vdots \\ \mathbb{E}[X_k \cdot (Y - X'\beta)] \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{7.6}$$

To see that $\mathbb{E}[\epsilon|X] = 0$ implies $\mathbb{E}[\epsilon \cdot X_j] = 0$ for any $j = 1 \ldots k$, use the law of iterated expectations:

$$\mathbb{E}[\epsilon \cdot X_j] = \mathbb{E}\left\{\mathbb{E}[\epsilon \cdot X_j |\, X]\right\} = \mathbb{E}\left\{\mathbb{E}[\epsilon|X] \cdot X_j\right\} = \mathbb{E}\left\{0 \cdot X_j\right\} = 0$$

It's probably a good idea to stare at this and make sure it makes sense. Conditional on any value $X = x$, the component X_j has some fixed value x_j. Thus, we can pull it out of the inner expectation, so that $\mathbb{E}[\epsilon \cdot X_j |\, X = x] = \mathbb{E}[\epsilon|X = x] \cdot x_j$. Then we take the outer expectation (curly braces) over values x.

Since (7.6) provides a system of k equations in the k unknowns $\beta_1 \ldots \beta_k$, it generally has a unique solution. A general expression for this solution is:

$$\beta = \mathbb{E}[XX']^{-1}\mathbb{E}[X \cdot Y] \tag{7.7}$$

We'll unpack the matrix notation of this equation later, so don't worry if it's not familiar to you right now. The important thing is that there is typically a single vector β that can satisfy all k lines of Eq. (7.6), and it is given by (7.7) above.

When people talk about "running a regression", the quantity they are estimating is (7.7), whether or not the conditional expectation function $\mathbb{E}[Y|X]$ is linear in X as the linear regression model assumes. Thus, rather than Eqs. (7.2) and (7.3) we could have gotten away with introducing β with a so-called *linear projection model*, which just says that

$$Y = X'\beta + \epsilon \qquad \text{where} \qquad \mathbb{E}[\epsilon \cdot X_j] = 0 \text{ for all } j = 1 \ldots k \tag{7.8}$$

Whether one starts from Eq. (7.4) or from (7.8), we're talking about the same β. We'll call this β, which has the explicit formula (7.7), the *coefficient vector* or the *linear regression vector.*

[1] The Hansen textbook reserves the term "residual" for an *estimated* value of ϵ that arises in the context of the ordinary least squares estimator. I'll refer to ϵ above as a residual, and what Hansen calls a residual a "fitted residual" in Sec. 7.5.

We can also write the linear regression vector in a second way: it minimizes the population mean-squared error between Y and a linear function of the components of X:

$$\beta = \underset{\gamma \in \mathbb{R}^k}{\operatorname{argmin}} \; \mathbb{E}[(Y - X'\gamma)^2] \tag{7.9}$$

This says that the value of the β appearing in Eq. (7.2) is exactly the one that minimizes the expectation of the squared difference between Y and the "regression line" $X'\beta$ implied by β and X. We'll establish Eq. (7.9) in Section 7.4.1.

Exercise: Using the fact that $\epsilon = Y - X'\beta$, show that Equation (7.6) is the first-order-condition of the minimization problem (7.9).

Note: to connect the linear regression model to the discussion of selection-on-observables in Section 7.1, we simply incorporate D as well as the constant in Eq. (7.1) into the vector X.

There are several motivations for caring about the linear regression vector β. linear regression will be a useful tool when the RHS of Eq. (7.7) answers or sheds light on some interesting question about the world. We'll now turn to several ways in which it can.

7.3 Five reasons to use the linear regression model

Here are several distinct motivations for using the linear regression model above. In all cases except the last one, we have the same parameter of interest: β, and we'll end up using the exact same estimator $\hat{\beta}$—referred to as the *ordinary-least-squares* or OLS estimator—for β.

7.3.1 As a structural model of the world

One way to arrive at the linear regression model is to simply assume that it describes a function that generates the outcome Y. For example, we might have a so-called *Mincer equation*, which explains log-wages as a function of education and job experience:

$$\log(wage_i) = \beta_0 + \beta_1 \cdot school_i + \beta_2 \cdot exp_i + \epsilon_i \tag{7.10}$$

where $wage_i$ is worker $i's$ hourly wage, $school_i$ is the number of years of schooling they completed, and exp_i is the their number of years of work experience. Here the $\beta = (\beta_0, \beta_1, \beta_2)'$ have a direct economic interpretation: we think of β_1 for example as telling us how much log wages would increase if student i's schooling was increased by one year.

In this structural model of the world, we can think of the residual ϵ_i as capturing everything else about i (or their employer) that also determines their wage: for example, their unobserved skills or "ability". If we are willing to believe that this ϵ_i is uncorrelated with schooling or experience, that is $\mathbb{E}[\epsilon_i \cdot school_i] = \mathbb{E}[\epsilon_i \cdot exp_i] = \mathbb{E}[\epsilon_i] = 0$, then we arrive at the linear projection model of Equations 7.2 and Equation 7.6.

When we interpret the regression equation as a story about how the world works, we give a particular interpretation to the vector β and to the error term ϵ: Eq. (7.10) constitutes an economic model and not just a statistical one. On this view, each expression $\mathbb{E}[\epsilon \cdot X_j] = 0$ is an assumption. Are unobservable factors like i's unobserved skills uncorrelated with schooling? If the answer is yes, then the parameters $\beta = (\beta_0, \beta_1, \beta_2)'$ can be estimated by the statistical technique of running a regression.

The above might be the way you were taught to think about regression in an undergraduate econometrics course. The next four sections will introduce motivations for using the linear regression model that have a different form. Instead of treating Equations 7.6 as an assumption, we use them to *define* β. While it may or may not have a direct economic interpretation like in the Mincer equation, β is always related to the conditional expectation function $\mathbb{E}[Y|X]$, which has a statistical interpretation. We turn now to this.

7.3.2 As an approximation to the conditional expectation function

As we've seen, Equations 7.2 and 7.6 are implied when the conditional expectation function (CEF) of Y on X is linear in X. That is, when

$$\mathbb{E}[Y|X] = X'\beta \tag{7.11}$$

Another way of saying this is that the function $m(x) := \mathbb{E}[Y|X = x]$ has the form

$$m(x) = x'\beta = \sum_{j=1}^{k} \beta_j \cdot x_j$$

Note that in the linear regression model, the residual ϵ is equal to the deviation of the random variable Y from its expectation given X; that is:

$$\epsilon := Y - \mathbb{E}[Y|X]$$

Then, by definition $Y = \mathbb{E}[Y|X] + \epsilon$. When the CEF takes the linear form of Equation 7.11, we get Equation 7.2. This is useful for example when we have selection-on-observables and are interested in the coefficient on treatment D, as in 7.1.

However, even if the assumption of Equation (7.4) that the CEF is linear in X is false, the function $X'\beta$ will still provide the *best linear approximation* to the true function $m(x) = \mathbb{E}[Y|X = x]$, in the sense of minimizing the mean squared approximation error:

Proposition 7.1. *$\beta = \underset{\gamma \in \mathbb{R}^k}{argmin}\ \mathbb{E}[(m(X) - X'\gamma)^2]$, where β is as defined in Equation 7.7.*

This means that even if the CEF is not quite linear, the linear projection coefficient β is the $k-$component vector such that $x'\beta$ best approximates $m(x)$ as a linear function.

7.3.3 As a way to summarize variation*

Linear regression is also often used as a descriptive tool to characterize the variation in one's data. Firstly, one can interpret Eq. (7.2) as decomposing the random variable Y into a term $X'\beta$ that is "explained" by X, and a residual ϵ that is not. One should be careful here: we're not saying that X causes Y, just that the two are correlated (see below). We'll see that the linear regression model also gives us a nice expression for $Var(Y)$, decomposing the variance of Y into a term that depends on the variance of X, and a second term that depends on the variance of the error. This can be helpful in establishing how much of the variance of Y can be "explained" by X.

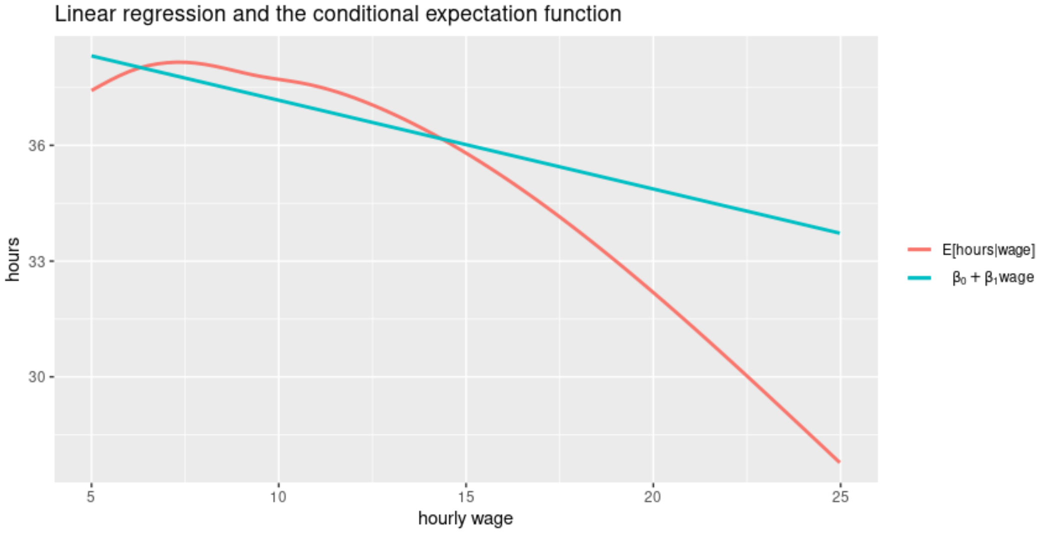
We previously defined the correlation coefficient $\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$ between two random variables. In Section 1.6 we generalized the notion of covariance to random vectors, but how does the notion of a correlation coefficient $\rho \in [-1, 1]$ extend to vectors? The regression coefficient vector β turns out to involve a version of this. The j^{th} component of the regression vector: β_j, can be expressed in terms of the so-called "partial-correlation" coefficient between Y and X_j, given the other variables in the regression (i.e. $X_1 \dots X_{j-1}$ and $X_{j+1} \dots X_k$). In general, the partial correlation between X and Y, given Z, is a version of ρ_{XY} that considers the residual variation in X and Y *after* accounting for the Z in a particular way. We'll study this in Section 7.4. This motivation is not entirely distinct from that of the last section, since correlations and partial correlations can always be written in terms of the CEF.

7.3.4 As a weighted average of something we care about*

Sometimes we start with a research question that has an complicated answer, but linear regression gives us a simple summary of that answer.

7.3.4.1 Example 1: the average derivative of a CEF*

Suppose I asked you to use the `nlswork` dataset from Chapter 2.2 to plot the conditional expectation function of a worker's weekly hours with respect to their hourly wage. Below I've computed this function for wages between \$5 an hour and \$25 and hour, displayed in red. As you can see, the function is non-monotonic: $\mathbb{E}[hours|wage = w]$ increases until w is about 8, then declines with w (quite steeply after $w = 15$).



In teal, I've plotted the regression function $\beta_0 - \beta_1 \cdot w$. We know from Proposition 7.1 that the vector (β_0, β_1) finds the best linear approximation to the conditional expectation function, which happens to be highly non-linear. That is: the slope of $\mathbb{E}[hours|wage = w]$ changes a lot with w. Although the linear approximation is not a particularly good one (the red and teal lines are far from one another for many values of w), we can see that the slope coefficient β_1 appears to "average out" the slope of $\mathbb{E}[hours|wage = w]$: sometimes the teal line is steeper, and sometimes the red line is steeper.

This is no accident: in a regression with one continuously-distributed X variable and a constant, the coefficient on X always captures a weighted-average of the derivative of the CEF with respect to X. In particular, $\beta_1 = \int m'(x) \cdot w(x) \cdot dx$, where $m'(x) = \frac{d}{dx}\mathbb{E}[Y|X = x]$, and $w(x)$ is a positive function that integrates to one. The weighting function $w(x)$ is proportional to $F(x)\left(\mathbb{E}[X] - \mathbb{E}[X|X \le x]\right)$, where $F(x)$ is the CDF of X. This result was originally derived by Yitzhaki (1996).

This result is nice, but be careful when appealing to it to say that a linear regression coefficient always captures an average CEF derivative. When there is more than one variable in the regression, say X_1 and X_2, the weights that β_1 places on $\frac{\partial}{\partial x_1}m(x_1, x_2)$ could be negative, if $\mathbb{E}[X_1|X_2]$ is not linear in X_2. Even with a single regressor, the averaging introduced by the linear regression vector can be misleading. If the CEF is sometimes increasing, and sometimes decreasing, we may get a regression coefficient of zero even in cases where X and Y are closely related.

Proof: This is jumping ahead a bit, but I include a proof here in case you are interested. To see this, note that by the law of iterated expectations, the slope coeffient is:

$$\beta_1 = \frac{Cov(X,Y)}{Var(X)} = \frac{1}{Var(X)} \cdot \mathbb{E}[Y \cdot X] - \mathbb{E}[X]\mathbb{E}[Y] = \frac{1}{Var(X)} \cdot \mathbb{E}[Y \cdot (X - \mathbb{E}[X])]$$
$$= \frac{1}{Var(X)} \cdot \int f(x) \cdot (x - \mathbb{E}[X]) \cdot \mathbb{E}[Y|X = x] = \frac{1}{Var(X)} \cdot \int f(x) \cdot (x - \mathbb{E}[X]) \cdot m(x)$$

where $m(x) = \mathbb{E}[Y|X = x]$ and $f(x)$ is the density of X. Now, we use the method of *integrattion by parts* with $u = f(x)(x - \mathbb{E}[X])$ and $dv = m(x) \cdot dx$ to write the integral as $v(x)g(x)|_{-\infty}^{\infty} - \int m'(x)v(x)dx$ where $v(x) = \int_{-\infty}^{x} f(t)(t - \mathbb{E}[X])$. The first term is zero because both $v(\infty)$ and $v(-\infty)$ are equal to zero (for $v(\infty)$ we assume X has a finite second moment). So, $\beta_1 = \int m'(x)w(x)dx$ where $w(x) = -v(x)/Var(X)$. To see that $w(x)$ integrates to one, substitute $Y = X$ in which case $m'(x) = 1$. To see that the $w(x) \ge 0$, rewrite $v(x) = F(x)\mathbb{E}[X|X \le x] - \mathbb{E}[X]F(x) = F(x)\left(\mathbb{E}[X|X \le x] - \mathbb{E}[X]\right)$ and note that $\mathbb{E}[X|X \le x] \le \mathbb{E}[X]$ for all x.

7.3.4.2 Example 2: weighted averages of treatment effects*

Consider a regression of Y on a binary treatment variable D and X:

$$Y = \beta_D \cdot D + \beta_X' X + \epsilon \tag{7.12}$$

where the vector X is a set of indicator variables for an underlying categorical variable G. By this, I mean that $X = (\mathbb{1}(G = 1), \mathbb{1}(G = 2), \dots, \mathbb{1}(G = N_g))'$, where $P(G \in \{1, 2, \dots N_G\}) = 1$. We'll return to this kind of regression later, which is sometimes referred to as "saturated". It turns out that the coefficient on D in this regression can be written as:

$$\beta_D = \frac{\mathbb{E}[\{\mathbb{E}[Y|D = 1, X] - \mathbb{E}[Y|D = 0, X]\} \cdot Var(D|X)]}{\mathbb{E}[Var(D|X)]}$$

If we assume selection-on-observables, then we know that the term in brackets is equal to $ATE(x) = E[Y(1) - Y(0)|X = x]$. Then, we have that $\beta_D = \sum_{j=1}^{N_G} w_j \cdot ATE(x_j)$ where$x_1, x_2, \dots$ are the values that X can take, i.e. x_j is a vector of N_G components composed of all zeros but a 1 in the j^{th} component. The $w_j = \frac{Var(D|X) \cdot P(X_i = x_j)}{E[Var(D|X)]}$ can be thought of as weights: they are positive and sum to one. This result can be found in Angrist and Pischke (2008), and we'll prove it later.

Thus, Eq. (7.12) recovers a weighted average of the conditional-on-X average treatment effects. It weights each group in proportion to the conditional variance of D given X. Intuitively, this puts more weights on the groups for which there is a more equal proportion of treatment and control units (since note that $Var(D|X) = P(D = 1|X) \cdot P(D = 0|X)$. It does not recover the average treatment effect $ATE = \mathbb{E}[Y(1) - Y(0)]$, which can be thought of as applying weights $w_j = P(X = x_j)$ (by the law of iterated expectations).

Recall that by contrast, we can always with a binary treatment get the ATE under selection-on-observables by estimating $\mathbb{E}[Y|D = 1, X = x] - \mathbb{E}[Y|D = 0, X = x]$, and we don't need to assume that X_i has a group structure for this result. What then is the value of the result in this section? In Section 7.2, we assumed that $\mathbb{E}[Y|D = 1, X = x]$ had the linear form of Eq. (7.1). This is a strong assumption; it implies for example that $ATE(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$ is the same for all x. The result in this section shows that when X as a group structure, we can still keep X and D separate in the regression, without assuming that $ATE(x)$ is constant in x (i.e. we can get away without including *interaction terms* between X and D).

7.3.5 As a tool for prediction*

One final way one might arrive at a linear regression model is that regression can be a tool for *prediction*. This is conceptually quite distinct from the other motivations discussed thus far, because in this case the goal is not to learn the value of β, but rather to use β to predict the outcome Y.

For example, suppose you are Netflix and want to decide whether to advertise the epic romance series *Econometrics* to a given consumer. In particular, you'd like to know whether i will click on the promotional display, and begin watching the (riveting) first episode. Let's indicate this by $Y = 1$. You know several things X about the consumer, for example their age and other shows that they have been watching. You don't know Y for this particular consumer (because you haven't shown them the ad yet), but you do know (Y_i, X_i) for a random sample of other individuals who did see the advertisement.

Now consider solving the above prediction problem for a randomly drawn consumer, where we attempt to use a linear function of X to predict Y. Because of Eq. (7.9), β will minimize the average such prediction error (squared). That is: $\beta = \text{argmin}_{\gamma \in \mathbb{R}^k} \; \mathbb{E}[(Y_i - X_i'\gamma)^2]$. Linear regression is not necessarily a particularly good tool for prediction, but can provide a starting point. Modern tools such as *machine-learning* algorithms are better-optimized for prediction.

7.4 Understanding the population regression coefficient vector

Let us now see the relationship between the β that satisfies the linear regression model, and the problem of minimizing the mean-squared error between Y and $X'\beta$. We repeat here Eq. (7.9):

$$\beta = \underset{\gamma \in \mathbb{R}^k}{\text{argmin}} \; \mathbb{E}[(Y - X'\gamma)^2] \tag{7.13}$$

Note that we are not constraining the values that γ can take in this minimization problem, rather we have an unconstrained minimization in which we search over *all* $\gamma \in \mathbb{R}^k$. That means that to minimize the mean squared error, it must satisfy the following k first-order-conditions (FOCs), one for each of its components β_j for $j = 1 \dots k$:

$$\frac{\partial \mathbb{E}[(Y - X'\beta)^2]}{\partial \beta_j} = \mathbb{E}[2(Y - X'\beta) \cdot X_j] = 0 \tag{7.14}$$

where we've used that $X'\beta = \sum_{j=1}^k X_j \cdot \beta_j$. This is equivalent to $\mathbb{E}[X_j \cdot \epsilon] = 0$, if we define $\epsilon = Y - X'\beta$. This leads exactly to the linear regression model of Equations 7.2 and 7.6.

Thus we've seen that the minimizer of the mean squared error between Y and a linear function of X must be equal to the regression coefficient vector β. The box at the end of Section 7.4.1 shows that this also goes in the other direction: the β defined by Equations 7.2 and 7.6 must be the β that solves (7.9).

Note: I've assumed in the above that $\mathbb{E}[(Y - X'\gamma)^2]$ is differentiable with respect to γ and that we can interchange the derivative and the expectation (this requires regularity conditions that allow us to appeal to the dominated convergence theorem, but we don't need to worry about these technicalities here).

7.4.1 Existence and uniqueness of β: no perfect multicollinearity

Is there always a β that minimizes the mean squared error, and could there be multiple values β and β' that both minimize it? These are the questions of *existence* and *uniqueness* of β, respectively.

A sufficient condition for β to exist and for it to be unique turns out to be that the matrix $\mathbb{E}[XX']$ is *invertible*, meaning that the inverse matrix $\mathbb{E}[XX']^{-1}$ exists. A convenient characterization of when $\mathbb{E}[X'X]$ will be invertible is given by the following proposition:

Proposition 7.2. *The matrix $\mathbb{E}[XX']$ has an inverse $\mathbb{E}[XX']^{-1}$, if and only if for all $\gamma \in \mathbb{R}^k$:*

$$P(X'\gamma \neq 0) > 0$$

Proof. We call a symmetric $k \times k$ matrix M *positive definite* if $\gamma' M \gamma > 0$. Any positive definite matrix is invertible. I won't prove this here, but for the curious: the eigenvalues of a positive definite matrix are strictly positive, and a matrix is invertible if and only if it does not have zero as an eigenvalue.

Thus, we'll show that $H = \mathbb{E}[XX']$ is positive definite in order to establish that it is invertible. For any $\gamma \in \mathbb{R}^k$, note that $\gamma' H \gamma = \mathbb{E}[\gamma' X X' \gamma] = \mathbb{E}[||X\gamma||^2]$, where for any vector $x \in \mathbb{R}^k$ we let $||x||^2$ denote it's Euclidean norm $||x||^2 = x'x = \sum_{j=1}^k (x_j)^2$. Since $||X'\gamma||^2 \geq 0$ for any realization of X, it follows that $\gamma' H \gamma > 0$ if and only if with some positive probability, $||X'\gamma||^2$ is strictly greater than zero. This occurs whenever $P(X'\gamma \neq 0) > 0$. □

Proposition 7.2 says that there exists no value γ that makes $X'\gamma$ equal to the zero vector, with probability one (remember that X here is a random vector). When there is such a γ, we say that there is *perfect multicollinearity* among our regressions $X = (X_1, X_2, \dots X_k)$.

Definition. *We say that there is* ***perfect multicollinearity*** *among our regressors (in the population) if there exists some $\gamma \in \mathbb{R}^k$ such that $P(X'\gamma = 0) = 1$.*

Example: Suppose that our regression includes a constant $X_1 = 1$, a binary variable indicating that a given individual is married: $X_2 = \textit{married}$, and a second binary variable X_3 that indicates that a given individual is not married. Then, since $X = (1, \textit{married}, 1 - \textit{married})'$, we have that $X'(-1, 1, 1) = 0$ for all realizations of X. Thus, we have perfect multicollinearity: $X'\gamma = 0$ regardless of the value of *married* and hence with probability one.

Note: Since we're talking about the population in this section (rather than a sample), definition 7.4.1 says that there is no perfect multicollinearity in the population. In Section 7.5, we'll use a sample analog of this definition, which will be required for us to define the OLS estimator of β.

Now let's see how the absence of perfect multicollinearity, which by Proposition 7.2 we know is equivalent to $\mathbb{E}[XX']$ being invertible, implies that β exists and is unique.

Proposition 7.3. *If there is no perfect multicollinearity, then there exists a $\beta \in \mathbb{R}^k$ that satisfies Equations 7.2 and 7.6, and it is unique.*

Proof. We can combine Equations 7.2 and 7.6 into a single matrix equation, which is equivalent to the system of FOCs 7.14:

$$\mathbb{E}[X(Y - X'\beta)] = \mathbf{0}$$

where $\mathbf{0}$ denotes a vector of k zeroes $\mathbf{0} = (0, 0, \dots 0)'$. This equation is the same as

$$\mathbb{E}[XX']\beta = \mathbb{E}[XY]$$

Since $\mathbb{E}[XX']$ is invertible by Proposition 7.2, we can multiply both sides of the above equation by $\mathbb{E}[XX']^{-1}$ (see box below) to obtain:

$$\beta = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$$

Of course, we can only do this if the matrix $\mathbb{E}[XX']^{-1}$ exists, which is guaranteed by no perfect multicollinearity. Note that the above expression for β was given previously in Equation 7.7.

□

Review: using matrix inverses to solve a system of linear equations

Suppose we have a system of k equations in k variables

$$\begin{aligned}
a_{11} \cdot x_1 + a_{21} \cdot x_2 + \cdots + a_{k1} \cdot x_k &= b_1 \\
a_{12} \cdot x_1 + a_{22} \cdot x_2 + \cdots + a_{k2} \cdot x_k &= b_1 \\
&\vdots \\
a_{1n} \cdot x_1 + a_{2n} \cdot x_2 + \cdots + a_{kk} \cdot x_k &= b_k
\end{aligned} \tag{7.15}$$

We seek a solution $\mathbf{x} = (x_1, x_2, \dots x_n)$ that satisfies all of the above equations. Let us gather all of the coefficients in to a $k \times k$ matrix and call it $\mathbf{A}$:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{21} & a_{22} & \dots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n1} & \dots & a_{nk} \end{pmatrix}$$

Our system of Equations (7.15) says, in vector notation, that $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{b} = (b_1, b_2, \dots b_k)'$ is a vector composed of the values appearing on the RHS in Eq. (7.15).

If the matrix $\mathbf{A}$ is *invertible*, this means that there exists a unique matrix $\mathbf{A}^{-1}$ such that $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_k$, where $\mathbf{I}_k$ is the $k \times k$ *identity matrix*. It has entries of one along the diagonal and zeros everywhere else:

$$\mathbf{I}_n = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

Note that the identity matrix $\mathbf{I}_k$ has the property that $\mathbf{I}_k\boldsymbol{\lambda} = \boldsymbol{\lambda}$ for any vector $\boldsymbol{\lambda} \in \mathbb{R}^n$.

Thus, if we start with the equation $\mathbf{Ax} = \mathbf{b}$ and multiply both sides by $\mathbf{A}^{-1}$, we get that

$$\mathbf{A}^{-1}(\mathbf{Ax}) = (\mathbf{A}^{-1}\mathbf{A})\mathbf{x} = \mathbf{I}_k\mathbf{x} = \mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

Thus, we've shown that $\mathbf{x}$ must be equal to $\mathbf{A}^{-1}\mathbf{b}$. This value definately satisfies (7.15), which we can verify by:

$$\mathbf{A}(\mathbf{A}^{-1}\mathbf{b}) = (\mathbf{AA}^{-1})\mathbf{b} = \mathbf{I}_k\mathbf{b} = \mathbf{b}$$

Also, it is the *only* value of $\mathbf{x}$ that satisfies the system (7.15). The solution exists and is unique, provided that $\mathbf{A}^{-1}$ exists.

Furthermore, one can show that the $\mathbf{x}$ solving $\mathbf{Ax} = \mathbf{b}$ is unique *only if* $\mathbf{A}$ is invertible. $\mathbf{A}$ is invertible if and only if there exists no $\boldsymbol{\lambda} \in \mathbb{R}^k$ that differs from the zero vector (i.e. it is not all zeros), for which $\mathbf{A}\boldsymbol{\lambda} = \mathbf{0}$ (here $\mathbf{0}_k$ is a vector composed of k zeros). Thus if $\mathbf{A}$ is not invertible, there is such a vector $\boldsymbol{\lambda}$. Suppose we have one solution $\mathbf{x}$ to $\mathbf{Ax} = \mathbf{b}$. Then $\mathbf{x} + \alpha\boldsymbol{\lambda}$ is another solution, for any value of α, because $\mathbf{A}(\mathbf{x} + \alpha\boldsymbol{\lambda}) = \mathbf{Ax} + \alpha\mathbf{A}\boldsymbol{\lambda} = \mathbf{b} + \mathbf{0}_k = \mathbf{b}ss$.

We are now also in a position to see why when $\mathbb{E}[X'X]^{-1}$ exists, the regression coefficient vector β *must* minimize the mean-squared error in Equation 7.13. Since $\mathbb{E}[(Y - X'\gamma)^2]$ is a convex function of the vector $\gamma = (\gamma_1, \gamma_2, \dots \gamma_k)$. That implies that any local minimum of $\mathbb{E}[(Y - X'\gamma)^2]$ is also global minimum. Therefore, we'd like to find any values of γ that might represent local minima of the mean-squared error. Sufficient conditions for γ to be a local minimum of the MSE are that: a) γ satisfies the FOCs 7.14 for each $j = 1 \dots k$; and b) the matrix H composed of components $H_{j\ell} = \frac{\partial^2}{\partial\gamma_j \partial\gamma_\ell}\mathbb{E}[(Y - X'\gamma)^2]$ is positive definite. The $k \times k$ matrix H is called the *Hessian* and it represents all of the second derivatives of a function. In the case of the MSE function $\mathbb{E}[(Y - X'\gamma)^2]$, the Hessian matrix turns out to be equal to $\mathbb{E}[X'X]$.

7.4.2 Simple linear regression in terms of covariances

When we just have a single regressor and a constant, we call this *simple linear regression*:

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon \tag{7.16}$$

where X is a scalar. Note that this is really a $k = 2$ instance of regression, in which one regressor is a constant and the other is a random variable. In this case we can derive a simple expression for β_0 and β_1, which do not require matrix notation.

Note that (7.6) provides two equations:

$$\mathbb{E}[\epsilon] = 0 \tag{7.17}$$

$$\mathbb{E}[X \cdot \epsilon] = 0 \tag{7.18}$$

We can take expectations of both sides of 7.16 and use 7.17 to obtain:

$$\mathbb{E}[Y] = \beta_0 + \beta_1 \cdot \mathbb{E}[X] + \cancel{\mathbb{E}[\epsilon]}$$

This expression says that the *regression line* $Y = \beta_0 + \beta_1 \cdot X$ passes through the point $(\mathbb{E}[X], \mathbb{E}[Y])$. If we plug in the mean value of X, our "predicted value" of Y is the mean of Y. Re-arranging, we have that $\beta_0 = \mathbb{E}[Y] - \beta_1 \cdot \mathbb{E}[X]$.

To make use of 7.18, let us multiply both sides of 7.16 by X and then take expectations. We have:

$$\mathbb{E}[X \cdot Y] = \beta_0 \cdot \mathbb{E}[X] + \beta_1 \cdot \mathbb{E}[X^2] + \cancel{\mathbb{E}[X \cdot \epsilon]}$$

We've now derived two equations for the two unknowns β_0 and β_1. If we substitute $\beta_0 = \mathbb{E}[Y] - \beta_1 \cdot \mathbb{E}[X]$ into the second equation above, we get that

$$\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y] - \beta_1 \cdot \mathbb{E}[X]^2 + \beta_1 \cdot \mathbb{E}[X^2]$$

Using that $Cov(X, Y) = \mathbb{E}[X \cdot Y] - \mathbb{E}[X] \cdot \mathbb{E}[Y]$ and $Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, we arrive at a nice simple formula for β_1:

$$\beta_1 = \frac{Cov(X, Y)}{Var(X)} \tag{7.19}$$

Note that this also gives us an explicit expression for β_0, by substituting the above expression for β_1 into

$$\beta_0 = \mathbb{E}[Y] - \beta_1 \cdot \mathbb{E}[X] \tag{7.20}$$

Looking back at Equatios 7.19, we see that in the case of simple linear regression β_1 is nothing more than a rescaled version of the covariance between X and Y. When $Cov(X, Y)$ is positive, β_1 will be positive, and when X and Y are negatively correlated β_1 will be negative. Looking at Equation 7.20, we see that β_0 is simply whatever it needs to be so that when we plot the regression line $\beta_0 + \beta_1 \cdot x$, it passes through the point $(\mathbb{E}[X], \mathbb{E}[Y])$.

There is also a simpler way to derive Eq. (7.19), which is to start with Eq. 7.16 and take the covariance of both sides with X. Since the covariance operator is linear, we have that

$$Cov(X,Y) = \cancel{Cov(X,\beta_0)} + \beta_1 \cdot Cov(X,X) + \cancel{Cov(X,\epsilon)}$$

The first term on the RHS is zero, because β_0 is a constant, which has a zero covariance with anything. The last term is zero because $Cov(X,\epsilon) = \cancel{\mathbb{E}[X\cdot\epsilon]} - \mathbb{E}[X]\cdot\cancel{\mathbb{E}[\epsilon]}$, where the crossed out terms are zero by 7.17 and 7.18. Using that $Cov(X,X) = Var(X)$, we can rearrange to obtain Eq. (7.19).

Note that Equations 7.19 and 7.20 imply that we can write the residual from simple linear regression as

$$\epsilon = (Y - \mathbb{E}[Y]) - \frac{Cov(X,Y)}{Var(X)}(X - \mathbb{E}[X]) \tag{7.21}$$

Example: Consider a simple linear regression in which X is a binary variable, for example if $X = 1$ indicates that an individual is female and $X = 0$ otherwise. In this case, recall from the homework that $Cov(X,Y) = Var(X)\cdot(\mathbb{E}[Y|X=1] - \mathbb{E}[Y|X=0])$, where $Var(X) = P(X = 1)\cdot(1 - P(X=1))$. Thus, $\beta_1 = \mathbb{E}[Y|X=1] - \mathbb{E}[Y|X=0]$. This means that β_0 must be:

$$\begin{aligned}
\beta_0 &= \mathbb{E}[Y] - (\mathbb{E}[Y|X=1] - \mathbb{E}[Y|X=0])\cdot\mathbb{E}[X] \\
&= \{P(X=1)\cdot E[Y|X=1] + P(X=0)\cdot\mathbb{E}[Y|X=0]\} - (\mathbb{E}[Y|X=1] - \mathbb{E}[Y|X=0])\cdot\mathbb{E}[X] \\
&= (1 - P(X=1))\cdot\mathbb{E}[Y|X=0] + P(X=1)\cdot\mathbb{E}[Y|X=0] \\
&= \mathbb{E}[Y|X=0]
\end{aligned}$$

where we've used the law of iterated expectations and that $E[X] = P(X=1)$.

Thus, when we have a single binary regression, linear regression gives us an intercept $\beta_0 = \mathbb{E}[Y|X=0]$ that recovers the condition mean of Y given $X = 0$. Since the slope coefficient is $\beta_1 = \mathbb{E}[Y|X=1] - \mathbb{E}[Y|X=0]$, this means that when $X = 1$, the regression line passes through $\beta_0 + \beta_1 = \mathbb{E}[Y|X=1]$. In other words, $\beta_0 + \beta_1\cdot X$ exactly recovers the CEF function $E[Y|X]$. We'll see that this nice property generalizes in multiple linear regression: whenever X contains indicators for a complete set of groups (or a constant and all but one of the groups, then regression exactly captures the CEF).

Exercise: Given the above, convince yourself that in a regression of Y on a binary treatment variable D, the coefficient on D is the difference-in-means estimand θ_{DM} that we met in Chapter 6. Under random assignment, this regression hence yields the ATE.

Exercise: Derive Equations 7.19 and 7.20 from the general formula (7.7) for the regression vector, which in this case reads $\beta = (\beta_0,\beta_1) = \mathbb{E}[(1,X)'(1,X)]^{-1}\mathbb{E}[(1,X)'Y]$. For this you will need the formula for the inverse of a 2×2 matrix:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

A hint to get you started: $\mathbb{E}[(1,X)'Y] = (\mathbb{E}[X], \mathbb{E}[XY])'$.

7.4.3 Multiple linear regression in terms of covariances

Now let's consider how the considerations of the last section generalize to a setting in which we have multiple regressors and a constant. We know the general formula (7.7) for the vector β, which involves inverting the matrix $\mathbb{E}[XX']^{-1}$ and multiplying it by the vector $\mathbb{E}[XY]$. While the matrix formula holds generally, it turns out that we can still write expressions for the individual components of β in terms of covariances and variances, which is helpful in understanding the mechanics of how regression works.

Two regressors and a constant

Consider first a case in which we have two regressors X_1 and X_2, plus a constant:

$$Y = \beta_0 + \beta_1\cdot X_1 + \beta_2\cdot X_2 + \epsilon \tag{7.22}$$

Recall that in this case the linear regression model gives us three restrictions: $\mathbb{E}[\epsilon] = 0$, $\mathbb{E}[X_1 \cdot \epsilon] = 0$ and $\mathbb{E}[X_2 \cdot \epsilon] = 0$.

Consider β_2, the coefficient on X_2. In turns out that we can write an expression for β_2 by imagining a sequence of two simple linear regressions. In the first step, we imagine running a regression of X_2 on X_1 and a constant. We know that in this case, the coefficient on X_1 will be $Cov(X_1, X_2)/Var(X_1)$, the constant will be $\mathbb{E}[X_2] - Cov(X_1, X_2)/Var(X_1) \cdot \mathbb{E}[X_1]$, and by Equation (7.21) the residual will be

$$\tilde{X}_2 = (X_2 - \mathbb{E}[X_2]) - \frac{Cov(X_1, X_2)}{Var(X_1)}(X_1 - \mathbb{E}[X_1]) \tag{7.23}$$

where we use the notation $\tilde{X}_2$ for the residual from this regression of X_2 on X_1 and a constant. Observe that since $\tilde{X}_2$ is a linear function of X_1 and X_2, we know that it is uncorrelated with the error ϵ from the "long-regression" Equation (7.22).

Exercise: Prove the claim above, that $Cov(\tilde{X}_2, \epsilon) = 0$, using that $\mathbb{E}[\epsilon] = 0$, $\mathbb{E}[X_1 \cdot \epsilon] = 0$ and $\mathbb{E}[X_2 \cdot \epsilon] = 0$.

Given that $Cov(\tilde{X}_2, \epsilon) = 0$, consider running a second simple linear regression, in which we regress Y on $\tilde{X}_2$ and a constant.

Proposition 7.4. *The slope coefficient from this second regression is equal to* β_2*.*

To demonstrate that this claim is true, let's begin by substituting in Eq. (7.22), the equation for the slope in this second regression will be

$$\frac{Cov(\tilde{X}_2, Y)}{Var(\tilde{X}_2)} = \frac{\cancel{Cov(\beta_0, X_1)}}{Var(\tilde{X}_2)} + \beta_1 \cdot \frac{Cov(\tilde{X}_2, X_1)}{Var(\tilde{X}_2)} + \beta_2 \cdot \frac{Cov(\tilde{X}_2, X_2)}{Var(\tilde{X}_2)} + \frac{\cancel{Cov(\tilde{X}_2, \epsilon)}}{Var(\tilde{X}_2)} \tag{7.24}$$

where the first term is zero since β_0 is a constant, and the last term is zero because we've seen that $Cov(\tilde{X}_2, \epsilon) = 0$.

However, we *also* have that $Cov(\tilde{X}_2, X_1) = 0$, so the second term above also vanishes. How do we know this? Remember that $\tilde{X}_2$ is the residual from a regression of something on X_1 and a constant, so it is uncorrelated with X_1 by construction.

Exercise: Convince yourself of the claim above, that $Cov(\tilde{X}_2, X_2) = 0$. Why does it follow from how we've defined $\tilde{X}_2$ that $\mathbb{E}[\tilde{X}_2] = 0$ and $\mathbb{E}[X_1 \cdot \tilde{X}_2] = 0$?

The last step in establishing Proposition 7.4 is showing that $Cov(\tilde{X}_2, X_2) = Var(\tilde{X}_2)$. To see this, substitute our explicit expression (7.23) for $\tilde{X}_2$ into the variance. Since $\mathbb{E}[\tilde{X}_2] = 0$, $Var(\tilde{X}_2) = \mathbb{E}[(\tilde{X}_2)^2]$, which is equal to

$$\begin{aligned}\mathbb{E}\left[\left((X_2 - \mathbb{E}[X_2]) - \frac{Cov(X_1, X_2)}{Var(X_1)}(X_1 - \mathbb{E}[X_1])\right)^2\right] &= Var(X_2) - 2\frac{Cov(X_1, X_2)}{Var(X_1)} \cdot Cov(X_1, X_2) + \left(\frac{Cov(X_1, X_2)}{Var(X_1)}\right)^2 \cdot Var(X_1) \\ &= Var(X_2) - \frac{Cov(X_1, X_2)}{Var(X_1)} \cdot Cov(X_1, X_2) = Cov(\tilde{X}_2, X_2)\end{aligned}$$

A more direct way to conclude that $Cov(\tilde{X}_2, X_2) = Var(\tilde{X}_2)$ is to notice that this statement is equivalent to $Cov(\tilde{X}_2, (X_2 - \tilde{X}_2)) = 0$, where $(X_2 - \tilde{X}_2)$ is the regression line from a regression of X_2 on X_1 and a constant. That this is uncorrelated with the error $\tilde{X}_2$ from this same regression follows then by definition.

Since the labeling of X_1 and X_2 is arbitrary, Proposition 7.4 also gives us a way to express the coefficient β_1: first, run a regression of X_1 on X_2, and then regress Y on the residuals $\tilde{X}_1$ from this initial regression.

Many regressors and a constant

This principle generalizes to the general setting in which we have a regression equation with a constant and k additional regressors $X_1, X_2, \ldots X_k$:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots \beta_k \cdot X_k + \epsilon \tag{7.25}$$

Note first that since one of our regressors is a constant, the system of equations (7.6) implies that $\mathbb{E}[\epsilon] = 0$. Then the remainder of the equations in (7.6) can be read as saying that each X_j is *uncorrelated* with the error, since $Cov(X_j, \epsilon) = \cancel{\mathbb{E}[X_j \cdot Y]} - \cancel{\mathbb{E}[X_j]} \cdot \mathbb{E}[Y] = 0$.

Proposition 7.5 ("regression anatomy" formula). *The coefficient on X_j in regression (7.25) is*

$$\beta_j = Cov(\tilde{X}_j, Y)/Var(\tilde{X}_j),$$

where $\tilde{X}_j$ is the residual from a regression of X_j on all of the other regressors and a constant.

The text *Mostly Harmless Econometrics* refers to Proposition 7.5 as the "regression anatomy" formula because it allows us to translate the complicated expression for the full vector $\beta = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$ into a simpler expression for each of the components β_j.

Note: We'll see when we get to estimation in Section 7.5 that Proposition 7.5 has a sample analog, referred to as the *Frisch-Waugh-Lovell* theorem. Proposition 7.5 constitutes a "population version" of this very useful result.

Note: A Corollary to Proposition 7.5 is that we can also write β_j as $Cov(\tilde{X}_j, \tilde{Y}_j)/Var(\tilde{X}_j)$, where we define $\tilde{Y}_j$ to be the residual from a regression of Y on all the regressors aside from X_j, and a constant. This follows because the difference between $Y - \tilde{Y}_j$ is uncorrelated with $\tilde{X}_j$.

Using Proposition 7.5 iteratively to get regression coefficients: Notice that Proposition 7.5 gives us an explicit expression for β_j, if we know the residuals $\tilde{X}_j$. But if $k > 2$, how do we compute the $\tilde{X}_j$, which involve running a regression on $k - 1$ variables (e.g. $X_1, X_2, \dots X_{j-1}, X_{j+1}, \dots X_k$) and a constant?

If one would like to avoid the matrix expression for β, the answer is to appeal to Proposition 7.5 iteratively. This allows us to build up an expression for β_j by running a series of several simple linear regressions.

For instance, suppose we are interested in β_k in regression (7.25). We can obtain it as follows:

1. To get β_k, we need the residuals $\tilde{X}_j$ from a regression of X_k on $X_1 \dots X_{k-1}$ and a constant.
2. Call the coefficient on X_j from this regression β_j^k. If we know all the β_j^k for $j = 1 \dots_{k=1}$, we can calculate $\tilde{X}_j$ (note that we can pin down the β_0^k from the fact that $\mathbb{E}[\tilde{X}_j] = 0$.
3. By Proposition 7.5, we know that we can calculate each β_j^k if we have the residuals from a regression of X_j on all regressors except X_j and X_k, and a constant. Call the coefficient on X_ℓ from this regression β_ℓ^{jk}.
4. By Proposition 7.5, we know that we can obtain β_ℓ^{jk} if we know the coefficients from a regression of X_ℓ on a regression on all variables except X_ℓ, X_j and X_k, and a constant.
5. Continue on in this way, until we have a set of simple linear regressions, for which we can apply Equations 7.19 and 7.20.

The box above describes an iterative algorithm that would allow us to use Proposition 7.5 to obtain an expression for each coefficient β_j in Equation (7.25). As you can see, it is tedious, and involves running *many* simple linear regressions if our original regression contains more than a couple of variables.

That was a mess–lets use matrix notation!

For this reason, we inevitably need to appeal to the general formula $\beta = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$, which in the context of regression (7.25) says that:

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} = \mathbb{E}[(1, X_1, X_2, \dots X_k)(1, X_1, X_2, \dots X_k)']^{-1}\mathbb{E}[(1, X_1, X_2, \dots X_k)Y]$$

$$= \begin{pmatrix} 1 & \mathbb{E}[X_1] & \mathbb{E}[X_2] & \dots & \mathbb{E}[X_k] \\ \mathbb{E}[X_1] & \mathbb{E}[X_1^2] & \mathbb{E}[X_1 \cdot X_2] & \dots & \mathbb{E}[X_1 \cdot X_k] \\ \mathbb{E}[X_2] & \mathbb{E}[X_1 \cdot X_2] & \mathbb{E}[X_2^2] & \dots & \mathbb{E}[X_2 \cdot X_k] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[X_k] & \mathbb{E}[X_k \cdot X_1] & \mathbb{E}[X_k \cdot X_2] & \dots & \mathbb{E}[X_k^2] \end{pmatrix}^{-1} \begin{pmatrix} \mathbb{E}[Y] \\ \mathbb{E}[X_1 \cdot Y] \\ \mathbb{E}[X_2 \cdot Y] \\ \vdots \\ \mathbb{E}[X_k \cdot Y] \end{pmatrix} \tag{7.26}$$

Proposition 7.5 can be derived from Equation 7.26, but it's not exactly pretty. One way to make this work is to apply the *block matrix inversion* formula over and over again, mirroring the iterative approach in the box above. We would eventually end up with a large number of 2 × 2 matrix inverse problems, which are each easy to deal with (see the exercise at the end of Section 7.4.2). But life's too short for that! Thankfully computers area great help in computing our matrix inverses in practice.

7.5 The ordinary least squares (OLS) estimator

Now let's turn to *estimation* in the linear regression model. That is, suppose that based on the many motivations considered in Section 7.3, we're interested in estimating the vector β from the linear regression model. The standard estimator for β in the linear regression model is referred to as the *ordinary least squares* (OLS) estimator $\hat{\beta}_{OLS}$. Since this is the only estimator for β that we'll consider, we'll just write it as $\hat{\beta}$, to avoid writing OLS over and over again.

I break this section into several headings, for quick reference.

7.5.1 Sample

To define the OLS estimator $\hat{\beta}$ we suppose that we have a sample $(Y_i, X_{1i}, X_{2i}, \dots X_{ki})$ of Y and some set of regressors X_1 to X_k. Let n be the number of observations in our sample. *Note:* we will later assume that our sample is $i.i.d$, but we don't need to use that fact right now.

7.5.2 OLS estimator

A simple way to define the OLS estimator $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \dots \hat{\beta}_k)$ is as the minimizer of the sample analog of the least squares minimization, in which we replace the population expectation with the sample mean:

$$\hat{\beta} = \underset{\gamma \in \mathbb{R}^k}{\operatorname{argmin}} \ \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\gamma)^2 \tag{7.27}$$

Given the OLS estimator $\hat{\beta}$, let us make the following definitions:

- The *fitted value* $\hat{Y}_i$ for observation i is $\hat{Y}_i = X_i'\hat{\beta} = \sum_{j=1}^{k} \hat{\beta}_j \cdot X_{ji}$
- The *fitted residual* for observation i is $\hat{\epsilon}_i = Y_i - \hat{Y}_i$
- Note that for each i, we have that $Y_i = \hat{Y}_i + \hat{\epsilon}_i$ (by definition)

Equation (7.27) explains the origin of the name "ordinary least squares", as $\hat{\beta}$ is defined as the value of γ that minimizes the sample sum of squares.

What is the solution to the minimization problem (7.27)? Taking the first-order-condition with respect to each γ_j, we obtain the following system of equations:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} X_{1i}\cdot(Y_i - X_i'\hat{\beta}) &= \frac{1}{n}\sum_{i=1}^{n} X_{1i}\cdot\hat{\epsilon}_i = 0\\
\frac{1}{n}\sum_{i=1}^{n} X_{2i}\cdot(Y_i - X_i'\hat{\beta}) &= \frac{1}{n}\sum_{i=1}^{n} X_{2i}\cdot\hat{\epsilon}_i = 0\\
&\vdots\\
\frac{1}{n}\sum_{i=1}^{n} X_{ki}\cdot(Y_i - X_i'\hat{\beta}) &= \frac{1}{n}\sum_{i=1}^{n} X_{ki}\cdot\hat{\epsilon}_i = 0 \qquad (7.28)
\end{aligned}$$

which can be summarized by the matrix equation

$$\frac{1}{n}\sum_{i=1}^{n} X_i(Y_i - X_i'\hat{\beta}) = \mathbf{0}$$

$\mathbf{0}$ is a vector of k zeros. This is exactly analogous to Eq. (7.6), except that we have replaced the population expectations $\mathbb{E}$ with sample averages $\frac{1}{n}\sum_i$. Rearranging the above:

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_iX_i'\right)\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} Y_iX_i \qquad (7.29)$$

where recall that $X_i = (X_{1i}, X_{2i}, \dots X_k)'$ is a vector and Y_i is a scalar for each i. Since X_i is $k \times 1$ and X_i' is $1 \times k$, X_iX_i' is a $k \times k$ matrix. In Equation 7.29 we've used that by the distributive property of matrix multiplication, we can sum over the observations i *and then* multiply by β, which is equivalent to multiplying and then summing the $k \times 1$ vector $X_iX_i'\hat{\beta}$ over observations.

"Physics intuition": A nice way to visualize what OLS does (in the case of a single regressor pus a constant) is to imagine a set of n springs. One end of each spring is attached to a data point (Y_i, X_i), and the other end is attached to a rigid rod, which will represent the regression line $\hat{\beta}_0 + \hat{\beta}_1 X$. The springs want to be as short as possible, and are constrained to move only in the vertical direction. The length of spring i is equal to the fitted residual $\hat{\epsilon}_i$. An approximation known as *Hooke's Law* says that the potential energy stored in a spring is proportional to the square of its distance. Thus, if the springs are identical to one another, the total potential energy will be equal to the sum of squares of $\hat{\epsilon}_i$. Classical physics (i.e. "Newton's laws") tell us that equilibrium position of the rod will minimize the potential energy stored across all of the springs: exactly the minimization problem that OLS solves! The following link provides an illustration of this: it generates a random dataset, and shows the rod coming to rest in exactly the position of the OLS regression line: https://sam.zhang.fyi/html/fullscreen/springs/. Here is another nice visualization (not animated though): https://joshualoftus.com/posts/2020-11-23-least-squares-as-springs/

7.5.3 Matrix notation

We can obtain a more compact notation for Equation 7.29 by introducing an $n \times k$ matrix $\mathbf{X}$, that records all of our observations of all of the regressors:

$$\mathbf{X} := \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix} = \underbrace{\begin{pmatrix} (X_{11}, X_{21}, \dots X_{k1}) \\ (X_{12}, X_{22}, \dots X_{k2}) \\ \vdots \\ (X_{1n}, X_{2n}, \dots X_{kn}) \end{pmatrix}}_{k \text{ columns}} \left.\vphantom{\begin{pmatrix} X \\ X \\ \vdots \\ X \end{pmatrix}}\right\} n \text{ rows}$$

The matrix $\mathbf{X}$ is often called the *design matrix*.

Similiarly, we define a $k \times 1$ vector of our observations of Y:

$$\mathbf{Y} := \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

In this notation, we can rewrite the matrix $\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)$ as $\frac{1}{n}\mathbf{X}'\mathbf{X}$. We can then write 7.29 in the compact form:

$$(\mathbf{X}'\mathbf{X})\hat{\beta} = \mathbf{X}'\mathbf{Y} \tag{7.30}$$

where we've multiplied both sides by n.

Exercise: Use the definition of matrix multiplication to verify Equation (7.30) from Equation (7.29). That is, show that the j^{th} component of $\mathbf{X}'\mathbf{X}$ is equal to the j^{th} component of $\hat{\beta} = \left(\sum_{i=1}^{n} X_i X_i\right)\hat{\beta}$, and similarly that the j^{th} component of $\mathbf{X}'\mathbf{Y}$ is equal to $\sum_{i=1}^{n} Y_i X_{ji}$, for all $j = 1 \dots k$.

Note: For a general matrix M, it is conventional to let M_{ij} denote the entry on row i, column j. In our notation above, we actually have the opposite order of indices, because we've let X_{ji} denote the i^{th} observation of variable X_j, which is the entry in the j^{th} *column* and i^{th} *row* of $\mathbf{X}$. To avoid getting confused, always remain mindful that your matrix expressions are *conformable*, e.g. if you're multiplying $\mathbf{X}$ by a vector v to obtain $\mathbf{X}v$, then you know that v must have k components, since $\mathbf{X}$ has k columns.

7.5.4 Existence and uniqueness of $\hat{\beta}$: no perfect-multicollinearity in sample

When does the sample least-squares problem have a unique solution, such that the OLS estimator is well-defined? In exact analogy with the question of the existence and uniqueness of the population regression vector β, studied in Section 7.4.1.

For Equation 7.30 to have a unique solution, we need for the $k \times k$ matrix $\mathbf{X}'\mathbf{X}$ to be invertible (see the box in Section 7.4.1 for a review of solving a sytem of linear equations). The following proposition provides a characterization of when this will be the true:

Proposition 7.6. *Provided that $n > k$, the matrix $\mathbf{X}'\mathbf{X}$ is invertible if none of the columns of $\mathbf{X}$ can be written as linear combinations of the other. That is: $X'\gamma \neq \mathbf{0}$ for all $\gamma \in \mathbb{R}^k$.*

This condition can be referred to as no perfect multicollinearity *in the sample.* When it holds, we obtain an explicit expression for the OLS estimator $\hat{\beta}$:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{7.31}$$

Recall from Proposition 7.2 that we have no perfect multicollinearity in the population if $P(X_i'\gamma \neq \mathbf{0}) < 1$ for all $\gamma \in \mathbb{R}^k$. It's not impossible that this could hold, but that we still end up with perfect multicollinearity failing in the sample. Technically, this is a "knife-edge" case that would happen with probability zero if the X_i are continuously distributed. But in practice $\hat{\beta}$ will be defined but numerically unstable if $\mathbf{X}'\mathbf{X}$ is close to being non-invertible. This is what statistical software sometimes complains about when it says that $\mathbf{X}$ is "highly singular".

The most common case in which perfect multicollinearity occurs is when you forget that some of your regressors are related to one another definitionally. For instance, recall the example from Section 7.4.1, in which our regression includes a constant $X_{1i} = 1$, a binary variable indicating that a given individual i is married: $X_{2i} = married_i$, and a second binary variable X_{3i} that indicates that a given individual is not married. Then, since $X_i = (1, married_i, 1 - married_i)'$, we have that $X_i'(-1, 1, 1) = 0$ for each i in the sample, and hence $\mathbf{X}(-1, 1, 1)' = (0, 0, 0)'$. In this case we have perfect multicollinearity both in sample and in the population, since $P(X_i'(-1, 1, 1)) = 1$. If you tried to compute the OLS estimator for this regression in statistical software like Stata, it will choose arbitrarily one of the variables and drop it from the regression equation, and give you a warning.

7.5.5 More matrix notation

Following the notation we've developed to define the OLS estimator, we can also define $k \times 1$ vectors of the fitted values $\hat{Y}_i$, the fitted residuals $\hat{\epsilon}_i$, and the population residuals ϵ_i:

$$\hat{\mathbf{e}} := \begin{pmatrix} \hat{\epsilon}_1 \\ \hat{\epsilon}_2 \\ \vdots \\ \hat{\epsilon}_n \end{pmatrix} \qquad \hat{\mathbf{Y}} := \begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{pmatrix} \qquad \boldsymbol{\epsilon} := \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

While $\hat{\mathbf{e}}$ and $\hat{\mathbf{Y}}$ are built with estimates from the data, note that $\boldsymbol{\epsilon}$ is not observable. However, under the assumption that the regression model $Y_i = X_i'\beta + \epsilon_i$ holds for each i, we have that

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\epsilon} \tag{7.32}$$

Note that we can also write

$$\mathbf{Y} = \mathbf{X}\hat{\beta} + \hat{\mathbf{e}} \tag{7.33}$$

where $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$.

7.5.6 Regression as projection*

Regression thus provides a decomposition of the vector of observed outcomes, $\mathbf{Y}$, into two pieces: the vector $\mathbf{X}\hat{\beta}$ and the vector $\hat{\mathbf{e}}$. This decomposition has a geometric interpretation. Make the following definitions

1. Let $\mathbf{I}_n$ be the $n \times n$ *identity matrix* (see box in Section 7.4.1 for definition).
2. Define the $n \times n$ matrix $\mathbf{P} := (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Following Hansen, we'll call this the *projector matrix*.
3. Define the $n \times n$ matrix $\mathbf{M} := \mathbf{I}_n - \mathbf{P} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Following Hansen, we'll call this the *annihilator matrix*.

Exercise: Check that both $\mathbf{P}$ and $\mathbf{M}$ are *idempotent*. We say that an $n \times n$ matrix $\mathbf{A}$ is idempotent when $\mathbf{AA} = \mathbf{A}$.

Exercise: Show that both $\mathbf{P}$ and $\mathbf{M}$ are *symmetric*. We say that an $n \times n$ matrix $\mathbf{A}$ is symmetric when it is equal to its transpose: $\mathbf{A}' = \mathbf{A}$.

Exercise: Use the definitions of the projector and annihilator matrices to verify that

$$\mathbf{X}\hat{\beta} = \mathbf{PY} \qquad \text{and} \qquad \hat{\mathbf{e}} = \mathbf{MY}$$

Consider any vector v in $\mathbb{R}^n$. Since $\mathbf{I}_n v = v$ (check this if you haven't seen it before), and $\mathbf{P} + \mathbf{M} = \mathbf{I}_n$, it follows that any vector v can be written as $v = \mathbf{P}v + \mathbf{M}v$. Applying this to the vector $\mathbf{Y}$ composed of observations of our outcome variable, we obtain Equation (7.33).

Geometrically, Equation (7.33) provides a decomposition of $\mathbf{Y}$ into the part of $\mathbb{R}^n$ that is spanned by the columns of the design matrix $\mathbf{X}$ (this is $\mathbf{X}\hat{\beta}$), and the part that is orthogonal to it (this is $\hat{\mathbf{e}}$).

In what sense does $\mathbf{P}$ "project" the vector $\mathbf{Y}$? Algebraically, we think of projection as something that throws away some information in $\mathbf{Y}$ in such a way that after we "project" once, projecting again wouldn't change anything. For example, projecting a two-dimensional vector (v_1, v_2) onto the x-axis means replacing v_2 with zero, but keeping the first value: $(v_1, 0)$. If we then projected $(v_1, 0)$ onto the x-axis again, we'd just get $(v_1, 0)$ again. In this case projection corresponds to the matrix $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$.

This is also what happens with regression, except that our projection operator is somewhat more complicated. If we regress $\mathbf{Y}$ on $\mathbf{X}$, and then run the regression a second time using the fitted values $\hat{\mathbf{Y}}$ on the left hand side, this second regression will result in the exact same vector of estimate. This is because $\mathbf{PP} = \mathbf{P}$, and hence $\mathbf{P}\hat{\mathbf{Y}} = \mathbf{PP}\hat{\mathbf{Y}} = \mathbf{P}\hat{\mathbf{Y}} = \hat{\mathbf{Y}}$ (see exercise above). A similar argument holds for the fitted residuals, given that $\mathbf{MM} = \mathbf{M}$.

We call $\mathbf{M}$ the "annihilator" matrix because it is contructed in such a way that it "annihilates" the matrix $\mathbf{X}$ upon mulitplication:

$$\mathbf{MX} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{X} = \mathbf{I}_n\mathbf{X} - \mathbf{X}\cancel{(\mathbf{X}'\mathbf{X})^{-1}}\cancel{(\mathbf{X}'\mathbf{X})} = \mathbf{X} - \mathbf{X} = \mathbf{0} \tag{7.34}$$

By a similar argument, the projector matrix does nothing to $\mathbf{X}$: $\mathbf{PX} = \mathbf{X}$.

The diagonal elements of the projection matrix $\mathbf{P}$ have a useful interpretation, and are called *leverage values*. Let

$$h_{ii} := [\mathbf{P}]_{ii} = [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']_{ii} = X_i'(\mathbf{X}'\mathbf{X})^{-1}X_i$$

Since $\mathbf{P}$ is an $n \times n$ matrix, there exists a leverage value corresponding to each observation i. It measures how "extreme" the regressors X_i are for that observation. Note that $X_i'X_i$ is the norm of the vector X_i—how "big" it is. h_{ii} inserts the $k \times k$ matrix $(\mathbf{X}'\mathbf{X})^{-1}$ inside the dot-product, which changes what we mean by "big". In the case of simple linear regression, h_{ii} ends up measuring whether the value of X_i is particularly high or low. Such observations may have a large influence on the values of the OLS estimate $\hat{\beta}$, since a small change in $\hat{\beta}$ can translate into a large change in the fitted residual $\hat{\epsilon}_i$ for that observation, when X_i is very far from the sample mean.

The projection matrix notation also gives us a nice way to express the *coefficient of determination*, or R^2, which you may have met in previous econometrics classes. The coefficient of determination can be defined as the ratio of the variance of the fitted estimates $\hat{Y}_i$ across the sample to the variance of the observed outcomes Y_i across the sample. Thus it measures the proportion of the variation in Y that is "explained" by the OLS regression line $\hat{Y}_i = X_i'\hat{\beta}$.

For simplicity, suppose that the average value of Y_i in the sample is zero, so that the sample variance of Y_i is $\frac{1}{n}\mathbf{Y}'\mathbf{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$. Note that we could always de-mean our data to satisfy this assumption without affecting the variance of Y_i or R^2 (provided that the regression contains a constant, for the latter claim). In our matrix notation, we could then write R^2 as:

$$R^2 = \frac{\hat{\mathbf{Y}}'\hat{\mathbf{Y}}}{\mathbf{Y}'\mathbf{Y}} = \frac{(\mathbf{PY})'(\mathbf{PY})}{\mathbf{Y}'\mathbf{Y}} = \frac{\mathbf{Y}'\mathbf{PY}}{\mathbf{Y}'\mathbf{Y}}$$

The coefficient of determination measures how much the Euclidean norm of the vector $\mathbf{Y}$ shrinks when we project it onto the regressors using $\mathbf{X}$. Note that since $\mathbf{P} = \mathbf{I}_n - \mathbf{M}$, we can also write R^2 as:

$$R^2 = \frac{\mathbf{Y}'(\mathbf{I}_n - \mathbf{M})\mathbf{Y}}{\mathbf{Y}'\mathbf{Y}} = 1 - \frac{\mathbf{Y}'\mathbf{MY}}{\mathbf{Y}'\mathbf{Y}} = 1 - \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{\mathbf{Y}'\mathbf{Y}}$$

7.5.7 The Frisch-Waugh-Lovell theorem: matrix version*

Suppose we're interested in just *part* of the vector $\hat{\beta}$. That is, we separate our regressors $X_1 \dots X_k$ into two groups, let's say $X_1 \dots X_j$ and $X_{j+1} \dots X_k$, for some j (this is without loss of generality since we could always re-order the indexing of the regressors). Our object of interest will be $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$, where we introduce the notation:

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_{\mathbf{1}} \\ \hat{\boldsymbol{\beta}}_{\mathbf{2}} \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \dots \\ \hat{\beta}_j \end{pmatrix} \\ \begin{pmatrix} \hat{\beta}_{j+1} \\ \hat{\beta}_{j+2} \\ \dots \\ \hat{\beta}_k \end{pmatrix} \end{pmatrix}$$

Analogously, define the matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ as

$$\mathbf{X}_1 := \underbrace{\begin{pmatrix} (X_{11}, X_{21}, \dots X_{j1}) \\ (X_{12}, X_{22}, \dots X_{j2}) \\ \vdots \\ (X_{1n}, X_{2n}, \dots X_{jn}) \end{pmatrix}}_{j \text{ columns}} \left.\vphantom{\begin{pmatrix} a \\ a \\ a \\ a \end{pmatrix}}\right\} n \text{ rows} \qquad \text{and} \qquad \mathbf{X}_2 := \underbrace{\begin{pmatrix} (X_{j+1,1}, X_{j+2,1}, \dots X_{k1}) \\ (X_{j+1,2}, X_{j+1,2}, \dots X_{k2}) \\ \vdots \\ (X_{j+1,n}, X_{j+2,n}, \dots X_{kn}) \end{pmatrix}}_{(k-j) \text{ columns}} \left.\vphantom{\begin{pmatrix} a \\ a \\ a \\ a \end{pmatrix}}\right\} n \text{ rows}$$

where $\mathbf{X}_1$ is a matrix of observations of the regressors $X_1 \dots X_j$ and $\mathbf{X}_2$ is a matrix of observations of the regressors $X_{j+1} \dots X_k$.

Define $n \times n$ projector matrices $\mathbf{P}_1 = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$ and $\mathbf{P}_2 = \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'$, and corresponding annihilator matrices $\mathbf{M}_1 = \mathbf{I}_n - \mathbf{P}_1$ and $\mathbf{M}_2 = \mathbf{I}_n - \mathbf{P}_2$. Note that by the same logic as Equation (7.34), the matrix $\mathbf{M}_1$ annihilates $\mathbf{X}_1$ (that is, $\mathbf{M}_1\mathbf{X}_1 = \mathbf{0}$, where $\mathbf{0}$, is a set of j zeroes), and similarly $\mathbf{M}_2\mathbf{X}_2 = \mathbf{0}$, where now $\mathbf{0}$, is a set of $k-j$ zeroes

The matrix $\mathbf{P}_1$ projects vectors in $\mathbb{R}^n$ into the subspace spanned by the columns of $\mathbf{X}_1$, which are the first j columns of $\mathbf{X}$. The matrix $\mathbf{M}_1$ projects vectors in $\mathbb{R}^n$ into the subspace orthogonal to the columns of $\mathbf{X}_1$. Similarly,$\mathbf{P}_2$ projects vectors in $\mathbb{R}^n$ into the subspace spanned by the columns of $\mathbf{X}_2$, which are the last $(k-j)$ columns of $\mathbf{X}$.

With the matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ in hand, we can now give an explicit formula for $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{2}}$, known famously as the *Frisch-Waugh-Lovell theorem*:

Proposition 7.7 (Frisch-Waugh-Lovell theorem).

$$\hat{\boldsymbol{\beta}}_{\mathbf{1}} = (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{Y}$$

and

$$\hat{\boldsymbol{\beta}}_{\mathbf{2}} = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{Y}$$

Proof. We'll prove the expression for $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$, as the proof for $\hat{\boldsymbol{\beta}}_{\mathbf{2}}$ is exactly analogous. Note that the full design matrix $\mathbf{X}$ can be written as $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$, and

$$\mathbf{X}\beta = [\mathbf{X}_1, \mathbf{X}_2](\hat{\boldsymbol{\beta}}_{\mathbf{1}}{}', \hat{\boldsymbol{\beta}}_{\mathbf{2}}{}')' = \mathbf{X}_1\hat{\boldsymbol{\beta}}_{\mathbf{1}} + \mathbf{X}_2\hat{\boldsymbol{\beta}}_{\mathbf{2}}$$

We can thus rewrite Equation (7.33) as:

$$\mathbf{Y} = \mathbf{X}\hat{\beta} + \hat{\mathbf{e}} = \mathbf{X}_1\hat{\boldsymbol{\beta}}_{\mathbf{1}} + \mathbf{X}_2\hat{\boldsymbol{\beta}}_{\mathbf{2}} + \hat{\mathbf{e}}$$

Consider multiplying both sides of this equation by $\mathbf{M}_2$. Since $\mathbf{M}_2\mathbf{X}_2 = \mathbf{0}$, where $\mathbf{0}$ is a vector of $k-j$ zeros, we have:

$$\mathbf{M}_2\mathbf{Y} = \mathbf{M}_2\mathbf{X}_1\hat{\boldsymbol{\beta}}_{\mathbf{1}} + \mathbf{M}_2\hat{\mathbf{e}}$$

Now mulitply this equation from the left by $\mathbf{X}_1'$:

$$\mathbf{X}_1'\mathbf{M}_2\mathbf{Y} = \mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1\hat{\boldsymbol{\beta}}_{\mathbf{1}} + \mathbf{X}_1'\mathbf{M}_2\hat{\mathbf{e}}$$

The final step of the proof is to show that $\mathbf{X}_1'\mathbf{M}_2\hat{\mathbf{e}} = \mathbf{0}$, and then multiply each side by $(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}$ to get the final expression for $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$. To see that $\mathbf{X}_1'\mathbf{M}_2\hat{\mathbf{e}} = \mathbf{0}$, recall that $\hat{\mathbf{e}} = \mathbf{MY}$, where $\mathbf{M}$ is the annihilator matrix for the full design matrix $\mathbf{X}$. Since the vector $\hat{\mathbf{e}} = \mathbf{MY}$ is orthogonal to all of the columns of $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$, then it is also orthogonal to all of the columns of $\mathbf{X}_1$, which is just a susbset of these. That implies that $\mathbf{M_1}\hat{\mathbf{e}} = \hat{\mathbf{e}}$. Now $\mathbf{X}_1'\mathbf{M}_2\hat{\mathbf{e}} = \mathbf{X}_1'\hat{\mathbf{e}} = \mathbf{X}_1'\mathbf{MY} = \mathbf{0}$, where in the last step we've used that $\mathbf{X}_1'\mathbf{M} = \mathbf{0}'$. □

Now let's see how the Frisch-Waugh-Lovell theorem relates to the "regression anatomy" result Proposition 7.5. Since $\mathbf{M}_2$ is idempotent, we can write

$$\hat{\boldsymbol{\beta}}_{\mathbf{1}} = (\mathbf{X}_1'\mathbf{M}_2\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{Y} = (\tilde{\mathbf{X}}_1'\tilde{\mathbf{X}}_1)^{-1}\tilde{\mathbf{X}}_1'\mathbf{Y}$$

where $\tilde{\mathbf{X}}_1 := \mathbf{M}_2\mathbf{X}_1$, and we've used that $\mathbf{M}_2$ is a symmetric matrix: $\mathbf{M}_2' = \mathbf{M}_2$. The $n \times k$ matrix $\tilde{\mathbf{X}}_1 := \mathbf{M}_2\mathbf{X}_1$ collects the residuals from a series of j regressions: for each $\ell = 1 \dots j$, column ℓ of $\tilde{\mathbf{X}}_1$ is composed of the residuals from a regression of X_ℓ on $X_{j+1} \dots X_k$.

An analogous formula applies for $\hat{\boldsymbol{\beta}}_\mathbf{2}$, where $\tilde{\mathbf{X}}_2$ collects the residuals from regressions of each X_ℓ on $X_1 \dots \dots X_j$, for $\ell = j+1 \dots k$. In the special case in which $\mathbf{X}_2$ has a single column (e.g. we're interested only in $\hat{\beta}_k$, and we include a constant in the regression (e.g. $X_1 = 1$), then we get exactly a sample version of Proposition 7.5.

7.5.8 For the matrix haters: OLS in terms of covariances*

If we specialize to a regression that includes a constant in addition to k regressors, as in Section 7.4.3, we can express the OLS slope coefficients in terms of *sample covariances*. The model is:

$$Y_i = \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \dots + X_{ki} + \epsilon_i$$

Let us now consider the OLS estimates $\hat{\beta}_0, \hat{\beta}_1 \dots \hat{\beta}_k$. One useful property of $\hat{\beta}$ is that the fitted residuals

$$\hat{\epsilon}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 \cdot X_{1i} + \dots + \hat{\beta}_k \cdot X_{ki})$$

will exactly average out to zero: that is $\sum_{i=1}^n \hat{\epsilon}_i = 0$. This can be seen from Eq. (7.28) since one of our regressors is equal to 1 for all observations. This holds for any regression with a constant, and will be a useful fact in what follows.

From the Frisch-Waugh-Lovell theorem, we can obtain an expression for each slope coefficient estimate $b\hat{e}taj$. In particular:

$$\hat{\beta}_j = \frac{\widehat{Cov}(\hat{\epsilon}^j, Y)}{\widehat{Var}(\hat{\epsilon}^j)} \tag{7.35}$$

where we let $\hat{\epsilon}^j$ denote the fitted residuals from a regression of X_j on all the other regressors and a constant. This is an analog of the population residual $\tilde{X}_j$ from this same regression. Equation (7.35) provides a "sample analog" to the regression anatomy formula: Proposition 7.5.

The operators $\widehat{Cov}$ and $\widehat{Var}$ appearing in Eq. (7.35) are defined as follows. Let $\mathbf{A} = (A_1, A_2, \dots A_n)$ $\mathbf{B} = (B_1, B_2, \dots B_n)$ be $n \times 1$ vectors composed of observations of a random variable A_i and B_i, respectively. Let $\bar{A}_n = \frac{1}{n}\sum_{i=1}^n A_i$ be the sample mean of A_i, and similarly for $\bar{B}_n$. Then, we define:

$$\widehat{Cov}(A, B) = \left(\frac{1}{n}\sum_{i=1}^n A_i \cdot B_i\right) - \bar{A}_n \cdot \bar{B}_n$$

and

$$\widehat{Var}(A) = \widehat{Cov}(A, B) = \left(\frac{1}{n}\sum_{i=1}^n A_i^2\right) - \left(\bar{A}_n\right)^2$$

As a special case of Eq. (7.35), we have that in simple linear regression

$$\hat{\beta}_1 = \frac{\widehat{Cov}(X, Y)}{\widehat{Var}(X)}$$

where in this case $\hat{\epsilon}_i^j$ is simply equal to X_i, the i^{th} observation of our single regressor X. We can work out the estimate of the constant β_0 from the fact that the fitted residual $\hat{\epsilon}_i$ satisfies $\frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot X_i) = 0$. $\hat{\beta}_0$ is thus $\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_0 \cdot \bar{X}_n$.

To see how to obtain Eq. (7.35), suppose we're interested in β_k, in which case $\mathbf{X}_2$ is a $n \times 1$ vector of observations of X_j and $\mathbf{X}_1$ is a matrix of observations of the other regressors and a constant. Note that since $\hat{\epsilon}_i^j$ is a fitted residual from a regression that includes a constant, $\frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^j = 0$.

Thus, we wish to show that

$$\hat{\beta}_k = \frac{\frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^j \cdot Y_i}{\frac{1}{n}\sum_{i=1}^{n}(\hat{\epsilon}_i^j)^2} = \frac{\frac{1}{n}\hat{\boldsymbol{\epsilon}}^{\boldsymbol{j}\prime}\mathbf{Y}}{\frac{1}{n}\hat{\boldsymbol{\epsilon}}^{\boldsymbol{j}\prime}\hat{\boldsymbol{\epsilon}}^{\boldsymbol{j}}} = \frac{(\mathbf{X}_2\mathbf{M}_1\mathbf{Y})}{\mathbf{X}_2\mathbf{M}_1\mathbf{X}_2}$$

where $\hat{\boldsymbol{\epsilon}}^{\boldsymbol{j}} := (\hat{\epsilon}_1^j, \hat{\epsilon}_2^j, \dots \hat{\epsilon}_n^j)'$ is a vector of the $\hat{\epsilon}_i^j$ across i. Since in this case $\mathbf{X}_2\mathbf{M}_1\mathbf{X}_2$ is a scalar (recall that we're looking at a single coefficient, and have thus defined $\mathbf{X}_2$ to be a vector rather than a matrix), and thus what we aim to show is $\hat{\beta}_k = (\mathbf{X}_2\mathbf{M}_1\mathbf{X}_2)^{-1}(\mathbf{X}_2\mathbf{M}_1\mathbf{Y})$. This is exactly what is given by Proposition 7.7.

7.5.9 A review of notation

Let's review the notation that we've introduced in this section, because it can be confusing.

- We began with a random variable Y and a random vector X, which are related by $Y = X'\beta + \epsilon$ in "the population". The random vector X can be written $X = (X_1, X_2, \dots X_k)'$, where each X_j is a different *regressor*. No i subscripts are necessary here.
- Then we draw a random *sample*, where observations are indexed by $i = 1 \dots n$. Y_i is a random variable reflecting the value of Y in the i^{th} observation, and X_i assembles the value of all regressors for observation i into a random vector: $X_i = (X_{1i}, X_{2i}, \dots X_{ki})'$.
- When discussing the OLS estimator, it is convenient to assemble information across all of the observations, leading to the $n \times 1$ vector $\mathbf{Y}$ and the $n \times k$ matrix $\mathbf{X}$.

Consider the following toy dataset, where $n = 4$ and $k = 3$. This reflects a realization of the random matrix $\mathbf{X}$ and the random vector $\mathbf{Y}$:

i	X1	X2	X3	Y
1	1	4	0	23
2	1	3	1	54
3	1	2	1	21
4	1	6	0	77

The 4×3 matrix framed by a large red box is $\mathbf{X}$ in our sample. The smaller green box inside indicates X_3 laid out as a row vector: the values of each of the three regressors in the third observation. The blue skinny rectangle indicates the $n \times 1$ vector $\mathbf{Y}$. Note that our first "regressor" X1 is simply one for each observation, and contributes a constant to our regression. Regressor X3 is a binary or "dummy" random variable: taking values of only zero or one for all obseravtions.

7.6 Statistical properties of the OLS estimator

In this section we'll see that the OLS estimator $\hat{\beta}$ has many of the desirable properties introduced in Section 5.3. It is consistent for the true population regression coefficient vector β, and has an asymptotically normal distribution. Knowing this will allow us to test hypotheses about the regression vector β. We also show in this esction that OLS is an unbiased estimator of β, and is an efficient estimator in a precise sense.

Recall from Section 5.3 that when considering the performance of an estimator, we want to compare it to the population parameter of interest, in this case β. How can we do this? Well, we know from Equation (7.31) that $\hat{\beta}$ is a function of our observations of the outcome $\mathbf{Y}$, and our observations of the regressor $\mathbf{X}$. So we need some way to relate these observations to the population parameter of interest β. Our model of Y_i does exactly that. Recall that Equation 7.2 describes how our observations of Y_i can be written in terms of the coefficients β. Equation (7.32) provides an equivalent statement of this in vector notation. Studying the statistical properties of $\hat{\beta}$ thus begins with the following crucial step: substitute our equation for $\mathbf{Y}$ (Eq. 7.32) into the definition of the estimator (Eq. 7.31):

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \boldsymbol{\epsilon}) \\
&= \cancel{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} \\
&= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} \qquad (7.36)
\end{aligned}$$

Eq. (7.36) is really quite remarkable: it says that regardless of whatever sample we ended up estimating $\hat{\beta}$ from, it is exactly equal to the true population parameter β, plus second term that depends on the vector of residuals $\boldsymbol{\epsilon}$ and the sample design matrix $\mathbf{X}$.

We'll now proceed in two steps. First, we'll study the distribution of $\hat{\beta}$ when our sample design matrix is held fixed. This allows us to establish that conditional on $\mathbf{X}$, the estimator $\hat{\beta}$ is unbiased and efficient. Then, we'll consider the properties of $\hat{\beta}$ as n gets very large.

Keep in mind what we're doing in this section: we're asking what the distribution of our estimator $\hat{\beta}$ is, given that the data in our sample was a random draw from an underlying population. This will allow us to think about questions like: how likely would we be to get an estimate $\hat{\beta}$ that is far from β, given that the sample we use to compute $\hat{\beta}$ is random (and thus could have been different that the one we actually see)?

7.6.1 Finite sample properties of $\hat{\beta}$, conditional on $\mathbf{X}$*

Understanding the finite-sample sampling distribution of an estimator is generally quite hard, since we can't rely on the asymptotic properties of Chapter 4. However, things become easier if we consider the *conditional* distribution of $\hat{\beta}$, given our observed design matrix $\mathbf{X}$ (which was itself a random draw). In this section we'll see that OLS is both unbiased and efficient, conditional on the realized design matrix.

For these results, we'll assume that the linear regression model holds and that we have an independent and identically distributed sample:

Assumption 1 (linear regression model and *i.i.d* sampling). *(Y_i, X_i) is an i.i.d. sample from the model: $Y = X'\beta + \epsilon$ with $\mathbb{E}[\epsilon|X] = 0$.*

Assumption 1 implies that for each $i = 1, 2, \ldots, n$:

$$Y_i = X_i'\beta + \epsilon_i$$

where $\mathbb{E}[\epsilon_i|X_i] = 0$.

With $\mathbf{X}$ fixed, we know from Eq. (7.36) that the only variation in $\hat{\beta}$ comes from variation in the residuals ϵ_i. What can we say about the distribution of the ϵ_i?

7.6.1.1 Unbiasedness

Proposition 7.8. *Given Assumption 1, OLS is conditionally unbiased for β; that is: $\mathbb{E}[\hat{\beta}|\mathbf{X}] = \beta$.*

Note that by the law of iterated expectations, this implies that $\hat{\beta}$ is also *unconditionally* unbiased:

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}\left\{\mathbb{E}[\hat{\beta}|\mathbf{X}]\right\} = \mathbb{E}\left\{\beta\right\} = \beta$$

Proof. Note that by 7.36, $\mathbb{E}[\hat{\beta}|\mathbf{X}] = \beta + \mathbb{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\middle|\,\mathbf{X}\right]$. Our goal will be to show that the second term is zero (for each component of β).

Since $\mathbf{X}$ is not random, conditional on $\mathbf{X}$, we can pull it out of the expectation:

$$\mathbb{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\middle|\,\mathbf{X}\right] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}\left[\boldsymbol{\epsilon}\middle|\,\mathbf{X}\right] \qquad (7.37)$$

Now consider the quantity $\mathbb{E}\left[\boldsymbol{\epsilon}\middle|\,\mathbf{X}\right]$. This is an $n \times 1$ vector, of which the i^{th} component is $\mathbb{E}[\epsilon_i|\mathbf{X}]$. Noting that conditioning on $\mathbf{X}$ is the same as conditioning on $X_1, X_2, \ldots, X_n$, we have that:

$$\mathbb{E}[\epsilon_i|\mathbf{X}] = \mathbb{E}[\epsilon_i|X_1, X_2, \ldots, X_n] = \mathbb{E}[\epsilon_i|X_i] = 0$$

The second equality above follows from $i.i.d$ sampling. Since Y_i and X_i are jointly independent of X_j for $X_j \neq i$ (recall that ϵ_i is a function of Y_i and X_i, i.e. $\epsilon_i = Y_i - X_i'\beta$), we can remove all of the X_j for $j \neq i$ from the conditioning event. The second equality then follows from the linear regression model assumption.

Considering that the above holds for each i, we have all together that $\mathbb{E}\left[\boldsymbol{\epsilon}|\,\mathbf{X}\right] = \mathbf{0}_n$, where we let $\mathbf{0}_n$ denote a vector of n zeroes. This then implies that $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}\left[\boldsymbol{\epsilon}|\,\mathbf{X}\right] = \mathbf{0}_k$, where we let $\mathbf{0}_k$ denote a vector of k zeroes. This completes the proof. □

7.6.1.2 Efficiency

Given that OLS is (conditionally) unbiased, we know that conditional on $\mathbf{X}$, it's mean squared error as an estimator of β is equal to it's variance (recall the bias-variance decomposition of Eq. 5.2). A well-known result about OLS is that it has the smallest variance among all unbiased estimators of β. To state this result, we will require an extra assumption, which is that the errors ϵ_i have the same variance for each observation:

Definition 7.1. *We call the linear regression model (conditionally)* ***homoskedastic*** *if for some value* σ, $Var(\epsilon_i|X_i) = \sigma^2$ *for all* i.

Homoskedasticity is a very strong assumption, and won't hold in practice in most settings. But it has some use as a simplifying assumption to help understand OLS.

Proposition (Gauss-Markov Theorem). *Consider an alternative estimator* $\tilde{\beta}$ *of* β *that is also conditionally unbiased:* $\mathbb{E}[\tilde{\beta}|\mathbf{X}] = \beta$. *Given Assumption 1 and homoskedasticity, OLS is efficient for* β *in the sense that:*

$$Var(\tilde{\beta}|\mathbf{X}) \geq Var(\hat{\beta}|\mathbf{X})$$

where for two $k \times k$ *matrices* $\mathbf{A}$ *and* $\mathbf{B}$, *we say that* $\mathbf{A} \geq \mathbf{B}$ *when the matrix* $\mathbf{A} - \mathbf{B}$ *is positive semi-definite.*

The above version of the Gauss-Markov Theorem is due to Hansen (2021), which was proved just this year (2021)! Most textbooks (with the exception of the Hansen book) add the assumption that the candidate estimator $\tilde{\beta}$ is a linear function of $\mathbf{Y}$ (like OLS is). In this context, the Gauss-Markov theorem is often described as saying that OLS is B.L.U.E: the *best linear unbiased estimator* of β. The above "modern Gauss-Markov" theorem is a stronger result than BLUE: it implies that considering non-linear estimators of β wouldn't help us reduce the variance of $\tilde{\hat{}}$ beyond the bound acheived by OLS.

What is the conditional variance of OLS $Var(\hat{\beta}|\mathbf{X})$ appearing in the Gauss-Markov Theorem? Note that homoskedasticity along with the $i.i.d$ assumption imply that:

$$Var(\boldsymbol{\epsilon}|\mathbf{X}) = \mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|\mathbf{X}\right] = \sigma^2\mathbf{I}_n$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

Exercise: Definition 7.1 implies that the diagonal elements of $Var(\boldsymbol{\epsilon}|\mathbf{X})$ are equal to σ^2, but how do we know that the off-diagonal elements are equal to zero?

The above in turn implies that the conditional variance of the OLS estimator is:

$$\begin{aligned} Var(\hat{\beta}|\mathbf{X}) &= \mathbb{E}[(\beta - \hat{\beta})(\beta - \hat{\beta})'] = \mathbb{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\middle|\,\mathbf{X}\right] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I}_n)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2\cancel{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

Thus the Gauss-Markov Theorem is typically written as saying that $Var(\tilde{\beta}|\mathbf{X}) \geq \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, where the RHS is the conditional variance of the OLS estimator.

What should we make of the Gauss-Markov Theorem given that homoskedasticity is usually not a reasonable assumption? It turns out that OLS is actually *not* efficient when homoskedasticity fails (we call this *heteroskedasticity*). In that setting, a closely-related estimator known as *weighted-least-squares* (WLS) becomes efficient (at least among linear estimators). Efficient WLS minimizes $\sum_{i=1}^{n} w_i(Y_i - X_i'\beta)^2$ where the w_i are chosen to be $1/Var(\epsilon_i|\mathbf{X})$. In practice, we don't generally know $Var(\epsilon_i|\mathbf{X})$ ex-ante, so a feasible version of WLS requires estimating this quantity. Basically, the re-weighting of each observation by w_i "undoes" heteroskedasticity, so that WLS mimics the properties that OLS has in the homoskedastic case. Thus you should think of Gauss-Markov as illustrating an idea that is more general that homoskedasticity, but is simpler to express under that strong assumption.

7.6.2 Asymptotic properties of $\hat{\beta}$

We now turn to deriving properties of the OLS estimator as the sample size n gets very large.

We'll show that $\hat{\beta}$ is a consistent estimator for β, and then that its sampling distribution is asymptotically normal. For these results, we don't need for the linear regression model with $\mathbb{E}[\epsilon|X] = 0$ to hold. The large sample properties hold for the linear projection coefficient β even if the CEF of Y on X is not linear. As before, we assume that we have an independent and identically distributed sample:

Assumption 2 (linear projection model and *i.i.d* sampling). *(Y_i, X_i) is an i.i.d. sample from the model: $Y = X'\beta + \epsilon$ with $\mathbb{E}[\epsilon \cdot X] = \mathbf{0}$.*

To make claims that involve convergence in probability and convergence in distribution, we will consider a sequence of estimators $\hat{\beta}$, indexed by the sample size n. For each $n = 1, \dots, \infty$ along the sequence, we assume that 2 holds. As a reminder (cf. Chapter 4), in reality sample sizes never actually "grow" to infinity. In practice, we always have an actual sample that has some actual finite size n. The idea of an asymptotic sequence exists only to provide an *approximation* to the sampling distribution of $\hat{\beta}$ given our fixed n, which we will take to be accurate when the sample size is big enough.

7.6.2.1 Consistency

We'll first see that given the asymptotic sequence described above, $\hat{\beta} \xrightarrow{p} \beta$. That is, $\hat{\beta}$ is a consistent estimator of β.

Subtracting β from each side of Equation 7.36:

$$\hat{\beta} - \beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} = \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon}$$

where in the second equality we've used that the factor of $\frac{1}{n}$ inside the matrix inverse cancels the one on $\mathbf{X}'\boldsymbol{\epsilon}$. Now let's consider this latter quantity alone. Expanding out the matrix product:

$$\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon} = \frac{1}{n}\sum_{i=1}^{n} X_i\epsilon_i,$$

i.e. it is equal to the sample average of the random variable $X_i\epsilon_i$. To see the above, note that $\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon}$ is a $k \times 1$ vector, whose j^{th} element is equal to the inner product between ϵ and the j^{th} row of $\mathbf{X}'$. The j^{th} row of $\mathbf{X}'$ is equal to the j^{th} column of $\mathbf{X}$, which is comprised of the n observations of regressor X_j.

Thus, by the law of large numbers, we have that $\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon} \xrightarrow{p} \mathbb{E}[X_i\epsilon_i]$, provided that $\mathbb{E}[X_i\epsilon_i] < \infty$. By the linear projection model (Assumption 2), $\mathbb{E}[X_i\epsilon_i] = \mathbf{0}$, where $\mathbf{0}$ is a vector of k zeroes.

Similarly, we have by the law of large numbers that

$$\frac{1}{n}\mathbf{X}'\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n} X_iX_i' \xrightarrow{p} \mathbb{E}[X_iX_i'],$$

In Chapter 4 we only considered the LLN for random vectors, not random *matrices* like X_iX_i'. But since you can always rewrite an $n \times m$ matrix as a vector with $n \cdot m$ elements, the LLN for vectors applies so long as each element of the matrix $\mathbb{E}[X_iX_i']$ is finite. In the box below, I state a set of assumptions, "regularity conditions", that ensure we can use the law of large numbers here, and that all expectations that appear in this section exist.

Given that $\frac{1}{n}\mathbf{X}'\mathbf{X} \xrightarrow{p} \mathbb{E}[X_iX_i']$, the continuous mapping theorem implies that

$$\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \xrightarrow{p} \mathbb{E}[X_iX_i']^{-1}$$

That's because for a general invertible matrix $\mathbf{M}$, the matrix inverse function $\mathbf{M}^{-1}$ is a continuous function of each of the elements of $\mathbf{M}$.

Finally, by the continuous mapping theorem, we have that

$$\hat{\beta} - \beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} = \underbrace{\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)}_{\xrightarrow{p}\mathbb{E}[X_iX_i']}^{-1} \underbrace{\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon}}_{\xrightarrow{p}\mathbf{0}} \xrightarrow{p} \mathbb{E}[X_iX_i']^{-1}\mathbf{0} = \mathbf{0}$$

Thus we have proved that $\hat{\beta} \xrightarrow{p} \beta$.

Proposition 7.9. *OLS is consistent for β given Assumption 2 and the regularity conditions 3 below.*

How do we know that the matrix $\mathbb{E}[X_iX_i']$ has only finite entries, so that we can use the LLN?

One might simply assume that this is true, but the conventional textbook approach is to state a set of conditions that are sufficient for $\mathbb{E}[X_iX_i']$ to be finite, but are easier or more natural to state. Technical assumptions of this kind are often referred to as "regularity conditions". The Hansen text uses the following:

Assumption 3 (regularity conditions for consistency). *Suppose that:*

1. *$\mathbb{E}[Y_i^2]$ is finite*
2. *$\mathbb{E}[||X_i||^2]$ is finite*
3. *We have no perfect multicollinearity in the population: that is, $\mathbb{E}[X_iX_i']$ is positive definite.*

where for any vector with k components $\mathbf{a}$, we let $||\mathbf{a}||$ denote its Euclidean norm, i.e. $||\mathbf{a}||^2 = \sum_{j=1}^k (a_j)^2$.

Let us now see how Assumption 3 get us what we need to use the LLN to analyze the OLS estimator. In particular, we'll use item 2, that $\mathbb{E}[||X_i||^2] = \mathbb{E}\left[\sum_{j=1}^k X_{ji}^2\right] = \sum_{j=1}^k \mathbb{E}\left[X_{ji}^2\right] < \infty$. Note that this implies that $\mathbb{E}\left[X_{ji}^2\right] < \infty$ for each j, since these quantities are all positive would could never have one of them be infinite but the sum finite.

Now consider a generic element of the matrix $\mathbb{E}[X_iX_i']$. For instance, the element in the j^{th} row, ℓ^{th} column is $\mathbb{E}[X_{ji}X_{\ell i}]$. To show that this must be finite, we'll use the following very useful inequality for expectations:

1. The *Cauchy-Schwarz inequality* says that for random variables X and Y: $|\mathbb{E}[X \cdot Y]|^2 \leq \mathbb{E}[X^2] \cdot \mathbb{E}[Y^2]$

Applying the Cauchy-Schwartz to the k, ℓ element of $\mathbb{E}[X_iX_i']$, we have that

$$\mathbb{E}[X_{ji}X_{\ell i}] \leq \sqrt{\mathbb{E}[X_{ji}^2] \cdot \mathbb{E}[X_{\ell i}^2]}$$

Since each of $\mathbb{E}[X_{ji}^2]$ and $\mathbb{E}[X_{\ell i}^2]$ are finite, their product must also be finite, and also its square root.

7.6.2.2 Asymptotic normality

Now let's use the central limit theorem to derive the asymptotic distribution of the OLS estimator. Let us pick up from the expression $\hat{\beta} - \beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}$. Knowing that the central limit theorem will involve a factor of $\sqrt{n}$, let's rewrite this as

$$\sqrt{n}(\hat{\beta}-\beta) = \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \cdot \sqrt{n}\left(\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon}\right)$$

Recall that $\mathbb{E}[X_i\epsilon_i] = \mathbf{0}$, where $\mathbf{0}$ is a vector of k zeroes, and that $\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon}$ is the sample mean of the random vector $X_i \cdot \epsilon_i$. Using the notation of Chapter 4, let's denote this as $\overline{(X\epsilon)}_n := \frac{1}{n}\sum_{i=1}^n X_i \cdot \epsilon_i$. Then we can write the above as:

$$\sqrt{n}(\hat{\beta}-\beta) = \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \cdot \sqrt{n}\left(\overline{(X\epsilon)}_n - \mathbb{E}[X_i\epsilon_i]\right)$$

The rightmost factor in the above expression has exactly the form that we need to apply the CLT, in particular:

$$\sqrt{n}\left(\overline{(X\epsilon)}_n - \mathbb{E}[X_i\epsilon_i]\right) \xrightarrow{d} N(\mathbf{0}, Var(X_i\epsilon_i)),$$

Note that since $\mathbb{E}[X_i\epsilon_i] = \mathbf{0}$, we can write the variance as

$$Var(X_i\epsilon_i) = \mathbb{E}[(X_i\epsilon_i)(X_i\epsilon_i)'] = \mathbb{E}[\epsilon_i^2 X_i X_i'] \tag{7.38}$$

Now we use the Slutsky theorem

$$\sqrt{n}(\hat{\beta}-\beta) = \underbrace{\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}}_{\xrightarrow{p}\mathbb{E}[X_iX_i']} \cdot \underbrace{\sqrt{n}\left(\overline{(X\epsilon)}_n - \mathbb{E}[X_i\epsilon_i]\right)}_{\xrightarrow{d}\mathbb{E}[\epsilon_i^2X_iX_i']} \xrightarrow{d} \mathbb{E}[X_iX_i']^{-1}N(\mathbf{0}, \mathbb{E}[\epsilon_i^2X_iX_i'])) \tag{7.39}$$

That is, $\sqrt{n}(\hat{\beta}-\beta)$ converges in distribution to a random vector whose distribution is that of the matrix $\mathbb{E}[X_iX_i']^{-1}$ times a normal vector with mean zero (for each component) and variance-covariance matrix $\mathbb{E}[\epsilon_i^2X_iX_i']$.

The RHS of (7.39) is thus equal to a linear combination of normal random vectors. Adapting Proposition 3.5, note the following: let $X \sim N(\mu, \mathbf{\Sigma})$ be a k-component random vector. Then for any $k \times k$ matrix $\mathbf{A}$:

$$\mathbf{A}'X \sim N(\mathbf{A}'\mu, \mathbf{A}'\mathbf{\Sigma}\mathbf{A}))$$

(this can also be seen as an example of the delta method, applied to a vector-valued function h). Thus, we can write (7.39) as

$$\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}) \tag{7.40}$$

where $\mathbf{V} := \mathbb{E}[X_iX_i']^{-1}\mathbb{E}[\epsilon_i^2X_iX_i']\mathbb{E}[X_iX_i']^{-1}$. We refer to $\mathbf{V}$ as the *asymptotic variance* of the OLS estimator.

Example: Note that in the special case of homoskedasticity studied in Section 7.6.1, we have that $\mathbb{E}[\epsilon_i^2|X_i] = \sigma^2$ for all i, and thus by the law of iterated expectations:

$$\mathbb{E}[\epsilon_i^2X_iX_i'] = \mathbb{E}\left\{\mathbb{E}[\epsilon_i^2|X_i]X_iX_i'\right\} = \mathbb{E}\left\{\sigma^2X_iX_i'\right\} = \sigma^2 \cdot \mathbb{E}[X_iX_i'] \tag{7.41}$$

In this case the asymptotic variance takes on a very simple form:

$$\mathbf{V} = \mathbb{E}[X_iX_i']^{-1}\sigma^2 \cdot \cancel{\mathbb{E}[X_iX_i']}\cancel{\mathbb{E}[X_iX_i']^{-1}} = \sigma^2 \cdot \mathbb{E}[X_iX_i']^{-1}$$

A sufficient condition for us to be able to apply the CLT (see Section 4.4) is that $\mathbb{E}[(X_i\epsilon_i)'(X_i\epsilon_i)] = \mathbb{E}[\epsilon_i^2X_i'X_i]$ be finite. This requires finite *fourth* moments of the data, rather than the finite second moments assumed to prove consistency of OLS. To see why, note that for any j and ℓ:

$$\mathbb{E}[\epsilon_i^2X_{ji}\cdot X_{\ell i}] = \mathbb{E}[(Y_i - X_i'\beta)^2X_{ji}\cdot X_{\ell i}] = \mathbb{E}[Y_i^2\cdot X_{ji}X_{\ell i}] - 2\mathbb{E}[X_i'\beta\cdot Y_i\cdot X_{ji}\cdot X_{\ell i}] + \mathbb{E}[\beta'X_iX_i'\beta\cdot X_{ji}\cdot X_{\ell i}]$$

which can be written out as a sum over expectations that each involve the product of four random variables. To keep all such terms finite, Hansen assumes the following:

Assumption 4 (regularity conditions for asymptotic normality). *Suppose that:*

1. $\mathbb{E}[Y_i^4]$ *is finite*
2. $\mathbb{E}[||X_i||^4]$ *is finite*
3. *We have no perfect multicollinearity in the population: that is,* $\mathbb{E}[X_i X_i']$ *is positive definite.*

7.6.2.3 Estimating the asymptotic variance

Equation 7.40 is not immediately useful, unless we know the asymptotic variance matrix $\mathbf{V}$. Since we don't know $\hat{\mathbf{V}}$ before seeing the data, we will estimate it! In this section we see that we can construct a consistent estimator $\hat{\mathbf{V}}$ such that $\hat{\mathbf{V}} \xrightarrow{p} \mathbf{V}$. Doing this will open the door to hypothesis testing, which we'll consider in the next section.

Before seeing how hypothesis testing will work, let's consider how to construct the estimator $\hat{\beta}$ for the asymptotic variance of OLS. Note that $\mathbf{V} = \mathbb{E}[X_i X_i']^{-1}\mathbb{E}[\epsilon_i^2 X_i X_i']\mathbb{E}[X_i X_i']^{-1}$ has a "sandwich" form: it puts the matrix $\mathbb{E}[\epsilon_i^2 X_i X_i']$ (the meat)[2], between two instances of the matrix $\mathbb{E}[X_i X_i']^{-1}$ (the bread). By the continuous mapping theorem, we can construct an estimator $\mathbf{V}$ by making a sandwich out of consistent estimators for the meat and for the bread.

We've already seen that $\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}$ is a consistent estimator for the bread: $\mathbb{E}[X_i X_i']^{-1}$. An estimator for the meat $\mathbb{E}[\epsilon_i^2 X_i X_i']$ is not quite as obvious. It's sample analog $\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 X_i X_i'$ would definitely work, but the true residuals ϵ_i are not observed. However, we can use the *fitted* residuals $\hat{\epsilon}_i$, which are a function of the observed data, instead. We can write this in matrix form as:

$$\hat{\mathbf{\Omega}} := \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2 X_i X_i'$$

One can verify that $\hat{\Omega} \xrightarrow{p} \mathbb{E}[\epsilon_i^2 X_i X_i']$. Thus, we can form a consistent variance estimator as

$$\hat{\mathbf{V}}_{HC0} := \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \hat{\mathbf{\Omega}} \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \tag{7.42}$$

Eq. (7.42) is referred to as the "HC0" estimator of $\mathbf{V}$, where HC stands for *heterokedasticity consistent*. This name comes from the fact that $\hat{\mathbf{V}}_{HC0}$ does not require the assumption of homoskedasticity (Definition 7.1) to be a consistent estimator of $\mathbf{V}$.

When you run a command like `regress y x, robust` in Stata, the default covariance estimator is the so-called "HC1" estimator of $\mathbf{V}$:

$$\hat{\mathbf{V}}_{HC1} := \frac{n}{n-k}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \hat{\mathbf{\Omega}} \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \tag{7.43}$$

Note that the additional factor $\frac{n}{n-k}$ will make very little difference when n is large compared with k, and will make no difference in the asymptotic limit, since $\frac{n}{n-k} \to 1$ as $n \to \infty$. Applying this rescaling however can be helpful when n is small. It's easiest to understand the justification in the case of homoskedasticity, which is left as an exercise (see box below).

Note: there are further estimators floating around, with names HC2, HC3, and HC4. These apply further modifications to $\hat{\mathbf{V}}_{HC0}$ (see the Hansen text for details). Other variance estimators exist for certain violations of the *i.i.d* sampling assumption, including *cluster-robust* variance estimators for clustered sampling and autocorrelation-consistent estimators for serially correlated panel data.

If we had reason to believe that homoskedasticity holds, then it is not necessary to use the matrix $\hat{\mathbf{\Omega}}$ when estimating $\mathbf{V}$. The alternative estimator given below will perform better provided that the assumption of homoskedasticity is true. But, it will be inconsistent if not.

[2] Ideally plant-based meat :)

Recall from Eq. 7.41 that under homoskedasticity $\mathbb{E}[\epsilon_i^2 X_i X_i'] = \sigma^2 \cdot \mathbb{E}[X_i X_i']$, where $\sigma^2 = \mathbb{E}[\epsilon_i^2|X_i] = \mathbb{E}[\epsilon_i^2]$. Thus, we just need a consistent estimator of $\mathbb{E}[\epsilon_i^2]$, which we can then multiply by $\frac{1}{n}\mathbf{X}'\mathbf{X}$. The standard estimator of $\mathbb{E}[\epsilon_i^2]$ is denoted by Hansen as s^2, where

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

Thus an estimator of $\mathbf{V}$ that is valid under the assumption of homoskedasticity is $s^2 \cdot \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}$. This is what Stata computes by default, if you don't include `, robust` at the end of your regression command.

Exercise: Why use the estimator s^2 as written rather than the simpler expression $\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2$? The reason is that dividing by $n-k$ rather than n makes s^2 an unbiased estimator of $\mathbb{E}[\epsilon_i^2]$. Derive the bias of the estimator $\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2$, and show that s^2 is unbiased.

7.7 Inference on the regression vector β

Given a consistent estimator of $\mathbf{V}$ like the HC0 or the HC1 estimator, we can transform the quantity $\sqrt{n}(\hat{\beta}-\beta)$ into one whose limiting distribution is well-understood, and contains no unknown parameters. This proves to be a much more useful result than Equation (7.40), because it allows us to test hypotheses about the population regression vector β. In particular, if we pre-multiply $\sqrt{n}(\hat{\beta}-\beta)$ by the matrix $\hat{\mathbf{V}}^{-1/2}$ (see box below for the definition of $\hat{\mathbf{V}}^{-1/2}$):

Proposition 7.10. *Given Assumption 2, the regularity conditions 4, and a $\hat{\mathbf{V}}$ such that $\hat{\mathbf{V}} \xrightarrow{p} \mathbf{V}$:*

$$\sqrt{n}\hat{\mathbf{V}}^{-1/2}(\hat{\beta}-\beta) \xrightarrow{d} N(\mathbf{0}, \mathbb{I}_k)$$

where $\mathbb{I}_k$ is the $k \times k$ identity matrix.

The distribution $N(\mathbf{0}, \mathbb{I}_k)$ is that of k standard normal random variables, each of which is independent of the others (the variance-covariance matrix $\mathbb{I}_k$ has an entry of zero for each off-diagonal element). The power of Proposition 7.10 lies in the fact that the distribution appearing on the RHS, $N(\mathbf{0}, \mathbb{I}_k)$, contains no unknown quantities. We know exactly the probability that it associates to any event. Thus, for large n, we have a very good approximation to the distribution of $\sqrt{n}\hat{\mathbf{V}}^{-1/2}(\hat{\beta}-\beta)$. This provides the foundation for us to quantify uncertainty in our estimates $\hat{\beta}$ and test hypotheses about the regression vector β.

The matrix square root. One can show that the matrix $\mathbf{V} = \mathbb{E}[X_i X_i']^{-1}\mathbb{E}[\epsilon_i^2 X_i X_i']\mathbb{E}[X_i X_i']^{-1}$ is invertible, symmetric and positive definite. A property from linear algebra is that symmetric positive definite matrices $\mathbf{M}$ have a "matrix square root", which is a unique matrix $M^{1/2}$ such that $M = M^{1/2}M^{1/2}$. Thus, we can let $\mathbf{V}^{-1/2}$ denote the inverse of the matrix square root of $\mathbf{V}$. Let $\hat{\mathbf{V}}^{-1/2}$ denote the analogous matrix for our estimator $\hat{\mathbf{V}}$. For example, $\mathbf{V}_{HC0}^{-1/2}$ turns out to be: $\hat{\mathbf{V}}_{HC0}^{-1/2} = \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\hat{\mathbf{\Omega}}^{-1/2}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}$.

Proof of Proposition 7.10. By the continuous mapping theorem $\hat{\mathbf{V}}^{-1/2} \xrightarrow{p} \mathbf{V}^{-1/2}$. Then, by the Slutsky theorem and Eq. (7.40):

$$\sqrt{n}\hat{\mathbf{V}}^{-1/2}(\hat{\beta}-\beta) \xrightarrow{d} \mathbf{V}^{-1/2}N(\mathbf{0}, \mathbf{V}) = N(\mathbf{V}^{-1/2}\mathbf{0}, \mathbf{V}^{-1/2}\mathbf{V}\mathbf{V}^{-1/2}) = N(\mathbf{0}, \mathbb{I}_k)$$

where in the last step we've used that $\mathbf{V}^{-1/2}\mathbf{V}\mathbf{V}^{-1/2} = \mathbf{V}\mathbf{V}^{-1/2}\mathbf{V}^{-1/2} = \mathbf{V}\mathbf{V}^{-1} = \mathbb{I}_k$ (we've made use of the fact that "matrix powers" like $\mathbf{M}^{-1/2}$ always *commute* with the original matrix $\mathbf{M}$, meaning that $\mathbf{M}^{-1/2}\mathbf{M} = \mathbf{M}\mathbf{M}^{-1/2}$).

Note: here we've simply assumed that $\hat{\mathbf{V}}^{-1/2}$ exists. However it's possible to argue that it must exist for large enough n by appealing to the strong law of large numbers.

7.7.1 Testing hypotheses about a single regression coefficient

To see how the logic of Proposition 7.10 is useful, let's consider a simple setting, which turns out to be the most common one in practice: we are interested in the true value of a single regression coefficient, say β_j, in a regression that contains k regressors.

Note that we can write β_j as $\mathbf{e}_j'\beta$, where $\mathbf{e}_j = (0, \dots, 1, \dots 0)'$ is a k-component vector that puts a one in position j, and zeros everywhere else. Similarly, $\mathbf{e}_j'\hat{\beta}$ picks out the single component $\hat{\beta}_j$ from the OLS estimator. It then follows from Equation (7.40) and the Delta method that

$$\sqrt{n}(\hat{\beta}_j - \beta_j) = \mathbf{e}_j'\sqrt{n}(\hat{\beta}_j - \beta_j) \xrightarrow{d} \mathbf{e}_j'N(\mathbf{0}, \mathbf{V}) = N(\mathbf{e}_j'\mathbf{0}, \mathbf{e}_j'\mathbf{V}\mathbf{e}_j) = N(0, V_{jj})$$

where $V_{jj} = \mathbf{e}_j'\mathbf{V}\mathbf{e}_j$ is the j^{th} element along the diagonal of the matrix $\mathbf{V}$.

This implies, analogously to Proposition 7.10, that

$$\sqrt{n} \cdot \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\mathbf{V}}_{jj}}} \xrightarrow{d} N(0,1) \tag{7.44}$$

where $\hat{\mathbf{V}}_{jj}$ is the j^{th} element along the diagonal of the matrix $\hat{\mathbf{V}}$, which is a consistent estimator of V_{jj}. Note that we could have written the LHS of Eq. (7.44) as $\sqrt{n}\hat{\mathbf{V}}_{jj}^{-1/2}(\hat{\beta}_j - \beta_j)$ as in Proposition 7.10, but since $\hat{\mathbf{V}}_{jj}$ is a scalar we may take its conventional square root and divide by it.

We define the *standard error* for the estimate $\hat{\beta}_j$ to be $se(\hat{\beta}_j) := \sqrt{\hat{\mathbf{V}}_{jj}/n}$. Note that the standard error is a quantity that is computed from the data, given $\hat{\mathbf{V}}$ (it is an *estimate*, rather than a population quantity). By Eq. (7.44), we know that the quantity $(\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j)$ converges in distribution to a standard normal.

This allows us to test hypotheses about the value of β_j, using our estimate $\hat{\beta}_j$ and $se(\hat{\beta}_j)$. Consider the null hypothesis: $\mathbf{H}_0 : \beta_j = \beta_0$ for some value β_0 (e.g. zero). Define the *T-statistic* for this hypothesis to be

$$T(\beta_0) = \frac{\hat{\beta}_j - \beta_0}{se(\hat{\beta}_j)}$$

If $\mathbf{H}_0$ is true, then we know that $T(\beta_0) \xrightarrow{d} N(0,1)$. Recall from Section 5.4.1 that the *size* of a hypothesis test is the maximum probability of rejecting the null hypothesis, when the null hypothesis is in fact true. We can form a test with size α in the following way:

$$\text{reject } \mathbf{H_0} \text{ iff } |T(\beta_0)| > c$$

where c is a value such that the probability of a standard normal random variable having a magnitude of at least c is less than α. To do this in a way that maximizes power, we choose c to be exactly the $1 - \alpha/2$ quantile of the standard normal distribution: $c = \Phi^{-1}(1 - \alpha/2)$.

Exercise: Show that if $Z \sim N(0,1)$, $P(|Z| > \Phi^{-1}(1 - \alpha/2)) = \alpha$.

Note: using the standard normal distribution Φ to form our critical value c makes our test a so-called *z-test*. When you use the `regress` command in Stata, it performs a *t-test*, which uses the same test statistic $T(\beta_0)$ but a different critical value, instead based on the *students' t-distribution*. Using the t-distribution will be more accurate if n is small an the residuals are approximately normal and homoskedastic, but as n becomes large critical values based on the t-distribution or the standard normal distribution become the same. In modern (large n) datasets, this distinction isn't very important, so we just develop theory for the *z-test* here.

Example: A common hypothesis to test is that $\beta_j = 0$. In this case, our T-statistic is simply $\hat{\beta}_j/se(\hat{\beta}_j)$. It is common to construct the test with a size of 5%, in which case we reject the null that $\beta_j = 0$

if $\hat{\beta}_j/se(\hat{\beta}_j) > 1.96$, where $1.96 \approx \Phi^{-1}(1-.05/2)$. If we reject, then we say that the estimate $\hat{\beta}_j$ is "significant at the 95% level".

Note: The above is an example of a so-called "two-sided" test. If we had a null-hypothesis like $\mathbf{H}_0 : \beta_j \geq \beta_0$, we might apply an asymmetric decision rule like: $\mathbf{H_0}$ iff $|T(\beta_0)| > c$, where now $c = \Phi^{-1}(1-\alpha)$ is the $1-\alpha$ quantile of the standard normal distribution. This test also has a size of α.
A $1-\alpha$ *confidence interval* $\mathcal{CI}^{1-\alpha}$ for β_j is the set of all values β_0 for which the null-hypothesis $\mathbf{H}_0 : \beta_j = \beta_0$ is *not* rejected by a test with size α. In other words, it is for a two-sided test the set of all β_0 such that

$$\frac{\hat{\beta}_j - \beta_0}{se(\hat{\beta}_j)} \leq \Phi^{-1}(1-\alpha/2)$$

Rearranging this, we have that $\mathcal{CI}^{1-\alpha} = [\hat{\beta}_j - c \cdot se(\hat{\beta}_j), \hat{\beta}_j + c \cdot se(\hat{\beta}_j)]$, where $c = \Phi^{-1}(1-\alpha/2)$ grows with the desired α.

Example: A 95% confidence interval, based on a two-sided test, is

$$\mathcal{CI}^{95\%} = [\hat{\beta}_j - 1.96 \cdot se(\hat{\beta}_j), \hat{\beta}_j + 1.96 \cdot se(\hat{\beta}_j)]$$

Exercise: Show that as $\lim_{n\to\infty} P(\beta_j \in \mathcal{CI}^{1-\alpha}) = 1-\alpha$. That is, the $1-\alpha$ confidence interval contains the true value of β_j with probability $1-\alpha$, in the asymptotic limit. You may take for granted that if a sequence of random variables $Z_n \xrightarrow{d} Z$, then for any closed interval A of the real line: $\lim_{n\to\infty} P(Z_n \in A) = P(Z \in A)$. This statement is a consequence of the so-called *Portmanteau Theorem*.

Note: It is often said on the basis of the above that their is a 95% change that the true value of β_j lies in an e.g. 95% confidence interval. This language is sort of sloppy and makes it sound like β_j is a random variable, that may or may not lie inside the confidence interval. This is backwards: it is the confidence interval $\mathcal{CI}^{1-\alpha}$) that is random (it depends on the random variables $\hat{\beta}_j$ and $se(\hat{\beta}_j)$), while β_j is not (it is just some number).

7.7.2 Testing a joint hypothesis about the regression coefficients*

Suppose now that we want to test a hypothesis about the value of the full regression vector β. What would be the analog of our t-test from the last section?

It follows from Proposition 7.10 and the continuous mapping theorem that

$$n(\hat{\beta}-\beta)'\hat{\mathbf{V}}^{-1}(\hat{\beta}-\beta) = \left(\sqrt{n}\hat{\mathbf{V}}^{-1/2}(\hat{\beta}-\beta)\right)' \left(\sqrt{n}\hat{\mathbf{V}}^{-1/2}(\hat{\beta}-\beta)\right) \xrightarrow{d} \chi^2_k \tag{7.45}$$

where χ^2_k indicates the *chi-squared* distribution with k degrees of freedom, which is the distribution that applies to the sum of the squares of k independent standard normal random variables. To see this, note that the LHS must converge to the distribution that would apply to $Z'Z$, if $Z \sim N(\mathbf{0}, \mathbb{I}_k)$ is a vector of k mutually independent standard normal random variables.

Thus, to test the null-hypothesis $\mathbf{H}_0 : \beta = \beta_0$, we can compute $n(\hat{\beta}-\beta)'\hat{\mathbf{V}}^{-1}(\hat{\beta}-\beta)$, and compare it's value to quantiles of the chi-squared distribution. A test of this kind is called a *Wald-test*, and the test-statistic $n(\hat{\beta}-\beta)'\hat{\mathbf{V}}^{-1}(\hat{\beta}-\beta)$ is called a *Wald statistic*. Since a chi-squared random variable can only be positive, Wald-tests are inherently two-sided tests.

More generally, suppose we want to test a null-hypothesis of the form $\mathbf{H}_0 : r(\beta) = \theta_0$ where $r : \mathbb{R}^k \to \mathbb{R}^q$ is some known and differentiable *function* of the regression vector β. Introduce the shorthand that $\theta = r(\beta)$.

By the continuous mapping theorem, we know that $\hat{\theta} := r(\hat{\beta})$ is a consistent estimator of the parameter θ, and by the Delta method we know that

$$\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{d} N(0, \mathbf{R}'\mathbf{V}\mathbf{R})$$

where the $k \times q$ matrix $\mathbf{R} = \nabla r(\theta)$ is the Jacobian matrix of r, composed of all of it's derivatives evaluated at θ. By the same steps as those that establish Proposition 7.10, we have that $\sqrt{n}(\mathbf{R}'\mathbf{V}\mathbf{R})^{-1/2}(\hat{\theta}-\theta) \xrightarrow{d} N(0, \mathbb{I}_k)$, where notice now that since θ has q rather than k components, we have the $q \times q$ identity matrix appearing in the asymptotic variance.

Similar to Eq. (7.45), we can now form a Wald statistic for the hypothesis $r(\beta) = \theta_0$, which has a chi-squared distribution with q, rather than k degrees of freedom:

$$n(\hat{\theta} - \theta)'(\mathbf{R}'\hat{\mathbf{V}}\mathbf{R})^{-1}(\hat{\theta} - \theta) \xrightarrow{d} \chi^2_q \tag{7.46}$$

We can perform tests of very general hypotheses about β based upon Eq. (7.46). Note that in implementing the Wald-test, it is important that the matrix $\mathbf{R}$ is known in order to construct the test statistic (7.46). Fortunately we do know $\mathbf{R}$ under the null-hypotheses, which this fixes the value of θ and the function r is provided by the researcher (and if the function r is linear, then $\mathbf{R}$ doesn't even depend on the value of θ). *Note:* it is common to use critical values from the so-called F-distribution rather than the χ^2 distribution when using (7.46) to construct tests (this also involves rescaling the test statistic by a factor of $1/q$). This is analagous to the distinction between t and z tests discussed above: the F test is more conservative and may do better if the residuals are close to normally distributed and n is small.

It is illustrative to see that a test based on (7.46) recovers a test that is equivalent to the *t-test* considered in the last section, when the function $r(\beta) = \mathbf{e}_j'\beta$ picks out a single component of β. In this case $\mathbf{R} = \mathbf{e}_j$. The Wald statistic becomes

$$n(\mathbf{e}_j'\beta - \mathbf{e}_j'\beta)'(\mathbf{e}_j'\hat{\mathbf{V}}\mathbf{e}_j)^{-1}(\mathbf{e}_j'\beta - \mathbf{e}_j'\beta)' = n(\beta_j - \beta_0)\hat{\mathbf{V}}_{jj}^{-1}(\beta_j - \beta_0) = \frac{(\beta_j - \beta_0)^2}{\hat{\mathbf{V}}_{jj}/n} = T(\beta_0)^2$$

exactly the square of the t-statistic for $\mathbf{H}_0 : \beta_j = \beta_0$. The $1 - \alpha$ quantile of the χ^2_1 distribution is equal to the square of the $1 - \alpha/2$ quantile of the standard normal distribution. Thus, a two-tailed t-test rejects the null that $\beta_j = \beta_0$ exactly when the Wald test does, and vice versa.

Bibliography

ANGRIST, J. D. and PISCHKE, J.-S. (2008). *Mostly Harmless Econometrics.*

DALE, S. B. and KRUEGER, A. B. (2002). "Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables*". *The Quarterly Journal of Economics* 117 (4), pp. 1491–1527. eprint: https://academic.oup.com/qje/article-pdf/117/4/1491/5304491/117-4-1491.pdf.

FAN, Y. and PARK, S. S. (2010). "Sharp Bounds on the Distribution of the Treatment E§ect and Their Statistical Inference". *Econometric Theory* 26 (3), 931–951.

HANSEN, B. (2021). "A Modern Gauss-Markov Theorem". *Econometrica.*

HECKMAN, J. J., SMITH, J. and CLEMENTS, N. (1997). "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts". *The Review of Economic Studies* 64 (4), pp. 487–535.

LEWIS, D. (1973). "Causation". *The Journal of Philosophy* 70 (17), pp. 556–567.

ROSENBAUM, P. and RUBIN, D. (1983). "The central role of the propensity score in observational studies for causal effects". *Biometrika* 70 (1), pp. 41–55. eprint: https://academic.oup.com/biomet/article-pdf/70/1/41/662954/70-1-41.pdf.

ROSENTHAL, J. (2006). *A First Look at Rigorous Probability Theory* -. World Scientific.

RUBIN, D. B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66 (5). Place: US Publisher: American Psychological Association, pp. 688–701.

YITZHAKI, S. (1996). "On Using Linear Regressions in Welfare Economics". *Journal of Business and Economic Statistics* 14 (4), pp. 478–486.